ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ISSN: 3041-0894 (Print)
ISSN: 3029-2387 (Online)
Volume 1 · Issue 4
October 2024

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

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

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

About the Publisher

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

Contact Us

Managing Editor
aih.office@accscience.sg

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

Volume 1 • Issue 4 • October 2024
ISSN 3041-0894 (print) ISSN 3029-2387 (online)

ARTIFICIAL INTELLIGENCE IN HEALTH

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

ACCSCIENCE PUBLISHING
Access Science Without Barriers

Artificial Intelligence in Health

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

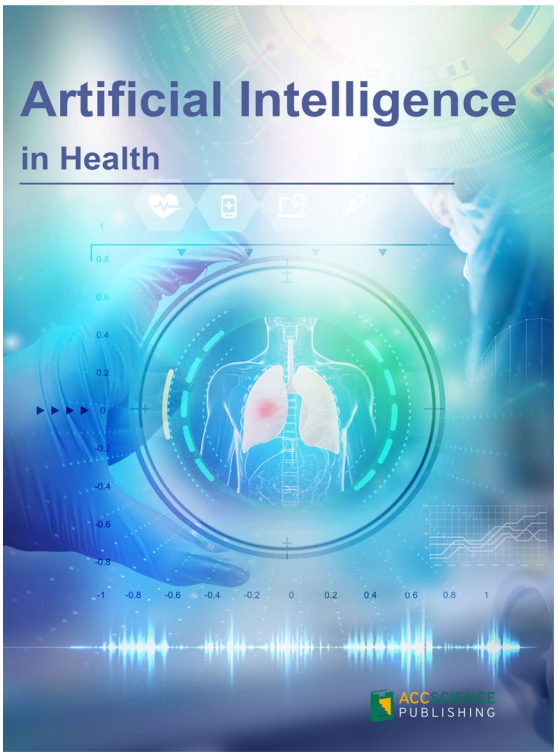
Editorial and Production Credits

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

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

Supplementary file

Supplementary files of articles can be obtained at https://accscience.com/journal/AIH/1/4.



Artificial Intelligence in Health

Editorial Board

ACCSCIENCE PUBLISHING

Volume 1 | Issue 4 | October 2024 | Page 1 – 122 |

Artificial Intelligence in Health

https://accscience.com/journal/AIH

CONTENTS

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

Prognostic evaluation using radiomics after stereotactic body radiotherapy in early-stage lung cancer

Melek Yakar*

Department of Radiation Oncology, Faculty of Medicine, Osmangazi University, Eskişehir, Turkey

Abstract

Non-small cell lung cancer (NSCLC), the leading cause of cancer-related deaths, is the most common subtype of lung cancer with an incidence of 85%. Stereotactic body radiotherapy (SBRT) is a curative treatment option for patients with early-stage NSCLC who cannot undergo surgery due to medical reasons or who refuse surgery. Radiomics non-invasively extracts advanced imaging features invisible to the human eye from medical images. Radiomics has prognostic value in predicting oncological outcomes after lung SBRT. Although studies on this subject are available in the literature, they are quite heterogeneous. There is a need for large-scale multicenter studies in which standard imaging techniques are used to obtain radiomic features, artificial intelligence-based segmentations are used to eliminate differences between contours, and SBRT dose schemes with appropriate therapeutic indexes are applied. This review aimed to interpret the existing studies and emphasize the clinical importance of radiomics, which can contribute to personalized treatment. A comprehensive literature search was conducted through the PubMed database using a wide range of keywords, which yielded 11 peer-reviewed articles published between 2017 and 2024. Seven articles evaluated computed tomography radiomics, and four evaluated fluorodeoxyglucose positron emission tomography-computed tomography radiomics. Oncological outcomes are not always identical in patients with a similar history receiving similar treatments at the same stage and age. Clinical, demographic, or treatment-related data are insufficient to predict prognosis and determine personalized treatment. Incorporating radiomics to these data can help establish models with higher accuracy and achieve personalized treatment.

Keywords: Artificial intelligence; Lung cancer; Stereotactic body radiotherapy; Prognosis; Radiomics

***Corresponding author:**
Melek Yakar
(myakar@ogu.edu.tr)

Citation: Yakar M. Prognostic evaluation using radiomics after stereotactic body radiotherapy in early-stage lung cancer. *Artif Intell Health*. 2024;1(4):1-11. doi: 10.36922/aih.3541

Received: April 30, 2024

Accepted: August 1, 2024

Published Online: October 16, 2024

1. Introduction

Non-small cell lung cancer (NSCLC) is the most common subtype of lung cancer with an incidence of 85% and has the highest cancer-related mortality. Stereotactic body radiotherapy (SBRT) is a curative treatment option for patients with early-stage NSCLC who cannot undergo surgery due to medical reasons or who refuse surgery. Although 92 – 98% of local tumor control can be achieved using SBRT in these patients, varying recurrence patterns have been reported in 18 – 20% of the patients.[1] Stereotactic irradiation focuses multiple X-rays at different angles on a small localized lesion, provided

that good immobilization is achieved. The most important difference between SBRT and conventional fractionation is that high doses are delivered to the target volume in several fractions, resulting in a high biological effective dose (BED). Li *et al*. reported a statistically significant difference in 3- and 5-year local control between the SBRT arm and conventional fractionated radiotherapy with SBRT arm.[2]

Recent advances in science and technology have revealed that each tumor, even within the same type of cancer, has several varying phenotypic and genotypic characteristics. This inhomogeneity between tumors leads to different oncological responses to standard treatments administered at the same cancer stage.[3] Identifying the factors associated with relapse is important to initiate salvage treatment for patients as soon as possible. Underdosing may be a reason for tumor recurrence after SBRT. Therefore, predicting tumor response before treatment can help in adjusting dose prescription to prevent relapses.

Radiomics quantitatively describes the characteristics of medical images. It calculates features and provides numerical values using mathematical formulas. Radiomics features are based on the distribution of pixels and voxels in the region of interest (ROI) and the relationship between them. Radiomics provides pixel and voxel characteristics of images that are indistinguishable to the human eye.[4] Personalized treatment involves treatment tailored to each patient according to tumor characteristics, with the aim of improving oncological outcomes and obtaining a good therapeutic index while reducing side effects. The principle of personalized treatment is based on determining the heterogeneous structure of the tumor and its characteristic features and elucidating the treatment option accordingly.[5]

Radiomics has prognostic value in predicting oncological outcomes after lung SBRT. However, prediction models have been created using parameters such as radiomics, dosimetrics, and patient and treatment characteristics to evaluate local tumor control, recurrence, and tumor-related survival in patients with lung cancer receiving SBRT.

This review aimed to summarize the role of radiomics features obtained from different imaging methods in predicting the prognosis of patients receiving lung SBRT. Moreover, it emphasizes the clinical importance of radiomics and its ability to contribute to personalized treatment by interpreting existing studies.

2. Material and methods

2.1. Search strategy and study selection

A comprehensive literature search was conducted in the PubMed database using a wide range of keywords, including "Lung cancer," "SBRT," "radiomics," "computed tomography," "FDG PET-CT," "prognosis," "prediction of response to treatment," "survival," "local control," and "recurrence."

This review included studies evaluating quantitative features extracted from baseline or follow-up computed tomography (CT) or positron emission tomography (PET)-CT scans against treatment response in patients treated with SBRT for NSCLC of any stage or lung metastases. Studies whose full texts were available were included. The exclusion criteria were (1) studies that did not evaluate radiological response as an end point, (2) studies focusing entirely on methodological aspects of radiomics, (3) studies using phantoms or animal models, and (4) studies without original data, such as reviews and editorials. In total, 11 peer-reviewed articles published during 2017 – 2024 were included, of which seven articles evaluated CT radiomics, and four articles evaluated fluorodeoxyglucose (FDG) PET-CT radiomics.

3. Prognosis evaluation using radiomics

According to studies in the literature, CT and even more frequently planning CTs for radiotherapy are generally used to extract radiomics features, and PET-CT is used less frequently. In some studies, clinical and dosimetric features were also added to the models established using radiomics features.

The effect of radiomics features obtained from different imaging methods on prognostication has not been explored comparatively. Some studies have reported the varying impact of different imaging techniques and estimation algorithms on radiomics features.[6-8] There is no standard imaging, scanning parameter, algorithm, or radiomics feature. Data from existing studies and subsequent multicenter randomized studies are required for standardization.

3.1. CT-based radiomics models

CT and rarely magnetic resonance imaging are used during radiotherapy planning, and extensive patient data are accumulated during the standard treatment planning process. Radiomics features obtained from CT are used to predict oncological outcomes such as evaluating treatment response and predicting prognosis and recurrence patterns. Although planning CT is frequently used because CT is currently performed at the treatment planning stage, some studies use diagnostic thorax CT acquired at the diagnosis stage or during follow-up.

Hyunh *et al*. evaluated 112 patients diagnosed with early-stage lung cancer who received SBRT using both static-free breathing (FB) CT and respiratory-gated CT (average intensity projection [AIP]). The median SBRT

BED was 151.2 (range: 100 – 151.2) Gy. Radiomics features of all patients were obtained using both FB CT and AIP CT. A total of 644 radiomics features were obtained using the MATLAB 2013 toolbox. Both FB CT and AIP CT yielded 19 significant radiomics features, with six features (two features of tumor shape, three features of intensity histogram or statistics of the tumor, and one feature of homogeneity of the tumor tissue) being similar in both image sets and the remaining 13 being completely different from each other but similar to each image type in both image sets. The unique radiomics features of FB CT images were statistical ($n = 6$) and textural ($n = 7$) features. Of the 13 unique radiomics features of AIP CT images, there were eight textural, four statistical, and one shape feature. The median follow-up duration was 20.8 months. Distant metastasis (DM) developed in 20.5% of the patients ($n = 23$), and the average time to metastasis was 10.0 months. Locoregional recurrence (LRR) developed in 21.4% of the patients ($n = 24$), and the median time until LRR developed was 8.8 months. The 2-year estimates for LRR and DM were 70.9% and 74.0%, respectively. The concordance index (CI) was used to evaluate prognostic performance. We found that FB CT radiomics features were unrelated to DM development, whereas AIP CT radiomics features describing tumor shape and heterogeneity were compatible with DM development (CI: 0.638 ± 0.676). No imaging radiomics features were associated with LRR in their study. Compared with FB CT images, AIP CT images contained more valuable information regarding disease recurrence in patients with early-stage NSCLC receiving SBRT. According to their study, AIP CT images may be a valuable non-invasive technique to predict recurrence.[9]

Kakino *et al*. conducted a local and distant recurrence prediction study using breath-hold CT-based radiomics features in 573 patients with early-stage NSCLC who received SBRT. The patients were categorized into two groups: 464 patients (from 10 centers) were evaluated as the training group and 109 patients (from 1 center) were evaluated as the test group. Important variables were determined using the adaptive least absolute shrinkage and selection operator (LASSO) method. Three prognostic models (clinical, radiomics, and combined) were trained using the random survival forest (RSF) algorithm. The CI values in the clinical, radiomics, and combined models constructed using clinical and radiomics features were 0.57, 0.55, and 0.61 for LRR and 0.59, 0.67, and 0.68 for DM, respectively. According to their study, although DM could be predicted through the RSF algorithm in patients with early-stage NSCLC who received SBRT using breath-hold CT-based radiomics features, the same algorithm was considered to have a lower potential to predict LRR.[10]

Sawayangi *et al*. randomly divided 358 patients treated with SBRT into training (250 patients) and validation (108 patients) groups and estimated the overall survival (OS) using both clinical variables and CT-based radiomics features. They applied 42 – 64 Gy SBRT in 4 – 10 fractions (BED10 range: 75 – 166 Gy). To extract radiomics features, the gross tumor volume (GTV) was contoured in the expiratory phase of the planning CT. Radiomics features categorized as gray level size region matrix features calculated from GTV in the pretreatment CT image had high accuracy for OS prediction in patients with early-stage NSCLC treated with curative SBRT. These findings indicate that the operating system predicted from multiple linear regression analysis is similar to the actual operating system. These data may be important in selecting patients who would benefit from increasing treatment intensity, such as increasing the radiotherapy dose or using different fractionations and adding other treatments such as chemotherapy/immunotherapy/targeted therapies.[11]

Luo *et al*. evaluated 119 patients and 129 lesions treated with SBRT with a median of 48 Gy (range: 18 – 70 Gy) in a median of four fractions (range: 1 – 12). Like other studies, they created clinical, radiomics, and combined models. They generated radiomics features based on planning CT. Of 1502 radiomics features, four were identified as important variables using the LASSO method. Logistic regression (LR), decision tree, and support vector machine (SVM) algorithms were used to create the optimal model for evaluating local tumor control. LR was determined as the prediction algorithm with the highest accuracy rate, with the accuracy rates of the radiomics, clinical, and combined models being 67.4%, 82.0%, and 85.4% in the training group and 92.9%, 77.5%, and 82.5% in the validation group, respectively. The combined model performed statistically significantly better than the radiomics ($P = 0.025$) and clinical ($P = 0.033$) models in the training group, whereas both radiomics and clinical models showed similar performance ($P = 0.613$). According to their study, the combined model based on radiomics features and clinical and dosimetric parameters can be used to predict 1-year local tumor control in patients with lung cancer receiving SBRT.[12]

Isoyama-Shirakawa *et al*. conducted a study on 125 patients with early-stage NSCLC treated with SBRT. They generated the radiomics score and investigated the effect of this score in predicting LC and metastasis-free survival (MFS). The median BED10 for SBRT was 88.9 (range: 61.2 – 119.0) Gy. Planning CT-based radiomics features were generated. From 432 radiomics features, five important variables were identified using the LASSO method. The radiomics score was obtained using five significant radiomics features, which was statistically

significantly associated with both LC (>0.043 vs. ≤0.043, $P = 0.042$) and MFS (>0.304 vs. ≤0.304, $P < 0.001$). Considering clinical factors, their study suggested that tumor histology, tumor diameter, and radiomics score are vital factors in predicting NSCLC recurrence patterns after SBRT.[13]

Lafata *et al.* evaluated the FB planning CT-based radiomics features of 70 patients with early-stage NSCLC who received SBRT with a mean of 51 Gy at Duke University. Only two of 43 radiomics features, namely, Homogeneity2 and Long-Run-High-Gray-Level-Emphasis, were considered significant variables and were associated with LC. The area under the curve (AUC) values of the multivariable LR prediction models for recurrence, local recurrence, and non-local recurrence were 0.72 ± 0.04, 0.83 ± 0.03, and 0.60 ± 0.04, respectively. These data suggest that relatively dense tumors with homogeneous rough tissue are associated with higher rates of local recurrence. This was supported by both univariate and multivariate analyses that showed that CT-based radiomics features may be associated with local recurrences after SBRT in patients with Stage 1 NSCLC.[14]

Unlike other studies, Li *et al.* evaluated radiomics features in follow-up CT after SBRT. In their study, a dose of 50 Gy was administered in five fractions to 54 patients with early-stage NSCLC, 48 Gy was administered in four fractions to three patients, and 60 Gy was administered in eight or five fractions to two patients. The first control CT was taken 1 – 3 months after the completion of SBRT (median 91 days, range: 33 – 112 days) and used to extract radiomics features. As imaging features, 34 manually determined radiological features (semantics) describing the lesion, lung, and thorax, and 219 quantitative imaging features (radiomics) for the lesion were extracted. Cox proportional hazards models and Harrell's C index were used to predict OS, relapse-free survival (RFS), and locoregional RFS (LR-RFS), and a five-fold cross-validation was performed for the prognostic model. Data from a median follow-up of 42 months were available. Eastern Cooperative Oncology for OS model Group performance status, vascular involvement, lymphadenopathy, and radiomics features considered important were included in the study. Vascular involvement, pleural retraction, lymphadenopathy, vessel attachment, and relative enhancement were included in the RFS model. In the LR-RFS model, vascular involvement, lymphadenopathy, circularity, and radiomics features were included in the study. The AUC, which was used to evaluate the performance of the models, was >0.8 for all models. According to their study, the disease progression can be predicted even 3 months after SBRT using the model established using CT imaging features, which might be helpful in clinical decision-making.[15] The abovementioned studies are summarized in Table 1.

The field of radiomics, which quantitatively analyzes extensive data obtained from medical images, has significant potential for predicting treatment outcomes. Nevertheless, a standardized imaging method has not yet been developed. Moreover, the segmentation of the target region for the collected data varies among researchers. Studies have been conducted using several varying radiomics features using different software, and no consensus has been achieved on this issue. Multicenter radiomics-based prediction algorithms with higher accuracy can be generated using a standard imaging method and artificial intelligence (AI)-based segmentation.

3.2. PET- and CT-based radiomics models

FDG PET-CT is a molecular imaging technique that combines metabolic and functional assessments, improving the diagnostic accuracy and initial staging and restaging of lung cancer and influencing treatment optimization and monitoring of treatment response. Although most studies on PET-CT have shown that the standardized uptake value can be used as a prognostic indicator of OS, some studies do not support this.[16,17]

Evaluation of tumor heterogeneity is gaining importance due to radiomics features obtained from different imaging modalities. If prognosis can be predicted accurately before treatment through radiomics, the most beneficial personalized treatment can be administered to patients. Prognostic prediction studies have been conducted using a combination of radiomics and more classical PET parameters, and current prognostic prediction studies are conducted using both PET and CT radiomics features to create models with higher accuracy.[18,19]

Dissaux *et al.* conducted a prognostic prediction study using 18F-FDG PET-CT radiomics features in patients with early-stage lung cancer receiving SBRT. They included 64 patients from three different centers in the training set and 23 patients from one center in the test set. The primary tumor was segmented semiautomatically using the fuzzy locally adaptive Bayesian algorithm on PET images and manually on low-dose CT images. ComBat was used to harmonize the radiomics features obtained from the four institutions, yielding 184 (92 PET and 92 CT) radiomics features. In the training set, significant variables in the univariate analysis were added to the multivariate regression model. Models were constructed by combining the obtained independent prognostic factors. Two important variables were each obtained from PET and CT. The median follow-up duration was 21.1 (range: 1.7 – 63.4) and 25.5 (range: 7.7 – 57.8) months in

Table 1. SBRT response prediction studies using CT-based radiomics

Study and year	Imaging method used	Number of patients	Number of radiomics features and radiomics extraction method	Number of significant radiomics features	Important radiomics identification method	Algorithm/statistical method used	Oncological outcomes
Huynh *et al*., 2017[9]	FB CT and AIP CT	112	644, MATLAB 2013 toolbox	19	PCA and factor analysis (FactoMineR package)	Univariate and multivariate analyses	Seven AIP radiomics features were associated with DM (CI: 0.638±0.676)
Li *et al*., 2017[15]	Follow-up thorax CT after SBRT	59	219, Definiens Developer	166	Pearson's correlation	LR	LR model AUC (CI): OS: 0.88 (0.78 – 0.97) RFS: 0.86 (0.76 – 0.96) LR-RFS: 0.85 (0.74 – 0.95)
Lafata *et al*., 2019[14]	FB planning CT	70	43, MATLAB	2	SVD, LASSO	Multivariable LR models	AUC values of 0.72±0.04, 0.83±0.03, and 0.60±0.04, for recurrence, local recurrence, and non-local recurrence, respectively
Kakino *et al*., 2020[10]	Breath-hold planning CT	573	944, PyRadiomics Ver. 2.2.0	16	Adaptive LASSO	RSF	CI values for LR in clinical, radiomics, and combined models were 0.57 (0.39 – 0.75), 0.55 (0.38 – 0.73), and 0.61 (0.43 – 0.78), respectively Corresponding value for DM 0.59 (0.54 – 0.79), 0.67 (0.54 – 0.79), and 0.68 (0.55 – 0.81)
Luo *et al*., 2022[12]	FB planning CT	119	1502, PyRadiomics	4	LASSO	LR, DT, SVM	Best result for 1-year LC prediction in the combined model with LR AUC of 0.911 (training) and 0.818 (validation)
Sawayanagi *et al*., 2022[11]	FB planning CT	358	107, PyRadiomics v3.0.1	4 (OS) 2 (LRFS) 3 (PFS)	Regression model	Multiple linear regression	One radiomics factor remained a significant prognostic factor of overall survival (OS) (P=0.044)
Isoyama-Shirakawa *et al*., 2023[13]	FB planning CT	125	432, MATLAB-based Radiomics tools package	5	LASSO	Kaplan–Meier analysis and log-rank test	Radiomics score was significantly associated with MFS (P<0.001) and LC (P=0.042)

Abbreviations: FB: Free breathing; AIP: Average intensity projection; PCA: Principal component analysis; DM: Distant metastasis; CI: Concordance index; LASSO: Least absolute shrinkage and selection operator; RSF: Random survival forest; OS: Overall survival; LRFS: Local relapse-free survival; PFS: Progression-free survival; LR: Logistic regression; DT: Decision tree; SVM: Support vector machine; LC: Local control; MFS: Metastasis-free survival; SVD: Singular value decomposition; LR-RFS: Locoregional recurrence-free survival.

the training and test sets, respectively. No clinical variable was predictive of local tumor control in the univariate analysis. Similarly, two PET radiomics features and two CT radiomics features did not significantly predict local tumor control. The best predictive models in the training set were obtained by combining one feature from PET with one feature from CT, achieving 100% sensitivity and 96% specificity. Another model combining two PET features achieved 100% sensitivity and 88% specificity. The second model achieved an accuracy of 0.91 (sensitivity, 100%; specificity, 81%) in the test set. According to their study, two radiomics features derived from 18F-FDG PET were independently associated with local tumor control in patients with NSCLC receiving SBRT and can be combined into an accurate predictive model. This model can provide information regarding local tumor recurrence and assist in clinical decision-making.[20]

Oikonomou *et al*. evaluated the recurrence prediction algorithm using both PET-CT radiomics features and maximum standardized uptake value (SUVmax) values. Their study included 150 patients and 172 tumors, and 42 features were obtained from CT and PET. There were 11 important variables in the prediction models. OS, disease-specific survival, and regional control were estimated in the model established using radiomics features; however, neither SUVmax nor DFS could be predicted using radiomics models.[21]

In another study, 60% of 464 patients with early-stage lung cancer who received SBRT were included in the training set, 40% were included in the test set, and 63 patients from another center were included in the external test set. The SBRT dose was 40 – 60 Gy administered in 3 – 5 fractions. Differences between images from the two centers were eliminated using the ComBat harmonization method. A total of 318 radiomics features (106 from each imaging) were obtained from PET, PET-CT, and planning CT using the PyRadiomics toolbox. In the training and test sets, the C-statistics value for predicting regional and/or distant recurrences using the clinical model was 0.53 – 0.59 (95% CI: 0.41 – 0.67), that using the radiomics model was 0.70 – 0.78 (95% CI: 0.63 – 0.88), and that using the combined model was 0.50 – 0.62 (95% CI: 0.37 – 0.69), indicating that the radiomics model showed the best prediction performance. According to this study, radiomics features obtained from FDG PET-CT before SBRT performed better than clinical parameters in predicting regional and/or distant recurrence and determining adjuvant systemic therapy for patients with early-stage NSCLC.[22]

Similarly, Nemoto *et al*. conducted a recurrence prediction study using radiomics features from PET and CT images. They applied SBRT to 82 patients with NSCLC at 48 Gy/4 Fr, 50 Gy/4 Fr, or 55 Gy/4 Fr for T1 tumors and 60 Gy/10 Fr or 70 Gy/10 Fr for T2 tumors. They extracted 111 radiomics features using PyRadiomics from both planning CT and PET images. Using three different methods (chi-square test, minimum redundancy maximum relevance, and ReliefF), they obtained 42 important variables and created a model using four different algorithms (random forest, SVM, K-nearest neighborhood, and naive Bayes). SVM with PET radiomics (mean AUC: 0.646), naive Bayes with PET radiomics (mean AUC: 0.611), and SVM with CT radiomics (mean AUC: 0.645) exhibited the highest performance for local recurrence, regional lymph node metastasis, and DM. Their study demonstrated that the model combining PET imaging features and SVM may be useful for predicting local and regional lymph node recurrence and the model combining CT imaging features and SVM may be useful for predicting distant recurrence.[23]

4. Can radiomics be used in clinical practice?

Radiomics can be considered a signature of tumors. With gradual advancements in technologies in the field of medicine, personalized treatments are becoming important. Oncological outcomes are not always similar in patients at the same stage and age, with the same performance score, and receiving the same treatment. Radiomics is gaining importance in this regard. Tumor features that are not visible to the clinician can be revealed through radiomics and play a key role in determining the most accurate personalized treatment. Various medical imaging technologies, which are non-invasive methods, are used for staging patients with lung cancer for treatment selection. Evaluation of medical images is not completely objective and may vary between clinicians depending on the person's experience. Some radiomics features in medical images are not visible to the human eye. As summarized earlier, studies suggest that radiomics can be used as a non-invasive adjunct tool for personalized treatment selection and prediction of oncological response. Nevertheless, several technical difficulties, especially in feature engineering and statistical modeling, and the use of different methods limit a standardized approach; hence, the clinical use of radiomics still remains under development.

A standard contouring should be performed for the tumor or tumor microenvironment where radiomics features are planned to be studied. Determining the correct radiomics features may be difficult because of differences between users. Therefore, users must be experienced in this field and follow accepted guidelines. A well-trained AI segmentation system will also help in standardization. Furthermore, if the radiation oncologist and radiologist work

together on segmentation, the accuracy of segmentation can be increased.

Differences between centers in terms of obtaining imaging studies should be eliminated. Several parameters require standardization, such as the device used for imaging, use of contrast in CT, slice thicknesses, and the time between nuclear material injection and imaging.

Extraction of features involves a comprehensive quantification of tumor phenotypes. After image acquisition, preprocessing, and segmentation, radiomics features are extracted from two-dimensional (2D) or 3D ROIs in the images. A standard method for selecting ROIs is also lacking. For metastatic diseases, there is still no clear consensus on whether the metastatic focus or the primary focus should be segmented. Moreover, only the tumor and tumor microenvironment, which is usually created at a 5-mm margin to the tumor, can be used as ROIs for feature extraction. Furthermore, there is no consensus on whether radiomics features should be obtained from the original image or filtered images.[24]

There are different types of radiomics features, such as shape, first-order, and textural features. As it remains unclear which feature should be used at which stage of tumor, most studies investigate all radiomics features. There is also no consensus on whether 2D or 3D features should be used. The 2D radiomics features have been reported to be superior in some studies, whereas 3D features are reportedly superior in other studies in terms of prognosis.[25,26]

Due to the lack of knowledge on which type of radiomics features should be used in which tumors, hundreds of radiomics features are extracted; however, if all of them are used to create a model, it will generate an excessive number of features, causing confusion and overfitting of the data and decreasing the actual accuracy. It is important to consider that the extracted radiomics features may be related to each other, and feature selection should be conducted before modeling. It is also necessary to use radiomics feature selection methods that identify the most important features and remove redundant ones. As a standardized method for determining important variables has not yet been developed, different methods are used in various studies.[27]

After determining the important variables among 100 of radiomics features using appropriate methods, an accurate model is established. Researchers use different machine learning algorithms to create models. Each study has reported different algorithms with the best accuracy. This leads to further confusion among clinicians. Thus, the most effective algorithm for determining prognosis remains unclear.

Furthermore, data must be distributed evenly, which otherwise would increase the risk of overfitting. Radiomics studies are still ongoing at the clinical trial level, and there exist several parameters that require standardization. In this regard, clinicians and engineers must standardize these parameters through a joint investigation. Standardization of parameters is essential for the use of radiomics in routine clinical practice; otherwise, each study may identify different radiomics features as important variables, resulting in varying accuracy rates of different algorithms in multiple studies. Considering this heterogeneity, it would be inappropriate to use radiomics as a standard in routine clinical practice.

As technology advances, the importance of data also increases. Performing tumor and organ-at-risk segmentation in the radiotherapy planning stage is a routine procedure for radiation oncologists and is included in the treatment planning of each patient. From this perspective, algorithms need to be trained on a large amount of processed data in radiation oncology clinics. These valuable data should not be ignored and must be considered to contribute significantly to standardization.

In summary, radiomics is a promising non-invasive method for prognostic prediction and post-treatment follow-up, wherein medical images can be analyzed. It has the potential to facilitate personalized treatment selection with low cost and high sensitivity. If the appropriate personalized treatments are administered, patients can avoid unnecessary treatment and high treatment-related costs. Radiomics can be used as an important biomarker in treatment decisions, prognosis determination, and follow-up after accurate standardization.

5. Delta radiomics

Radiomics has been developed to evaluate feature changes at different time points, an approach often referred to as "delta radiomics." This method is used to examine the effects of feature changes after certain steps in the patient's workflow (for instance, after a certain treatment).[28-30] Delta radiomics is a promising field of research and allows changing and modulating the treatment approach due to its predictive power.[31]

The vast majority of radiomics methods used in existing studies depend on imaging data obtained at a single time point, often imaging tumors before the start of treatment. Delta radiomics reveals the changes in feature values during treatment by extracting quantitative features from image sets obtained throughout the treatment process.[32,33] Delta radiomics may improve cancer diagnosis, prognosis, patient follow-up, or evaluation of treatment response.[34,35] Some studies have shown that delta radiomics is effective

in investigating immunotherapy response or predicting relapse in oncological patients.[36-38]

Delta radiomics has been used and found to be useful in evaluating the response to chemotherapies in colorectal cancer, liver metastases, and metastatic renal cell cancer.[39,40] Delta radiomics has also been used to predict the risk of developing radiation pneumonitis during treatment in patients with esophageal cancer.[41] Delta radiomics features obtained from PET-CT images have been used to estimate prognosis in patients with NSCLC.[32]

Changes in the tumor during treatment can be detected using delta radiomics, which can thus help modify treatment strategies. For instance, in a patient planned for neoadjuvant radiotherapy, radioresistant tumors can be detected using delta radiomics, and the patient can be referred to surgery earlier. When unresponsive patients for lung SBRT are detected during this process, they can be protected from unnecessary treatment toxicities. Hence, standardizing the stages of obtaining delta radiomics can both contribute to personalized treatments and protect patients from unnecessary treatment toxicities.

6. AI and ethical issues in cancer treatment

Ethical issues surrounding AI in healthcare concern privacy, bias, and discrimination, as well as whether it can replace human judgment. Where there is technology, there is always the risk of inaccuracy and data breach. Moreover, wrong decisions can result in undesirable and devastating consequences in the treatment of patients with cancer. There is no clear regulation on legal and ethical issues regarding the role played by AI in healthcare; therefore, this issue needs careful consideration.

Innovations and new developments in technology accelerate scientific progress. It is important to explore strategies to eliminate the potentially disastrous problems of AI technologies.

Machine learning algorithms are effective in identifying and analyzing or classifying large amounts of data, referred to as "Big Data." Big Data are used to train algorithms. Using more data to train an algorithm generally increases the accuracy rate of the algorithm. The machine requires a set of rules and instructions, generally written in the form of algorithms, to perform tasks. Nevertheless, when newly acquired data are used, the machine gradually gains the ability to become more flexible and operate in different situations accordingly. This situation may increase the demand for data and sometimes cause sharing of personal or public information without considering user privacy. Ethics and moral values vary across countries and even regions within countries. Ethnic groups and nations have different norms. When defining values from humans to machines, ethical rules such as deontology and virtue found in humans should also be considered for using machines.

Our current moral systems are derived fundamentally from our responsibilities to other people. Therefore, a non-human system, *that is*, AI, cannot be expected to understand existing moral systems.[42] These issues must be resolved before using AI in human-related decisions.

AI systems should be data stewards. Only the required data may be used and then deleted, which is also defined as "data minimization." Data should be encrypted and used only by authorized individuals. Data must be collected, used, and shared according to privacy and personal data laws. Before using patient data, patients' consent must be obtained, and Ethics Committee permission should also be obtained from the concerned authorities.[43]

7. Conclusion

SBRT is considered the first treatment option with similar oncological outcomes in patients with early-stage NSCLC who cannot undergo surgery or refuse surgery for medical reasons.

It is important to determine the patient's risk of recurrence during the treatment planning stage to determine the most ideal personalized treatment. If patients with a high risk of recurrence can be selected in advance, the treatment intensity can be increased by changing the radiotherapy dose or schedule.

Acknowledgments

None.

Funding

None.

Conflict of interest

The author declares having 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.

References

1. Chang JY, Mehran RJ, Feng L, *et al*. Stereotactic ablative radiotherapy for operable stage I non-small-cell lung cancer (revised STARS): Long-term results of a single-arm, prospective trial with prespecified comparison to surgery. *Lancet Oncol*. 2021;22(10):1448-1457.

   doi: 10.1016/S1470-2045(21)00401-0

2. Li C, Wang L, Wu Q, *et al*. A meta-analysis comparing stereotactic body radiotherapy vs conventional radiotherapy in inoperable stage I non-small cell lung cancer. *Medicine* (*Baltimore*). 2020;99(34):e21715.

   doi: 10.1097/MD.0000000000021715

3. Jameson JL, Longo DL. Precision medicine--personalized, problematic, and promising. *N Engl J Med*. 2015;372(23):2229-2234.

   doi: 10.1056/NEJMsb1503104

4. Bertolini M, Trojani V, Botti A, *et al*. Novel harmonization method for multi-centric radiomic studies in non-small cell lung cancer. *Curr Oncol*. 2022;29(8):5179-5194.

   doi: 10.3390/curroncol29080410

5. Lambin P, Rios-Velazquez E, Leijenaar R, *et al.* Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441-446.

   doi: 10.1016/j.ejca.2011.11.036

6. Fave X, Cook M, Frederick A, *et al*. Preliminary investigation into sources of uncertainty in quantitative imaging features. *Comput Med Imaging Graph*. 2015;44:54-61.

   doi: 10.1016/j.compmedimag.2015.04.006

7. Zhao B, Tan Y, Tsai WY, *et al*. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep*. 2016;6:23428.

   doi: 10.1038/srep23428

8. Mackin D, Fave X, Zhang L, *et al*. Measuring computed tomography scanner variability of radiomics features. *Invest Radiol*. 2015;50(11):757-65.

   doi: 10.1097/RLI.0000000000000180

9. Huynh E, Coroller TP, Narayan V, *et al*. Associations of radiomic data extracted from static and respiratory-gated CT scans with disease recurrence in lung cancer patients treated with SBRT. *PLoS One*. 2017;12(1):e0169172.

   doi: 10.1371/journal.pone.0169172

10. Kakino R, Nakamura M, Mitsuyoshi T, *et al.* Application and limitation of radiomics approach to prognostic prediction for lung stereotactic body radiotherapy using breath-hold CT images with random survival forest: A multi-institutional study. *Med Phys*. 2020;47(9):4634-4643.

    doi: 10.1002/mp.14380

11. Sawayanagi S, Yamashita H, Nozawa Y, *et al*. Establishment of a prediction model for overall survival after stereotactic body radiation therapy for primary non-small cell lung cancer using radiomics analysis. *Cancers* (*Basel*). 2022;14(16):3859.

    doi: 10.3390/cancers14163859

12. Luo LM, Huang BT, Chen CZ, *et al*. A combined model to improve the prediction of local control for lung cancer patients undergoing stereotactic body radiotherapy based on radiomic signature plus clinical and dosimetric parameters. *Front Oncol*. 2022;11:819047.

    doi: 10.3389/fonc.2021.819047

13. Isoyama-Shirakawa Y, Yoshitake T, Ninomiya K, *et al*. Combination of clinical factors and radiomics can predict local recurrence and metastasis after stereotactic body radiotherapy for non-small cell lung cancer. *Anticancer Res*. 2023;43(11):5003-5013.

    doi: 10.21873/anticanres.16699

14. Lafata KJ, Hong JC, Geng R, *et al*. Association of pre-treatment radiomic features with lung cancer recurrence following stereotactic body radiation therapy. *Phys Med Biol*. 2019;64(2):025007.

    doi: 10.1088/1361-6560/aaf5a5

15. Li Q, Kim J, Balagurunathan Y, Qi J, *et al*. CT imaging features associated with recurrence in non-small cell lung cancer patients after stereotactic body radiotherapy. *Radiat Oncol*. 2017;12(1):158.

    doi: 10.1186/s13014-017-0892-y

16. Na F, Wang J, Li C, Deng L, Xue J, Lu Y. Primary tumor standardized uptake value measured on F18-Fluorodeoxyglucose positron emission tomography is of prediction value for survival and local control in non-small-cell lung cancer receiving radiotherapy: Meta-analysis. *J Thorac Oncol*. 2014;9(6):834-842.

    doi: 10.1097/JTO.0000000000000185

17. Agarwal M, Brahmanday G, Bajaj SK, Ravikrishnan KP, Wong CY. Revisiting the prognostic value of preoperative (18)F-fluoro-2-deoxyglucose ((18)F-FDG) positron emission tomography (PET) in early-stage (I & II) non-small cell lung cancers (NSCLC). *Eur J Nucl Med Mol Imaging*. 2010;37(4):691-698.

    doi: 10.1007/s00259-009-1291-x

18. Wu J, Aguilera T, Shultz D, *et al*. Early-stage non-small cell lung cancer: Quantitative imaging characteristics of (18)F fluorodeoxyglucose PET/CT allow prediction of distant metastasis. *Radiology*. 2016;281(1):270-278.

    doi: 10.1148/radiol.2016151829

19. Pyka T, Bundschuh RA, Andratschke N, *et al*. Textural features in pre-treatment [F18]-FDG-PET/CT are correlated with risk of local recurrence and disease-specific survival in early stage NSCLC patients receiving primary stereotactic

radiation therapy. *Radiat Oncol*. 2015;10:100.

doi: 10.1186/s13014-015-0407-7

20. Dissaux G, Visvikis D, Da-Ano R, *et al*. Pretreatment ^{18}F-FDG PET/CT radiomics predict local recurrence in patients treated with stereotactic body radiotherapy for early-stage non-small cell lung cancer: A multicentric study. *J Nucl Med*. 2020;61(6):814-820.

doi: 10.2967/jnumed.119.228106

21. Oikonomou A, Khalvati F, Tyrrell PN, *et al*. Radiomics analysis at PET/CT contributes to prognosis of recurrence and survival in lung cancer treated with stereotactic body radiotherapy. *Sci Rep*. 2018;8(1):4003.

doi: 10.1038/s41598-018-22357-y

22. Lucia F, Louis T, Cousin F, *et al*. Multicentric development and evaluation of [^{18}F]FDG PET/CT and CT radiomic models to predict regional and/or distant recurrence in early-stage non-small cell lung cancer treated by stereotactic body radiation therapy. *Eur J Nucl Med Mol Imaging*. 2024;51(4):1097-1108.

doi: 10.1007/s00259-023-06510-y

23. Nemoto H, Saito M, Satoh Y, *et al*. Evaluation of the performance of both machine learning models using PET and CT radiomics for predicting recurrence following lung stereotactic body radiation therapy: A single-institutional study. *J Appl Clin Med Phys*. 2024;25:e14322.

doi: 10.1002/acm2.14322

24. Zhang YP, Zhang XY, Cheng YT, *et al*. Artificial intelligence-driven radiomics study in cancer: The role of feature engineering and modeling. *Mil Med Res*. 2023;10(1):22.

doi: 10.1186/s40779-023-00458-8

25. Shen C, Liu Z, Guan M, *et al*. 2D and 3D CT radiomics features prognostic performance comparison in non-small cell lung cancer. *Transl Oncol*. 2017;10(6):886-894.

doi: 10.1016/j.tranon.2017.08.007

26. Zhu Y, Yao W, Xu BC, *et al*. Predicting response to immunotherapy plus chemotherapy in patients with esophageal squamous cell carcinoma using non-invasive radiomic biomarkers. *BMC Cancer*. 2021;21(1):1167.

doi: 10.1186/s12885-021-08899-x

27. Avanzo M, Wei L, Stancanello J, *et al*. Machine and deep learning methods for radiomics. *Med Phys*. 2020;47(5):e185-e202.

doi: 10.1002/mp.13678

28. Gao Y, Kalbasi A, Hsu W, *et al*. Treatment effect prediction for sarcoma patients treated with preoperative radiotherapy using radiomics features from longitudinal diffusion-weighted MRIs. *Phys Med Biol*. 2020;65(17):175006.

doi: 10.1088/1361-6560/ab9e58

29. Lorenz JW, Schott D, Rein L, *et al*. Serial T2-weighted magnetic resonance images acquired on a 1.5 tesla magnetic resonance linear accelerator reveal radiomic feature variation in organs at risk: An exploratory analysis of novel metrics of tissue response in prostate cancer. *Cureus*. 2019;11(4):e4510.

doi: 10.7759/cureus.4510

30. Mazzei MA, Nardone V, Di Giacomo L, *et al*. The role of delta radiomics in gastric cancer. *Quant Imaging Med Surg*. 2018;8(7):719-721.

doi: 10.21037/qims.2018.07.08

31. Ravanelli M, Agazzi GM, Tononcelli E, *et al*. Texture features of colorectal liver metastases on pretreatment contrast-enhanced CT may predict response and prognosis in patients treated with bevacizumab-containing chemotherapy: A pilot study including comparison with standard chemotherapy. *Radiol Med*. 2019;124(9):877-886.

doi: 10.1007/s11547-019-01046-4

32. Fave X, Zhang L, Yang J, *et al*. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. *Sci Rep*. 2017;7(1):588.

doi: 10.1038/s41598-017-00665-z

33. Ma Y, Ma W, Xu X, Cao F. How does the delta-radiomics better differentiate pre-invasive GGNs from invasive GGNs? *Front Oncol*. 2020;10:1017.

doi: 10.3389/fonc.2020.01017

34. Nasief H, Zheng C, Schott D, *et al*. A machine learning based delta-radiomics process for early prediction of treatment response of pancreatic cancer. *NPJ Precis Oncol*. 2019;3:25.

doi: 10.1038/s41698-019-0096-z

35. Lin P, Yang PF, Chen S, *et al*. A Delta-radiomics model for preoperative evaluation of neoadjuvant chemotherapy response in high-grade osteosarcoma. *Cancer Imaging*. 2020;20(1):7.

doi: 10.1186/s40644-019-0283-8

36. Liu Y, Wu M, Zhang Y, *et al*. Imaging biomarkers to predict and evaluate the effectiveness of immunotherapy in advanced non-small-cell lung cancer. *Front Oncol*. 2021;11:657615.

doi: 10.3389/fonc.2021.657615

37. Colen RR, Rolfo C, Ak M, *et al*. Radiomics analysis for predicting pembrolizumab response in patients with advanced rare cancers. *J Immunother Cancer*. 2021;9(4):e001752.

doi: 10.1136/jitc-2020-001752

38. Fatima K, Dasgupta A, DiCenzo D, *et al*. Ultrasound delta-radiomics during radiotherapy to predict recurrence in patients with head and neck squamous cell carcinoma. *Clin Transl Radiat Oncol*. 2021;28:62-70.

doi: 10.1016/j.ctro.2021.03.002

39. Rao SX, Lambregts DM, Schnerr RS, *et al*. CT texture analysis in colorectal liver metastases: A better way than size and volume measurements to assess response to chemotherapy? *United European Gastroenterol J*. 2016;4(2):257-263.

doi: 10.1177/2050640615601603

40. Goh V, Ganeshan B, Nathan P, Juttla JK, Vinayan A, Miles KA. Assessment of response to tyrosine kinase inhibitors in metastatic renal cell cancer: CT texture as a predictive biomarker. *Radiology*. 2011;261(1):165-171.

doi: 10.1148/radiol.11110264

41. Cunliffe A, Armato SG 3rd, Castillo R, Pham N, Guerrero T, Al-Hallaq HA. Lung texture in serial thoracic computed tomography scans: Correlation of radiomics-based features with radiation therapy dose and radiation pneumonitis development. *Int J Radiat Oncol Biol Phys*. 2015;91(5):1048-1056.

doi: 10.1016/j.ijrobp.2014.11.030

42. Keskinbora KH. Medical ethics considerations on artificial intelligence. *J Clin Neurosci*. 2019;64:277-282.

doi: 10.1016/j.jocn.2019.03.001

43. IEEE Global Initiative. A Vision for Prioritizing Human Well-being with Artificial Intelligence and Autonomous Systems. In: *The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems IEEE Glob Initiat Ethical Considerations*. Vol. 13. 2016.

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

PERSPECTIVE ARTICLE

Artificial intelligence scribe: A new era in medical documentation

Khalid Nawab*

Department of Internal Medicine, Penn State Holy Spirit Medical Center, Camp Hill, Pennsylvania, United States of America

Abstract

The high workloads involved in clinical documentation represent one of the major factors contributing to the significant escalation of clinician burnout. The emergence of artificial intelligence (AI) has provided new avenues for relieving this burden by automating certain tasks like clinical documentation through the generation of clinical notes from a transcript of a clinical encounter. The advances in large language models (LLMs) have led to the emergence of such startups, but they come with their own set of challenges, predominantly surrounding the concerns of documentation accuracy, completeness, and data security. These can be addressed with a multi-faceted approach which could include fine-tuning the currently available models; using domain-specific models and in-house AI systems to ensure data security; and involving smaller LLMs and clinicians in the development and implementation of such systems. We can imagine a future where these systems are deeply incorporated into electronic health records, providing not only automated clinical documentation but also improving Clinical Decision Support systems, research, and patient communication.

Keywords: Artificial intelligence; Large language models; Clinical documentation; Automation; Clinician burnout

***Corresponding author:**
Khalid Nawab
(knawab@pennstatehealth.psu.edu)

Citation: Nawab K. Artificial intelligence scribe: A new era in medical documentation. *Artif Intell Health*. 2024;1(4):12-15. doi: 10.36922/aih.3103

Received: March 6, 2024

Accepted: June 19, 2024

Published Online: September 27, 2024

1. Introduction

The American Medical Association reports that in the United States of America, physician burnout is an epidemic with about 63% of physicians reporting signs of burnout at least once per week.[1] Clinical documentation using electronic health record (EHR) is perceived as a significant contributor to clinicians' burnout mostly due to poor usability and excessive time spent on EHRs.[2]

Artificial intelligence (AI) has emerged as a potential solution to various tasks including documentation in healthcare. The idea can be traced back to 2017, with "DeepScribe" being one of the earliest companies offering such a service.[3] However, increased adoption likely happened after the attention was drawn to AI by ChatGPT, a publicly available online application that is optimized for human-like conversation.[4] Access to such powerful models through an application programming interface (API) opened new venues to easily incorporate natural language processing and AI in healthcare. An AI-based scribe application, incorporating speech-to-text transcription, and then using that to generate a clinical summary or other forms of notes, sounds

like a low-hanging fruit, that would be easy to sell to the burnt-out clinician with the promise of alleviating some of the burden of clinical documentation.[5] This has led to an explosion of startups offering such applications, with additional features such as recommendation of the International Classification of Disease codes, patient instructions in simple language, and even some degree of clinical decision-making.

2. The challenges

The AI scribe offers a potential solution to a problem that seemed impossible. However, this rapid adoption has not been devoid of challenges. Current popular large language models (LLM) like those that power ChatGPT and Google's Gemini are very good at general tasks, but their performance is suboptimal in domain-specific tasks. Thus, despite early excitement and some good feedback, it was soon realized that the level of diligence and accuracy of these models needed in clinical documentation may be out of reach. Some of the challenges are as follows:

2.1. Hallucinations and unfaithfulness

LLMs are known to hallucinate, meaning that they can "make up" information that may not be accurate.[6,7] This is because they have been trained on a large amount of textual data, the model tries to "fill in the gaps" with generated text based on its dataset. This can be very helpful in some tasks where accuracy is not a big concern; however, in healthcare, this poses a significant risk of introducing inaccuracies in clinical documentation which may compromise patient care.

2.2. Omission of information

The transcript of a clinical encounter may contain information that is not clinically relevant, such as small talk between the patient and the clinician. The LLM may decide to include that information, or conversely, decide not to include information that is clinically relevant, generating a note deficient in clinical information.

2.3. Note formatting inconsistencies

Even though there are generally accepted formats for clinical notes, each clinician has their own unique style of note-taking. Some may prefer to document problem-wise, while others may like it system-wise; some like their notes in a descriptive format while others in bullets. The LLMs can be prompted to draft a note in a certain format; however, their response is not always consistent, potentially leading to frustration for clinicians who expect their notes to be laid out in their preferred format.

2.4. Context window limitations

LLMs have a context window, which means they will take into consideration a certain amount of input textual data to craft a response for the user. If the length of the input data exceeds the context window, some information will likely be missed. In the context of AI scribes, if the encounter goes on for too long and there is a large amount of text in the input transcript, it is possible that the LLM misses information because of the narrow context window, leading to incomplete documentation.

2.5. Data security/Health Insurance Portability and Accountability Act (HIPAA) compliance

Many of the AI scribe applications utilize third-party LLMs through APIs, requiring data to be passed on to external servers. This poses a data security risk, as the organization loses control of the security and privacy of the data once it leaves their systems. In addition, if the organization that owns the AI scribe application does not implement HIPAA-compliant technologies[8] for data transmission and storage, the confidentiality of patient data may be compromised.

3. The way forward

Even though AI scribes come with a unique set of challenges, their place in healthcare is undeniable. Therefore, a lot of work is being done to improve their performance. Some of the potential solutions are as follows:

3.1. Fine-tuning

The LLMs, like OpenAI's Generative Pre-trained Transformer, can be fine-tuned for a specific task with the right data. In this process, the model is provided with sample input data and the expected response. The model then learns from this data and modifies future output to match the desired output. Fine-tuning is relatively easy to implement and may improve performance.

3.2. Selective information extraction

One potential solution to improve accuracy could be labeling information in the transcript based on their relevance, then omitting information that is labeled as not clinically relevant and including relevant information. This could provide the model with data that is relevant and concise, reducing the amount of data provided as input, fitting it in the context as well as reducing computation time.

3.3. Domain-specific models

Models trained on curated medical data and designed specifically for the task of clinical note generation from

transcripts of clinical encounters will likely produce significantly improved accurate results. Some of these models are already available, for example, Alphabet's MedPalm.[9] Some companies are even offering end-to-end pipelines starting with speech-to-text transcription to note generation.[10]

3.4. Retrieval-augmented generation (RAG)

LLMs store data in their parameters. However, their ability to retrieve and present precise information remains limited, leading to subpar performance in knowledge-intensive tasks compared to more task-specific architectures. This can be overcome by providing the model access to "non-parametric" data, known as RAG. The combination of parametric information with explicit non-parametric information can lead to much more accurate output.[11] This, when applied to AI scribes, can potentially improve the quality of the generated note significantly.

3.5. Small LLMs

Another potential is the use of "tiny LLMs" or "small LLMs." These are LLMs with a smaller number of parameters. The idea is that LLMs contain large amounts of generic data that may add little value to a specific task. Therefore, the models are trained and fine-tuned on smaller amounts of more specific high-quality data to improve their performance while keeping their size and thus computational expense low. The performance of these smaller LLMs for text summarization has been shown to be poor compared to larger LLMs.[12] However, there is potential for improvement through various methodologies. For example, knowledge can be transferred from a larger LLM to a smaller LLM to achieve better performance through improving "reasoning" by the smaller model. This methodology showed that the smaller LLM can even outperform some of the larger LLMs for certain tasks.[13] In the context of AI scribes, this can be very beneficial. Not only can the output be improved, the financial burden involved with implementing in-house LLMs can be reduced significantly.

3.6. In-house AI solutions

To ensure true data security, complete control over data and customization, in-house AI solutions could be implemented. Training and implementing LLMs is a computationally heavy task, which necessitates a significant financial investment in the initial phase. However, this initial investment will pay off in the long term and may even prove more profitable by reducing physician burnout, improving efficiency, and ensuring data security. It will also offer seamless integration with in-house systems, reducing technical difficulties.

3.7. Involvement of clinicians in co-design and implementation

Clinicians' input in the design and implementation of any new system in healthcare is crucial for its success. As the primary users of AI scribe applications, clinicians will have a deep understanding of their operational needs and workflow requirements. This will not only make such applications more effective but also improve adoption by clinicians. Similarly, ongoing education and training of clinicians along with the acquisition of their feedback, will ensure seamless integration as well as improvement of the application.

4. Conclusion and prospects

Clinical documentation is a crucial component of modern healthcare, but it also contributes significantly to the burnout of clinicians. AI-based technologies, like AI scribes provide a potential solution to alleviate this burden. Even though current technology is not without challenges, the prospects are promising. Anticipating widespread demand, EHR vendors will likely incorporate AI models in the core of their software, enabling not just AI scribes but also improving clinical decision support systems, automatic summarization of medical history, and research. Furthermore, AI also enables the advancements in patient-facing EHR systems that allow for documentation and provision of personalized information.

Acknowledgments

None.

Funding

None.

Conflict of interest

The author is the founder of a company that specializes in AI scribe services, which is relevant to the topic of this article. This has not influenced the content of the manuscript. No reference to the author's company is made, but it is declared for full transparency.

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.

References

1. *Physician Burnout: Warning Signs, Treatment & Recent News. American Medical Association*. Available from: https://www.ama-assn.org/topics/physician-burnout?page=4 [Last accessed on 2024 Mar 03].
2. Alobayli F, O'Connor S, Holloway A, Cresswell K. Electronic health record stress and burnout among clinicians in hospital settings: A systematic review. *Digit Health*. 2023;9:20552076231220241.

   doi: 10.1177/20552076231220241
3. *About Us - DeepScribe*. Available from: https://www.deepscribe.ai/about [Last accessed on 2024 Mar 04].
4. *Introducing ChatGPT*. Available from: https://openai.com/blog/chatgpt [Last accessed on 2024 Mar 04].
5. Coiera E, Kocaballi B, Halamaka J, Laranjo L. The digital scribe. *NPJ Digit Med*. 2018;1(1):1-5.

   doi: 10.1038/s41746-018-0066-9
6. Deng J, Lin Y. The benefits and challenges of ChatGPT: An overview. *Front Comput Intell Syst*. 2022;2(2):81-83.

   doi: 10.54097/FCIS.V2I2.4465
7. Borji A, Ai Q. *A Categorical Archive of ChatGPT Failures*; 2023. Available from: https://arxiv.org/abs/2302.03494v8 [Last accessed on 2024 Mar 04].
8. *Health Insurance Portability and Accountability Act of 1996*. ASPE. Available from: https://aspe.hhs.gov/reports/health-insurance-portability-accountability-act-1996 [Last accessed on 2024 May 23].
9. *Med-PaLM: A Medical Large Language*. Available from: https://sites.research.google/med-palm [Last accessed on 2024 Mar 04].
10. *Generate Clinical Notes with AI - AWS HealthScribe - AWS*. Available from: https://aws.amazon.com/healthscribe [Last accessed on 2024 Mar 04].
11. Lewis P, Perez E, Piktus A, *et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*; 2020. Available from: https://arxiv.org/abs/2005.11401 [Last accessed on 2024 May 23].
12. Fu XY, Laskar MTR, Khasanova E, Chen C, Shashi Bhushan TN. *Tiny Titans: Can Smaller Large Language Models Punch Above their Weight in the Real World for Meeting Summarization*?; 2024. Available from: https://arxiv.org/abs/2402.00841 [Last accessed on 2024 May 23].
13. Tian Y, Han Y, Chen X, Wang W, Chawla NV. *TinyLLM: Learning a Small Student from Multiple Large Language Models*; 2024. Available from: http://arxiv.org/abs/2402.04616 [Last accessed on 2024 May 23].

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Health-care app detection using optimized clustering

Ciza Thomas[1]* and Rendhir R. Prasad[2]

[1]School of Computer Science and Technology, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India

[2]Department of Information Technology, Government College of Engineering, Trivandrum, Kerala, India

Abstract

Medical health-care apps have become ubiquitous in today's world, enhancing health-care quality at affordable costs. The continuous development of new apps underscores their high acceptance and popularity. Machine learning techniques offer effective app identification owing to their high prediction accuracy, particularly with a training dataset of known apps. Although machine learning techniques provide high detection accuracy for known apps, they exhibit abysmal accuracy in detecting unknown and novel apps. This research proposes a novel approach to optimizing the K-means clustering algorithm for detecting zero-day apps. The proposed technique integrates a perceptron feed-forward neural network to determine the coordinates of the centroids of the clusters in K-means clustering. Experimental evaluations demonstrate the efficacy of the proposed approach in enhancing the performance of K-means clustering, providing improved detection for both known and unknown medical health-care apps. A total of 30 health-care apps was utilized in this evaluation. This research enhances the detection accuracy of medical health-care apps, particularly zero-day apps. The intercluster similarity of the benign class improved to 0.99, and that of the malicious class improved to 0.91, highlighting the improved classification of the apps. The major contribution of this work is achieving an intercluster similarity of 0.89 for detecting novel apps.

Keywords: Medical apps; Health-care apps; Machine learning; Artificial neural network; K-means clustering; Euclidean distance measure; Within-class similarity

***Corresponding author:**
Ciza Thomas
(cizathomas@karunya.edu)

Citation: Thomas C, Prasad RR. Health-care app detection using optimized clustering. *Artif Intell Health*. 2024;1(4):16-29. doi: 10.36922/aih.2585

Received: December 30, 2023

Accepted: June 13, 2024

Published Online: August 16, 2024

1. Introduction

Medical health-care apps have become ubiquitous in today's world, playing a pivotal role in enhancing health-care quality at an affordable cost. Health-care apps are software programs designed for mobile devices, including laptops, tablets, and smartphones, which often integrate with wearable devices. The functionalities of health-care apps are diverse, ranging from displaying vital information to assisting patients in monitoring their health. The constant influx of new apps, driven by high acceptance and popularity, is evident, including contributions from startups. However, the rapid development has raised concerns about inadequate security measures, potentially compromising privacy, authenticity, and data integrity for both patients and medical professionals. Addressing these security concerns is crucial to ensuring the safety and confidentiality of health-

related data, as well as the overall well-being of individuals relying on these apps. The increasing prevalence and importance of medical health-care apps in contemporary health-care systems is the driving force behind this study. While these apps offer significant benefits by improving health-care quality while remaining cost-effective, the rapid proliferation of new apps presents a challenge in accurately detecting and classifying them, particularly zero-day apps that are new and unseen. Conventional machine learning techniques exhibit high accuracy in detecting known apps but struggle with unknown ones. Hence, there is a pressing need to develop novel approaches capable of effectively identifying both known and unknown medical health-care apps. We aim to bridge this gap through an optimized clustering approach that leverages artificial neural networks (ANNs) to enhance detection accuracy, thereby improving the overall efficiency and reliability of health-care app detection systems.

Accurate detection of medical apps is crucial to maintaining the quality and continuity of systems, particularly in environments prioritizing bandwidth allocation for health-care purposes. App detection plays a pivotal role in enforcing network security policies. However, traditional methods that rely on port-based or payload-based techniques face challenges due to emerging obfuscation techniques such as port spoofing and encrypted traffic.

Identifying medical apps often involves well-known port-addressing techniques, such as matching traffic with ports registered by the Internet Assigned Numbers Authority (IANA). However, the growing prevalence of port spoofing poses a novel challenge, where attackers attempt to evade detection by utilizing exceptional port numbers for malicious network traffic. Relying solely on port values for app identification is insufficient. Therefore, a deeper examination of packet content using payload-based methods becomes necessary to understand the unique characteristics that define specific applications for effective detection. The conventional approach to app detection using well-known ports has limitations, especially when faced with sophisticated evasion techniques. Port spoofing, for instance, involves various tactics where attackers aim to bypass perimeter defenses using unconventional port numbers. Consequently, adopting a comprehensive approach that extends beyond port values becomes imperative.

Payload-based methods offer a deeper insight into packet content to identify the unique features that define a particular application. However, even this approach encounters challenges when dealing with encrypted packets. Encryption poses a hurdle to payload-based detection methods, hindering the ability to inspect packet contents effectively.

To overcome the limitations of traditional detection methods, a more nuanced and multifaceted approach is required. Combining port-based, payload-based, and potentially heuristic methods can enhance the accuracy and reliability of detecting medical apps within the dynamic realm of network security. By continuously adapting detection strategies to emerging threats, we can better safeguard the integrity and security of health-care systems that rely on medical apps. Flow-based approaches for measuring traffic statistics have been explored in existing literature to overcome common challenges encountered in conventional approaches. However, a significant challenge that is still unaddressed is the poor detection rate of new or zero-day apps.

Zero-day apps, which are previously unknown and emerging apps, pose a challenge for supervised machine learning algorithms that rely on training data from known apps. To address this issue, the proposed scheme introduces a semi-supervised method for categorizing novel apps into a distinct class that receives labels after analysis. The scheme comprises three key modules: known app detection, novel app segregation, and model updating with a new app class. The approach utilizes a hybrid detector that combines an ANN as a universal classifier with the K-means algorithm for enhanced detection.

The objective of this research is to enhance the accuracy of detecting medical health-care apps, with a particular focus on identifying zero-day apps. By integrating a novel approach that optimizes the K-means clustering algorithm with a perceptron feed-forward neural network, this study aims to improve the detection capabilities for both known and unknown medical health-care apps.

The innovation lies in optimizing K-means clustering using a neural network to fix centroids, facilitated by correlation-based feature selection to identify relevant clustering features. The proposed approach ensures superior results by determining the optimal value for K in K-means clustering and selecting appropriate choices for the number of input and output nodes in ANNs. Furthermore, the model includes a method to update the detector through retraining with a new class.

The experimental results demonstrate the superior performance of the proposed approach in detecting medical apps overall and specifically in identifying novel apps. This methodology offers an effective solution to the evolving landscape of app detection, particularly addressing the continuous emergence of previously unknown apps.

The subsequent sections of the paper are organized as follows: Section 2 delves into a comprehensive literature

review, offering insights into existing knowledge on the subject. Section 3 provides a theoretical background on machine learning techniques commonly applied in app detection, offering a foundational understanding of the methods employed in the field. Section 4 outlines and elucidates the proposed methodology, shedding light on the innovative approach introduced in this research. Section 5 meticulously examines and discusses the results obtained, providing a thorough analysis of the outcomes. Finally, Section 6 serves as the conclusive segment, summarizing the key findings and implications derived from the study.

2. Related works

In the literature, the detection of medical apps primarily relies on three prominent methods: the port-based approach, the payload-based approach, and the machine-learning approach. In the port-based approach, medical apps leverage well-known ports, as registered with IANA, for easy and conventional identification.[1] The original medical apps are registered with specific ports in IANA, and these well-known ports are advertised, facilitating proper and trivial identification. However, this approach has declined in popularity due to its susceptibility to inaccurate results caused by port obfuscation, particularly evident in cases where peer-to-peer (P2P) apps obfuscate their identity using well-known ports.[2]

When the limitations of port-based identification become apparent, the payload-based approach becomes crucial.[3] This method involves monitoring the entire packet content to identify unique and distinctive characteristics. While the payload-based approach exhibits high classification accuracy, it faces several challenges:

(i) Deep inspection is time-consuming, which limits real-time detection in today's high-speed networks.[4]
(ii) The approach is ineffective with encrypted traffic, allowing P2P apps to escape detection.[5]
(iii) Privacy concerns exacerbate the challenges associated with this approach.[6]

Advanced prediction techniques and data analytics are increasingly employed to enhance productivity and efficiency in detecting medical apps, moving beyond conventional port-based and payload-based approaches. This is because high-speed network connectivity and big data transfers between sensors and monitoring systems demand the use of machine learning techniques and data analytics. These technologies contribute to cost reduction and minimize downtime. In the current literature, app detection leverages machine learning as a core technology to improve the detection performance of novel apps. Unlike signature-based detection algorithms, which struggle to identify novel or zero-day apps and often exhibit low detection rates with real-world data containing numerous zero-day apps. However, the high detection rate achieved with anomaly-based machine algorithms is often associated with a large false alarm rate, which greatly affects their usability and overall performance. The unsupervised machine learning algorithms are the best at detecting unseen and novel samples in the data. Hence, clustering methods are usually used to detect zero-day apps. The disadvantage of traditional clustering techniques such as K-means is the possibility of an incorrect initial choice of the number of clusters, which can prevent the convergence of the output clusters. In the K-means algorithm, deciding the number of clusters and determining the centroid for each cluster are vital and often challenging tasks, as they directly affect the quality of the resultant clusters. Ahmad and Dey[7] presented a modified description of cluster centers to overcome the limitation of handling only numeric data in the K-means algorithm, thereby enhancing cluster characterization. The intended results were to overcome the limitation of K-means in dealing with numeric data, whereby a modified description of the cluster center was presented. Another approach using fuzzy c-means has been proposed by Bezdek *et al.*[8] The clustering results obtained were integrated into a judgment matrix, which was then iteratively partitioned to identify the desired cluster number and the result. Zhou *et al.*[9] proposed a modified neural network backpropagation algorithm to improve detection rates, particularly in cases where there is an imbalance in the data, with the class of interest being a minority class. Anand *et al.*[10] modified the placement of the clustering class to overcome the class imbalance. Their modified backpropagation algorithm accelerated the convergence of the neural network. Kumar *et al.*[11] proposed the under-sampled K-means technique, effectively removing noisy and weak instances from large volumes of the majority class. In the work of Wu,[12] clusters were seen to be uniform in size despite variations in input data sizes.

3. Theoretical background

3.1. Need for medical apps

Before delving into the key factors contributing to the essential nature of health-care apps, it is crucial to explore noteworthy statistics and facts that underscore the industry's growth trajectory. According to Statistica, the health-care sector is projected to be one of the top revenue contributors, with estimates suggesting it will increase from $25.39 billion in 2017 to $58.8 billion by 2020.[13] The report by Research 2 Guidance indicates that there are 3,25,000 health-care apps available worldwide, with Android leading the way forward on the mHealth platform. A recent

survey by Accenture Consultants revealed that downloads of health-care apps have doubled over the past couple of years. As of early 2020, the Apple App Store had 2.2 million apps available for download, whereas the Google Play Store had 2.8 million apps available. Interestingly, among them, there are more than 97,000 health and fitness apps available for download on mobile or tablet devices. In recent times, the health-care setting has undergone drastic positive changes due to the rapid rise of health-care apps on mobile devices. The adoption of mobile devices by health-care professionals (HCPs) has transformed many facets of clinical practice. Health-care apps have been the vital forerunner for surpassing and backing the condition of the health-care trade. Medical health-care apps provide many benefits for HCPs, such as increased access to point-of-care tools that enhance patient care and facilitate rapid and precise clinical decision-making. Today, numerous apps are available to assist HCPs with important tasks, such as accessing, retrieving, monitoring, and managing patient data; maintaining records; communicating with patients; facilitating clinical decision-making; and providing medical training and education. Thus, these mobile apps enable doctors, nurses, and other health-care workers to communicate easily with patients and access necessary data. It can also simplify coordination between departments, labs, and staff, irrespective of their physical location. Doctors can communicate directly with patients, make informed decisions on diagnoses and treatments, and promptly prescribe medications. Health-care apps contribute to improved health outcomes through customization, increased access to health-care, secure and streamlined clinical communication between providers and patients, cost reduction, time savings, 24/7 service availability, and improved hospital workflow management (Figure 1).

The present-day mobile health-care apps include interoperable platforms, secure bi-directional communication, and patient-provider interactivity based on accessed data. Some of these mobile apps are presented in Table 1.

3.2. Risk of using medical apps

While medical apps provide numerous benefits, their use involves significant risk factors, leading many medical practitioners to remain reluctant about adopting them. Mobile medical apps used in health-care settings must be accurate and reliable, as critical decisions by HCPs and patients are often based on the information from these apps. Several medical apps have been found to compromise patient safety, proving dangerous for clinical use. For instance, apps designed for opioid dosage conversion or melanoma detection have shown dangerously low accuracy,[14,15] whereas several other health-care apps fail to adhere to evidence-based guidelines.[16,17]

These risks underscore the need for increased regulation before medical apps are used in clinical practice.[18-20] For instance, the ArogyaSetu app was developed in India during the COVID-19 pandemic to spread awareness and connect people in India with essential COVID-19-related health services. The app collected user location data and cross-referenced it with the Indian Council of Medical Research database to alert users about close proximity to infected individuals. However, Indian security officials identified that hackers had developed a fake ArogyaSetu app to steal information, targeting Indian defense forces. These hackers impersonated the Indian government, sending emails containing malware to victims. The malware included bogus health advisories on coronavirus, which, when clicked, allowed hackers to access sensitive information such as passwords, credit card information, and location data from users' browsers without their knowledge. Thus, contact lists or any other sensitive information stored on the device could be accessed by hostile hackers.

People seeking information about diabetes and other conditions could be at risk of having their private information stolen and their privacy invaded by cyber

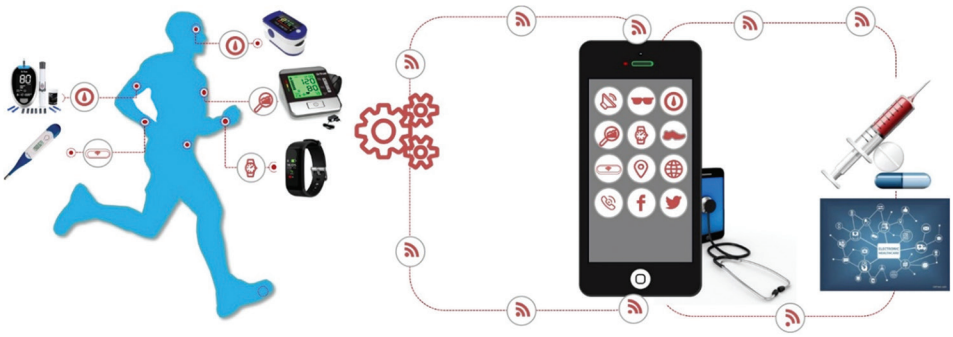


Figure 1. Picturization of mobile apps and their interactions

Table 1. Details of some health-care-related mobile apps[13]

Names	Details
AirStrip	Permits health-care coordination between several devices and numerous care settings
Arogya Setu app	Spreads awareness of COVID-19 by linking crucial COVID-19-associated health services with the Indian community
ITriage	Permits patients to get information on their health condition and provides them with proper guidance on treatment
CareAware Connect	Links patients in a mobile directory by administering medication with the help of bar codes
DSS Inc.	Offers both medical and organizational tools that automate billing and scheduling in emergency rooms and home health-care
MyChart Mobile	Allows patients to confirm appointments, pay the due health-care bills, and upload data generated from any wearable health device
Marbella	Used to indicate patient data collection external to the general rounds for nurses and caregivers within a health system
Ambulatory EHR	Health-care personnel can view the test results, schedule appointments, etc., by tapping on a patient's record
PatientKeeper	Permits health-care personnel to keep track of patient's health data such as allergies, test results, patient vitals, etc.
PatientTouch System	Brings clinical workflows, clinical communication, and coordinated care under one umbrella
Spok Mobile	Health-care personnel can prioritize and securely message members of the care team
22otters	Health-care personnel can dictate instructions such as medication information into the app and also set alerts for patients
AmWell	Connects physicians with patients through the Online Care Group.
BetterDoctor	Enables physicians to increase practices through their online presence and also supports patients in locating clinicians nearby
Blue Star	Analyzes the diabetes data entered by the patient and provides a self-management plan for the patient
CareConnect	Enables the parents to have an around-the-clock, face-to-face consultation with pediatricians for their child's health
CareZone	Manages one's own care by automatically updating medication information and relevant health news
Docphin	Records user preferences and sends specialty- and topic-specific journal articles, authors, and PDFs as they are released.
Doctor On Demand	Beyond medical and pediatric care, the app includes a model for 25/50-min psychology sessions and lactation consultations
Doximity	In-demand tools to make clinicians' lives easier by sourcing problems from the user community and developing solutions
FairCare	Encourages users to anonymously share the cost of care they have received to help others make decisions
Figure 1	The app may share uploaded photos with medical journals or other educational sites
FollowMyHealth Mobile	Automatically updates information such as blood pressure, changes in weight, and glucose readings
Heal	Heal's physicians can be ordered for house calls when patients are sick and looking for a checkup
Healow	Enables users to access data from all of their health-care providers in one place
HealthLoop	Enables automated check-ins that help monitor and guide patients as well as improve medication adherence
HealthTap	Enables users to connect with clinicians anywhere through video or text consultations through a pay-as-you-go plan
Human Dx	Clinicians using the app can access insights from colleagues and the medical community at large to apply knowledge
Isabel	Provider inputs information about a patient, and the app compiles a list of likely diagnoses to help make an informed decision
Jan Aushadhi Sugam	Ensuring the availability of quality generic medicines at affordable prices for all in India
LiveHealth Online	A telemedicine app that enables two-way chat between patient and physician
Medigram	Aims to give clinicians the power of fast, secure image sharing. Encrypted messaging enables providers to share images
Medscape	Allows physicians, nurses, medical students, and other clinicians to quickly look up medications and dosing guidelines
Medisafe	A "virtual pillbox," the app reminds users exactly when to take their prescription medications
MyChart	Available for patient download, enabling them to access their medical records by smartphone or tablet.
Nursing Central	Tools to help frontline healthcare workers do their jobs more efficiently and with greater ease
Pager	Enables users to page physicians and nurses for on-demand care wherever they may be
Patient IO	Enables physicians to program daily tasks for patients based on their treatment plans
PediaQ	Enables parents to request a nurse practitioner make a house call
PillPack	Manage user refills, including phone calls and fax follow-ups. The app includes a full schedule of medication intake
Pingmd	Helps parents communicate with their children's pediatricians and has since grown into a multi-use communication platform
Practo	Seeks to provide patients everywhere with the tools they need to find care and wellness

(Cont'd...)

Table 1. ***(Continued)***

Names	Details
referralMD	Aims to standardize referral network communication between primary care physicians and specialists
RevUp	A chronic pain management tool that enables patients to access, log, and monitor their health information for improved care outcomes
Sherpaa	Enables staff physicians to provide real-time medical advice to their clients' employees
SmartConcierge	Helps users understand health plans and benefit packages, schedule appointments, and provides 24/7 support from registered nurses
Teladoc	Resolves medical issues between patients and physicians via phone or video consults
Touch Surgery	Uses graphic 3-D images to give surgeon trainees an idea of what they will see during the actual procedure
Twine	Helps clinicians provide continuous care to chronically ill patients through their devices, anytime, anywhere
UpToDate	Helps clinicians make the best evidence-based, point-of-care decisions from the most current medical information
YouPlus	Provides daily lifestyle coaching to make healthy living easier with science-driven exercise programs, meal recommendations, etc.
ZocDoc	Helps patients schedule an appointment that will get them in front of a physician within 24 h

criminals. Axelle Apvrille, a principal security researcher at Fortinet, highlighted the details of such malicious Android apps during a presentation at the Virus Bulletin 2019 conference in London. Many free diabetes management apps, while seemingly helpful, require users to download additional apps that are loaded with adware to function as advertised. Meanwhile, another type of malicious app, posing as a diabetes advisor, tracks almost all the user activities, including the GPS location of the device, the IP address, and the other installed apps on the device, putting the privacy of the user at total risk. All these apps also bombard users, including vulnerable patients who rely on them, with persistent pop-up advertisements.

In addition, a concerning trend involves malicious apps claiming to predict users' life expectancy within minutes based on health-related questions while secretly transmitting these details to remote servers, raising significant privacy concerns. Such stolen medical records often end up for sale on dark web forums, which in turn results in financial gain for cybercriminals. The hackers create malicious health-related apps as they serve as an easy way to steal data, install malware, or both, affecting numerous individuals. This trend is particularly concerning as cybercriminals increasingly target individuals using health-care apps, a demographic that continues to grow. Many app developers lack formal medical training and do not involve clinicians in the development process. Consequently, they may be unaware of patient safety issues arising from inappropriate content or app functionality.[21-23]

Moreover, the exponential growth of medical apps has made it impossible to thoroughly assess each one.[24] Despite this, evidence suggests that even a small number of medical apps can pose a risk to patient safety, underscoring the necessity of developing robust detection models to help detect these apps. Increasing the quality and safety of medical apps through better standards and validation practices needs to be established to ensure the proper use and integration of these increasingly sophisticated tools into medical practice. Understanding and quantifying the risks associated with medical apps depends on two critical factors: (i) The frequency of events that cause damage, and (ii) the severity of the resulting damages. These risks include potential harm such as damage to a doctor's/hospital's reputation, privacy issues, or clinical decision errors, in increasing order of severity. The decision to use any health-care app depends on whether the risk associated with the app is less in comparison to its benefits.

3.3. Digital health apps during COVID-19

Along with the adverse effects of the pandemic, a concurrent "infodemic" has emerged, characterized by widespread misinformation circulating online about the coronavirus. Therefore, it is necessary to choose reliable apps. The Apple Corporation has adopted a cautious approach by cracking down on potentially malicious software in its app store, allowing apps only from recognized institutions such as governments, health organizations, or hospitals, excluding independent developers. Similarly, Google has also implemented proactive measures; Google Play launched a section dedicated to COVID-19 with a curated list of certified apps.[13]

4. Methods

The proposed methodology relies on the integration of machine learning techniques, specifically the ANN and the K-means clustering algorithm, to form an advanced clustering method for the detection of medical apps. The ANN, inspired by the structure and functioning of the human brain, serves as a universal classifier in the proposed

methodology. The ANN is employed for its ability to learn complex patterns and relationships within the dataset, making it an effective tool for classification tasks. In this context, the ANN acts as a fundamental element in the hybrid detector, contributing to the enhanced detection performance of the overall system. The K-means clustering algorithm is a widely used unsupervised machine learning technique employed for clustering similar data points. K-means clustering aims to partition the dataset into distinct clusters, with each cluster representing data points that share similarities. The neural network aids in fixing the centroids of each cluster within the K-means clustering, contributing to superior detection performance.

In this study, the input to the ANN is a feature vector derived from data representing medical health-care apps. The specific features used as input to the ANN depend largely on the characteristics of the apps under analysis. Key input features considered in this study include:

(i) App metadata: Information such as app name, description, category (e.g., medical, fitness, wellness), and developer information
(ii) User engagement metrics: Metrics such as app ratings, reviews, download counts, and active user counts
(iii) App functionality: Features provided by the app such as symptom tracking, medication reminders, and telemedicine services
(iv) Technical characteristics: Attributes such as app size, update frequency, and compatibility with different platforms.

As the ANN used in the study is integrated with the K-means clustering algorithm to determine the coordinates of the centroids, it is used as a component within the clustering process rather than for standalone classification.

The architecture of the ANN is a simple feedforward neural network with one hidden layer, where the input layer receives the feature vector representing the apps and the output layer provides coordinates or weights that influence the clustering process. The specific configuration of the neural network, including the number of layers, neurons per layer, activation functions, and training parameters, depends on the requirements of the clustering task and is determined through experimentation to optimize performance. The choice of the K-means clustering algorithm in this work for medical health-care application detection is motivated by several factors such as scalability, simplicity, speed, effectiveness for spherical clusters, and compatibility with integration with ANN, making it a suitable choice for the task at hand. K-means is known for its scalability and efficiency, making it suitable for handling large datasets with many data points, even though we work with small data in this work. In the context of medical health-care apps, where the dataset may contain a significant number of instances, K-means can efficiently handle the clustering task without excessive computational resources.

K-means is relatively simple to implement and understand compared to more complex clustering algorithms. This simplicity can make it an attractive choice, especially if the goal is to develop an approach that is straightforward to interpret. K-means is known for its computational speed, particularly for low-dimensional data, making it well-suited for real-time or near-real-time apps where quick processing and response times are important, such as in health-care settings where timely decision-making is crucial.

One of K-means' strengths lies in its effectiveness with spherical clusters. K-means performs well when the underlying clusters in the data are spherical or globular in shape. In many cases, medical health-care apps may exhibit clusters that are relatively well-separated and have spherical shapes in the feature space, making K-means an appropriate choice.

In addition, K-means is compatible with ANN integration with other machine learning techniques, such as ANNs, as described in the paper. This integration allows for leveraging the strengths of both approaches to enhance the performance of medical health-care application detection.

To address potential issues associated with K-means clustering, such as sensitivity to initial centroids, assumptions about cluster shapes, and determining the appropriate number of clusters (K), several strategies are employed in this work. Firstly, instead of relying on a single random initialization for the centroids, the algorithm is run multiple times with different initializations. By averaging the results or selecting the best clustering solution based on a predefined criterion such as the lowest within-cluster variance, the risk of converging to a suboptimal solution is mitigated. In addition, domain knowledge or prior information about the apps is leveraged to initialize the centroids in a more meaningful way. For example, hierarchical clustering or density-based clustering techniques can be used to identify initial cluster centers.

When K-means assumptions about cluster shapes or the number of clusters are unknown, an alternative clustering algorithm such as density-based spatial clustering of apps with noise (DBSCAN) is considered. This method is found to be more flexible in handling non-spherical clusters and automatically determining the number of clusters based on the data structure. Hierarchical clustering is also used to explore different levels of granularity in the clustering solution, allowing for more flexibility in determining the

number of clusters and identifying clusters of various shapes and sizes. However, the particular data used in this work are spherical, and hence, no improvement is seen in using alternative clustering algorithms. To address the challenges posed by outliers, the non-deterministic nature of K-means, and the potential issue of points near cluster boundaries being assigned to different clusters, several strategies are employed. Before clustering, DBSCAN, the outlier detection technique, is applied to identify and remove outliers from the dataset. This helps prevent outliers from unduly influencing the clustering results and improves the robustness of the algorithm. Furthermore, as already discussed, instead of relying on a single random initialization for the centroids, the algorithm can be run multiple times with different initializations. By averaging the results or selecting the best clustering solution based on a predefined criterion, the impact of the initial centroid choice can be mitigated. In addition, post-clustering and post-processing techniques such as cluster merging, splitting, or reassignment based on proximity or density can be applied to refine the clustering solution and address any misassignments near cluster boundaries. This can help improve the overall quality of the clustering results. However, these techniques were not used in our experiments, as we did not encounter such misassignments.

Effective management of hyperparameter tuning for ANNs used in conjunction with K-means clustering was crucial in this study. We utilized cross-validation techniques, such as k-fold cross-validation, to assess the performance of different hyperparameter configurations on multiple subsets of the data. This approach ensures that the hyperparameters selected (app name, description, category [e.g., medical, fitness, and wellness], developer information, etc.) generalize well to unseen data and mitigate the risk of overfitting.

The section also introduces the performance matrices employed to assess the effectiveness of the proposed method. The chosen performance metrics provide a quantitative measure of the system's efficiency, allowing for a comprehensive evaluation of its detection capabilities.

4.1. Theoretical formulation

The proposed methodology involves utilizing the ANN to classify medical apps and the K-means clustering algorithm to partition the dataset into clusters, with the neural network aiding in determining the centroids of the clusters. The performance of the method is evaluated using a set of performance metrics to assess its detection capabilities.

Let $X= x_1, x_2, x_n$ denote the dataset consisting of n data points, where each xi represents a feature vector describing a medical application.

The proposed methodology integrates two main components: The ANN and the K-means clustering algorithm.

The ANN is represented as a function ANN: $f_{ANN}: R^m \rightarrow R^k$, where m is the dimensionality of the feature space and k is the number of classes. The ANN learns complex patterns and relationships within dataset X to classify medical apps into different categories.

The K-means clustering algorithm partitions the dataset X into K clusters, $C= C_1, C_K$, where each cluster represents data points that share similarities.

Let $\mu = \mu_1, \mu_2, \mu_K$ denote the centroids of the clusters. The objective of K-means clustering is to minimize the within-cluster sum of squared distances, which can be formulated as Equation I:

$$\text{Minimise}_C, \mu \sum_{i=1}^{K} \sum_{x \epsilon C_i} x - \mu_i^2 \quad \text{(I)}$$

The neural network aids in fixing the centroids μ of each cluster within the K-means clustering process, contributing to superior detection performance.

To assess the effectiveness of the proposed method, various performance metrics were employed, including but not limited to accuracy, precision, recall, F1 score, specificity, and area under the receiver operating characteristics (ROC) curve (area under the ROC curve). These metrics provide a quantitative measure of the system's efficiency and enable a comprehensive evaluation of its detection capabilities.

The theoretical formulation presented here encompasses several key aspects related to the challenges and approaches to app detection using machine-learning techniques:

(i) Parametric versus non-parametric classifiers: Parametric classifiers, reliant on model building, are noted for their sluggishness, posing challenges for real-time app detection. Non-parametric classifiers, while requiring a set of training data for estimating app distribution, suffer from the drawback of necessitating substantial training data for effective model-building

(ii) Accuracy paradox: Machine learning techniques typically prioritize the detection of known classes, leading to satisfactory prediction accuracy for established apps. The tendency to concentrate on larger classes with known apps can result in poor prediction accuracies for novel or zero-day apps

(iii) Zero-day application clustering: Even in cases such as clustering, the focus would be placed on labeling

the available apps, which are larger in number. This is because it could greatly affect the accuracy of the prediction. The detection of zero-day apps is given the least priority, mainly because of this reason. Hence, the novel zero-day attacks are not considered in the case of clustering, and they are forced to get clustered in any of the known clusters depending on their similarity with any of the available clusters. However, in this work, we include an additional cluster to group all these zero-day attacks that do not belong to any of the existing clusters.

(iv) Dynamic classifiers versus thresholding: The detection systems that are normally used are the ones that classify using thresholding. Even though simple, this has the disadvantage that the detection systems fail to be adaptive with a fixed threshold.[25] A successful detection system needs to be contextual, adapting to changing conditions. To overcome the generality of the static thresholding classifiers, dynamic classifiers, which are based on ANN, are proposed in this work. Dynamic classifiers, or leveraging ANNs, are advocated for their adaptability to changing conditions, especially with a multitude of apps

(v) Perceptron feed-forward neural network in K-means clustering: The focus of this research is to determine the coordinates of the centroid of every cluster in the K-means clustering process and to analyze its effect on class imbalance. In determining the centroid in K-means clustering, we propose the use of a perceptron feed-forward neural network. As a supervised learning algorithm, this method is well-known for the efficient handling of large amounts of data. The approach has been proposed to minimize the mean square error.

The theoretical framework underscores the need for adaptive, context-aware detection systems, addressing issues related to known and unknown app classifications, and introducing innovative solutions, such as dynamic classifiers and enhanced K-means clustering with neural networks.

4.2. Performance matrices

The evaluation utilized two similarity measures: the Euclidean distance and the Manhattan distance. These measures were selected to assess their influence on the clustering outcomes across multiple iterations, the within-cluster sum of squared errors, and the overall model-building time. The computational complexity of K-means clustering is determined by three primary factors: the number of data points (n), the number of clusters (k), and the dimensionality of the data (d).

In each iteration of K-means, the algorithm assigns each data point to its nearest centroid and then updates the centroids based on the mean of the data points assigned to each cluster. The computational complexity of a single iteration can be broken down into two main steps: assignment and update. In the assignment step, each data point is assigned to the nearest centroid. The computational complexity of this step is O ($n \times k \times d$), as for each data point, we calculate its distance to each of the k centroids in d-dimensional space. In the update step, the centroids are updated by calculating the mean of the data points assigned to each cluster. The computational complexity of this step is O ($n \times d \times k$), and for each centroid, we calculate the mean of the d-dimensional data points assigned to that cluster.

Therefore, the overall computational complexity of K-means clustering is often given as O ($I \times n \times k \times d$), where I is the number of iterations required for convergence. Typically, the number of iterations is relatively small, and the algorithm converges quickly, especially if the data is well-clustered. It is only for large datasets or a large number of clusters that the computational complexity can become significant, which is not the case in this work. In addition, the initialization of centroids can also affect the computational complexity and convergence speed of the algorithm.

4.3. System model

In this work, a hybrid classifier with an ANN, which is a universal classifier, combined with the K-means clustering method is proposed to accurately detect the apps. The proposed classifier exploits the statistical variation of the distinguishing features of the different apps. The model is updated frequently by training with the zero-day apps belonging to that cluster. Hence, an updated model building happened in this case, which reduced the overall error of the app detection. The methodology consisted of two stages: the data preparation stage and the clustering stage. In the preparation stage, the data set containing 30 medical apps was trained by an ANN to generate k initial cluster centers (centroids) for the classes that are benign, malicious, and zero-day. In this research, the generated k centroids were used as the initial cluster centers for the K-means clustering. The architecture of the training and test phases of the proposed method with ANN followed by K-means clustering is presented in Figures 2A and B.

The modification of the K-means clustering algorithm in determining centroid using ANN is given as a flowchart in Figure 3. The algorithm for calculating the initial cluster centers (centroid) of *n* objects using ANN is given in algorithm 1.

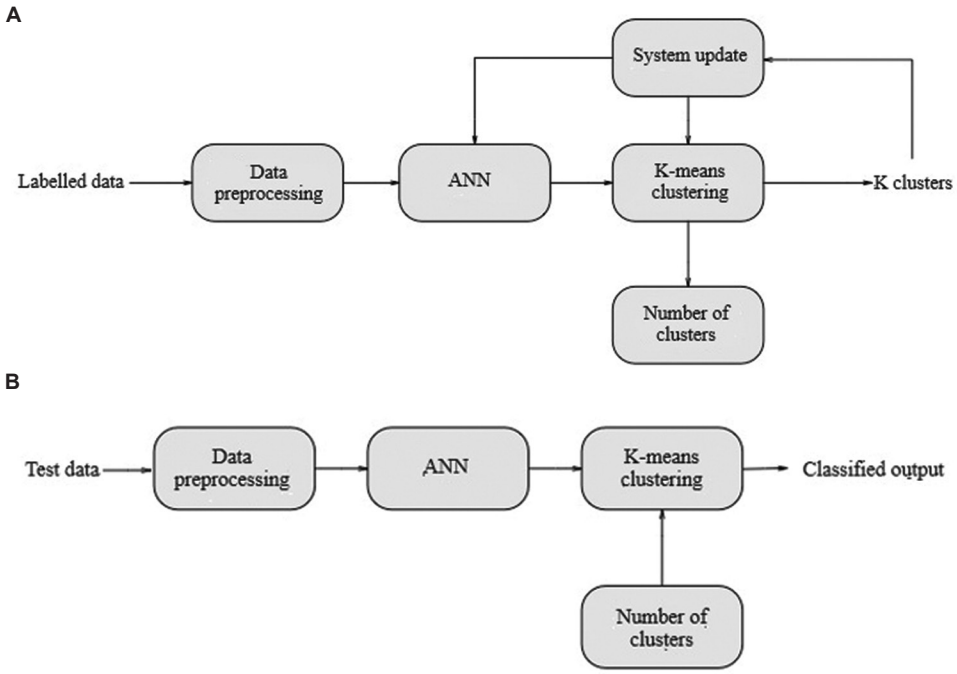


Figure 2. Proposed methodology for the (A) training and (B) test phase

Algorithm 1 Cluster centroids using ANN

1: Procedure ANN
2: *Initialize the bias and weights*
3: *Set the learning rate*
4: Set input conferring to dataset
5: *Set cluster output conferring to number of clusters from dataset*
6: **repeat**
7: **for** each training pair **do**
8: *Set activation of input units*
9: Compute response of output unit
10: Update bias and weights if an error occur for this data
11: **until** terminating condition is true
12: *Test terminating condition*
13: **if** weights are unchanged **then**
14: *stop*
15: **else**
16: *Continue step 6*

5. Results and discussion

Finding a dataset specifically focused on medical apps is very challenging. However, some datasets include information about app usage, reviews, ratings, or features related to medical apps within larger repositories or platforms. Potential sources for finding such datasets in the proposed work include platforms such as the Apple App Store and Google Play Store, which offer APIs or datasets containing information about various mobile apps, including medical apps. We can scrape this data or use APIs to retrieve information such as app names, descriptions, ratings, reviews, and categories. In addition, some of the publicly available APIs may offer access to data about medical apps or health-related services. For example, APIs provided by health-care organizations or platforms such as HealthKit might include information about app usage, health data integration, or user interactions with medical apps. Open data initiatives such as Kaggle host various datasets contributed by users or organizations. While we might not find a dedicated dataset for medical apps, we could find related datasets containing app usage or user behavior data that includes medical apps.

The modified clustering algorithm is trained using a training dataset. Semi-supervised classification uses a significant amount of labeled data together with unlabeled data for classification. The training dataset was created by considering 20 medical apps and malware and 10 unknown benign apps and malware samples. We used the platforms Weka and MATLAB for the whole training and validation procedures. With input data fed to the ANN, the number of iterations and nodes need to be specified during training. Outputs with the same node numbers were assumed to be in the same cluster, resulting in intracluster similarity being the maximum and intercluster similarity being the minimum. While increasing the number of nodes in the ANN can improve performance, it also adversely affects the time complexity. The maximum number of epochs in this study is 1000.

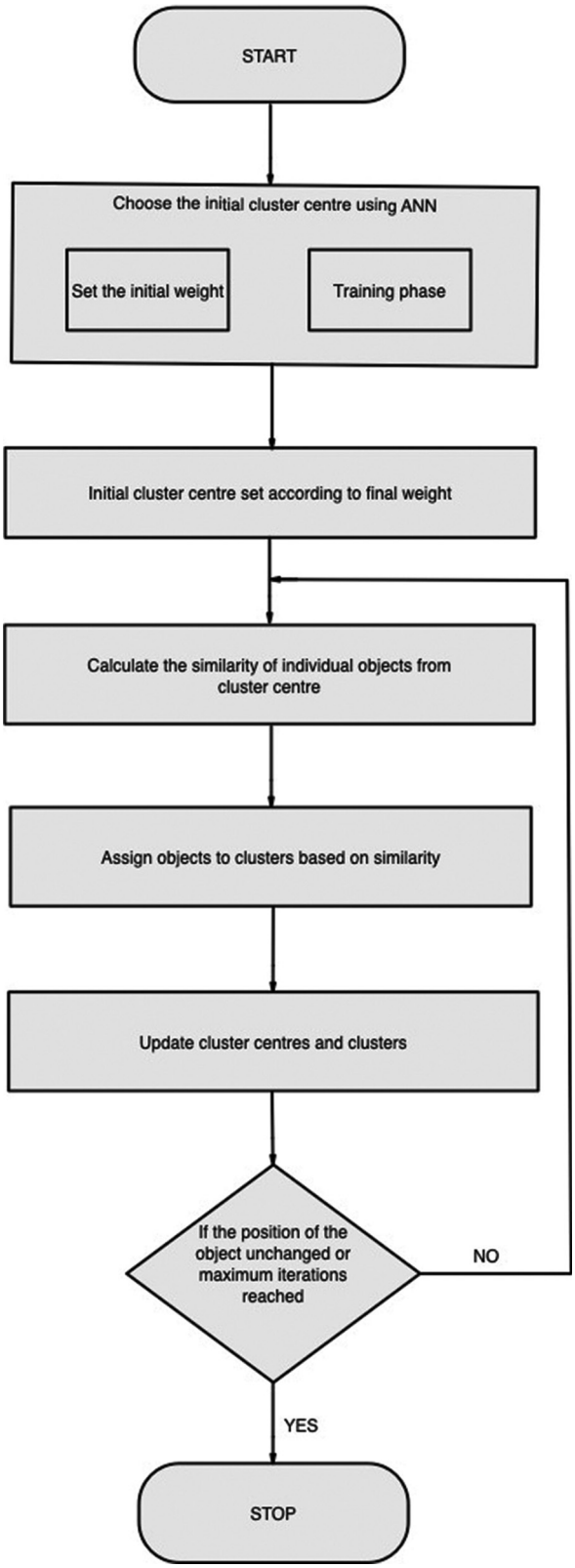


Figure 3. Flowchart showing modified K-means clustering with artificial neural network

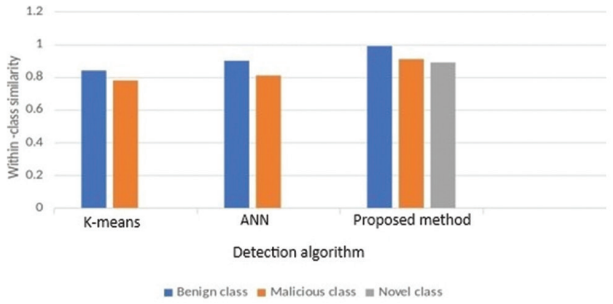


Figure 4. Intracluster similarity for various methods
Abbreviation: ANN: Artificial neural network.

Table 2. Performance comparison of advanced K-means with artificial neural network and K-means

Intra-cluster similarity of cluster types			
Method	**Benign class**	**Malicious class**	**Novel class**
K-means	0.84	0.78	0
Artificial neural network	0.9	0.81	0
Proposed method	0.99	0.91	0.89

5.1. Results

Table 2 and Figure 4 show the superior performance of the proposed algorithm with intracluster similarity of 0.99, 0.91, and 0.89 for the clusters benign, malicious, and zero-day, respectively. The intercluster similarity of the proposed algorithm is acceptably small compared to the individual techniques of K-means and ANN. By comparing the detection performance results with the ANN classifier, clustering, and advanced clustering, the number of errors in the datasets has been reduced using the optimization model of K-means clustering. The advanced K-means algorithm performed better than the individual ANN classifier or the K-means clustering, showing a minimum error rate.

5.2. Discussion

While our initial evaluation primarily focused on accuracy, we recognize the importance of assessing additional metrics such as precision, recall, and F1-score to provide a thorough evaluation of detection performance. These metrics are essential for understanding the nuances of classification performance, especially in the context of imbalanced datasets. These metrics offer a more comprehensive assessment of our proposed methodology.

Furthermore, the choice of a shallow ANN in our study is deliberate due to the specific characteristics of our dataset and application. Shallow ANNs are

computationally less intensive and reduce the risk of overfitting, which is crucial given our limited sample size. However, we acknowledge that more complex models could potentially yield better performance and will consider this in future work.

5.3. Limitations

One significant limitation of our study is the small sample size of 20 medical apps and malware samples, along with 10 unknown benign apps and malware samples. This small sample size limits the generalizability of our findings. In future work, we plan to expand our dataset to include a more diverse and larger set of samples to validate our results further.

Another limitation is the lack of direct comparison with existing state-of-the-art methods for app detection. While we conducted a comprehensive literature review to understand the current landscape, direct empirical comparisons are necessary to validate the effectiveness of our approach rigorously. We aim to address this in future studies by benchmarking our method against established techniques using larger and more diverse datasets.

While our proposed method demonstrates promising results in terms of intracluster similarity and error reduction, further research with larger datasets and more complex models is needed to fully validate its effectiveness and generalizability.

6. Conclusion

The paper has successfully addressed the challenge of detecting zero-day health-care apps, a prevalent issue where conventional app detection techniques struggle with misclassifying zero-day traffic into predefined known classes. Our approach proposes a scheme that can identify zero-day apps while accurately classifying those belonging to predefined application classes. The proposed scheme encompasses three crucial modules: unknown discovery, app classification, and system update. By leveraging ANNs to determine centroids in K-means clustering, our study reveals that the hybrid model of K-means clustering using ANN enhances app detection, particularly for zero-day apps. We highlight the impact of unknown apps on the classification accuracy of supervised methods, validating the effectiveness of correlation-based feature selection for clustering essential features. With a focus on unknown discovery and (N + 1) class classification, the proposed model efficiently identifies zero-day traffic and undergoes frequent updates through training with zero-day apps within the respective cluster. This continuous model refinement contributes to an overall reduction in app detection errors. By optimizing K in K-means and the number of nodes in ANN, substantial improvements in results are attainable.

To further enhance the effectiveness of health-care app detection, future work should explore several avenues. Real-time data capture could improve classification accuracy in dynamic environments. Investigating advanced feature selection algorithms holds promise for achieving greater accuracy in app detection. In addition, the incorporation of weighted sampling techniques in training flows may provide more representative and effective models. The obtained results suggest that substantial improvements in the performance of health-care app detection are feasible. Future studies should focus on finding an optimal method for determining the number of clusters, a critical aspect of refining the proposed scheme. In addition, extending the study to encompass diverse health-care scenarios and data sources would enhance the robustness and applicability of the proposed detection model. Collecting more data is essential to strengthening the conclusions and reliability of the proposed methods.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Ciza Thomas
Formal Analysis: Ciza Thomas
Investigation: Ciza Thomas
Methodology: All authors
Writing – original draft: All authors
Writing – review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Not applicable.

References

1. Moore AW, Papagiannaki K. Toward the accurate identification of network applications. *Lect Notes Comput Sci*. 2005;3431:41-54.

   doi: 10.1007/978-3-540-31966-5_4

2. Zink T, Maier M. *Analysis and Efficient Classification of P2P File Sharing Traffic. Technical Report KN-2010-DiSy-02*. Germany: University of Konstanz; 2010.

   doi: 10.5281/zenodo.1234567

3. Finsterbusch M, Richter C, Rocha E, Müller JA, Hanssgen K. A survey of payload-based traffic classification approaches. *IEEE Commun Surv Tuto*. 2014;16(2):1135-1156.

   doi: 10.1109/SURV.2013.100613.00161

4. Smith R, Estan C, Jha S, Kong S. Deflating the Big Bang: Fast and Scalable Deep Packet Inspection with Extended Finite Automata. In: *Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication*; 2008. p. 207-218.

   doi: 10.1145/1402958.1402983

5. Esteves AF, Inácio PR, Pereira M, *et al*. On-Line Detection of Encrypted Traffic Generated by Mesh-Based Peer-to-Peer Live Streaming Applications: The case of Goalbit. In: *2011 IEEE 10th International Symposium on Network Computing and Applications*. United States: IEEE; 2011. p. 223-228.

   doi: 10.1109/nca.2011.38.

6. Parekh JJ, Wang K, Stolfo SJ. Privacy-Preserving Payload-Based Correlation for Accurate Malicious Traffic Detection. In: *Proceedings of the 2006 SIGCOMM Workshop on Large-Scale Attack Defense*; 2006. p. 99-106.

   doi: 10.1145/1162666.1162679

7. Ahmad A, Dey L. A k-mean clustering algorithm for mixed numeric and categorical data. *Data Knowl Eng*. 2007;63(2):503-527.

   doi: 10.1016/j.datak.2007.03.016

8. Bezdek J, Ehrlich R. Numerical methods for fuzzy clustering. *Comput Geosci*. 1984;10:191-203.

   doi: 10.1016/0098-3004(84)90020-X

9. Zhou ZH, Liu XY. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Trans Knowl Data Eng*. 2005;18(1):63-77.

   doi: 10.1109/TKDE.2005.14

10. Anand R, Mehrotra KG, Mohan CK, Ranka S. An improved algorithm for neural network classification of imbalanced training sets. *IEEE Trans Neural Netw*. 1993;4(6):962-969.

    doi: 10.1109/72.258821

11. Kumar NS, Rao KN, Govardhan A, Reddy KS, Mahmood AM. Undersampled k-means approach for handling imbalanced distributed data. *Prog Artif Intell*. 2014;3(1):29-38.

    doi: 10.1007/s13748-014-0040-7

12. Wu J. The uniform effect of k-means clustering. In: *Advances in K-means Clustering*. Germany: Springer; 2012. p. 17-35.

    doi: 10.1007/978-3-642-31559-4_2

13. Mindinventory. *Mobile App for Healthcare*. Available from: https://www.mindinventory.com/blog/advantages-mobile-app-for-healthcare [Last accessed on 2022 Dec 20].

14. Haffey F, Brady RR, Maxwell S. A comparison of the reliability of smartphone apps for opioid conversion. *Drug Saf*. 2013;36(2):111-117.

    doi: 10.1007/s40264-013-0021-5

15. Wolf JA, Moreau JF, Akilov O, *et al*. Diagnostic inaccuracy of smartphone applications for melanoma detection. *JAMA Dermatol*. 2013;149(4):422-426.

    doi: 10.1001/jamadermatol.2013.1241

16. Rosser BA, Eccleston C. Smartphone applications for pain management. *J Telemed Telecare*. 2011;17(6):308-312.

    doi: 10.1258/jtt.2011.110202

17. Ferrero NA, Morrell DS, Burkhart CN. Skin scan: A demonstration of the need for FDA regulation of medical apps on iPhone. *J Am Acad Dermatol*. 2013;68(3):515-516.

    doi: 10.1016/j.jaad.2012.08.049

18. Misra S, Lewis TL, Aungst TD. Medical application use and the need for further research and assessment for clinical practice: Creation and integration of standards for best practice to alleviate poor application design. *JAMA Dermatol*. 2013;149(6):661-662.

    doi: 10.1001/jamadermatol.2013.351

19. Buijink AW, Visser BJ, Marshall L. Medical apps for smartphones: Lack of evidence undermines quality and safety. *BMJ Evid Based Med*. 2013;18(3):90-92.

    doi: 10.1136/eb-2013-101375

20. McCartney M. How do we know whether medical apps work? *BMJ*. 2013;346:f1811.

    doi: 10.1136/bmj.f1190

21. Hamilton A, Brady RW. Medical professional involvement in smartphone "apps" in dermatology. *Br J Dermatol*. 2012;167(1):220-221.

    doi: 10.1111/j.1365-2133.2012.11081.x

22. Huckvale K, Car M, Morrison C, Car J. Apps for asthma self-management: A systematic assessment of content and tools. *BMC Med*. 2012;10:144.

doi: 10.1186/1741-7015-10-144

23. Rodrigues M, Visvanathan A, Murchison J, Brady R. Radiology smartphone applications; current provision and cautions. *Insights Imaging*. 2013;4(5):555-562.

    doi: 10.1007/s13244-013-0275-2

24. Van Velsen L, Beaujean DJ, Van Gemert-Pijnen JE. Why mobile health app overload drives us crazy, and how to restore the sanity. *BMC Med Inform Decis Mak*. 2013; 13:23.

    doi: 10.1186/1472-6947-13-1

25. David J, Thomas C. Efficient DDoS flood attack detection using dynamic thresholding on flow-based network traffic. *Comput Secur*. 2019;82:284-295.

    doi: 10.1016/j.cose.2018.11.00

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Deep learning-powered segmentation and classification of diabetic retinopathy for enhanced diagnostic precision

Manoj Saligrama Harisha, Arya Arun Bhosale*, and M. Narender

Department of Computer Science and Engineering, The National Institute of Engineering, Mysuru, Karnataka, India

Abstract

This study addresses the critical challenge of diabetic retinopathy (DR), a severe complication of diabetes that potentially leads to blindness. We introduce a novel approach to DR detection using transfer learning, leveraging a single fundus photograph to automatically identify the disease's stage. DR progresses through four stages, posing challenges for early detection, with existing methods often inefficient and prone to disagreements among clinicians. The proposed approach demonstrated in the APTOS 2019 Blindness Detection Competition employs convolutional neural networks (CNNs) and achieved a high quadratic weighted kappa score of 0.92546, highlighting its effectiveness in automatic DR detection and emphasizing the need for timely intervention. This paper first reviews related work, spanning classical computer vision methods to deep learning approaches, with a focus on CNNs. Transfer learning with CNN architectures is explored, showcasing promising results from various studies. Identifying two critical gaps in existing literature, the research emphasizes the need for comprehensive exploration into integrating pre-trained large language models (LLMs) with segmented image inputs for generating test/treatment recommendations. In addition, understanding the dynamic interactions among integrated components, including lesion segmentations, disease classification, and LLMs within web applications, remains essential. The objectives of the study include developing a comprehensive DR detection methodology, exploring and implementing model integration, evaluating performance through competition ranking, contributing significantly to DR detection methodologies, and identifying research gaps. The study encompasses revolutionizing DR detection by integrating cutting-edge technologies, focusing on transfer learning and various model integrations within web applications. The methodology covers data pre-processing, augmentation, segmentation using U-Net neural network architecture, and a detailed training process. The U-Net model demonstrates efficient segmentation of retinal structures with high accuracy and an impressive frames-per-second rate. The results highlight the model's effectiveness in segmenting blood vessels, hard exudates, soft exudates, hemorrhages, microaneurysms, and the optical disc, with high Jaccard, F1, recall, precision, and accuracy scores. These findings underscore the model's potential to enhance diagnostic capabilities in retinal pathology assessment, promising improved patient outcomes through timely diagnosis and intervention in combating DR.

Keywords: Diabetic retinopathy; Deep learning; Segmentation; Transfer learning; Convolutional neural networks; Lesion segmentations; Disease classification; U-Net architecture

***Corresponding author:**
Arya Arun Bhosale
(2020cs_aryaarunbhosale_a@nie.ac.in)

Citation: Harisha MS, Bhosale AA, Narender M. Deep learning-powered segmentation and classification of diabetic retinopathy for enhanced diagnostic precision. *Artif Intell Health*. 2024;1(4):30-42. doi:10.36922/aih.2783

Received: January 19, 2024

Accepted: April 1, 2024

Published Online: September 6, 2024

1. Introduction

Diabetic retinopathy (DR) is a severe complication of diabetes, posing a threat of blindness by damaging the delicate blood vessels in the retina. This condition progresses through four distinct stages: mild non-proliferative retinopathy, moderate non-proliferative retinopathy, severe non-proliferative retinopathy, and proliferative DR. Each stage presents unique characteristics, complicating the diagnostic process, especially in the initial stage where warning signs are absent.[1]

The potential to reduce new cases of DR by a substantial 56% through timely treatment and monitoring emphasizes the gravity of the situation. However, accurately identifying the disease's early stages remains a challenging task for clinicians, even those well trained in the field. Manual examination of diagnostic fundus images for early-stage detection is intricate, and the existing diagnostic methods are plagued by inefficiencies, resulting in disagreements among ophthalmologists and the provision of inaccurate ground-truth data for research purposes. In response to these challenges, various algorithms have emerged to improve DR detection. Initially, these algorithms were grounded in classical computer vision approaches. However, recent years have witnessed the rise of deep learning, with convolutional neural networks (CNNs) demonstrating their prowess in tasks such as classification and object detection, including the diagnosis of DR.[2]

This research paper presents a novel approach to addressing the complexities of detecting DR. The proposed method, which employs transfer learning, leverages a single fundus photograph to automatically identify the stage of DR. Notably, the approach is designed to learn essential features from a dataset that is both limited and noisy, presenting itself as a valuable screening tool within automated solutions, as depicted in Hann *et al*.[3]

Highlighting the method's effectiveness, the proposed approach achieved a commendable ranking in the APTOS 2019 Blindness Detection Competition, underscoring its capability with a high quadratic weighted kappa score of 0.92546. This research aims to significantly advance DR detection methodologies, particularly in automated systems, addressing the critical need for early diagnosis and intervention in the fight against DR.[4]

DR is a significant concern in Singapore due to its high prevalence among individuals with diabetes mellitus, affecting approximately one in nine adults in the country. Globally, DR is a leading cause of vision loss and blindness, imposing a substantial burden on Singapore's health-care system and economy. With Singapore's aging population and rising rates of diabetes, the incidence of DR is expected to increase, emphasizing the importance of early detection and intervention. To address this challenge, Singapore has implemented various initiatives, including nationwide DR screening programs, enhanced diabetic management protocols, and the integration of telemedicine technologies for remote retinal imaging and diagnosis. Despite these efforts, barriers to accessing eye care services, especially among underserved populations, remain a concern. Continued investment in preventive measures, health-care infrastructure, and public awareness campaigns is crucial to effectively managing DR and mitigating its impact on vision health in Singapore.[5]

The stages of DR progress as follows:

(i) Mild non-proliferative retinopathy: This is the earliest stage of DR and is characterized by the occurrence of microaneurysms (MAs), which have a limited impact on blood vessels and cause minimal distortion.
(ii) Moderate non-proliferative retinopathy: Progression to this stage involves the loss of blood vessels' ability to transport blood due to increased distortion and swelling. As abnormalities in the blood vessels become more pronounced, the distortion and swelling hinder the normal transportation of blood, significantly impacting overall retinal health.
(iii) Severe non-proliferative retinopathy: This stage results in a depleted blood supply to the retina. Increased blockage of blood vessels exacerbates the condition, prompting the retina to stimulate the growth of new blood vessels in an attempt to compensate for the reduced supply.
(iv) Proliferative DR: This advanced stage is marked by the proliferation of new blood vessels. Growth features secreted by the retina activate the proliferation of new blood vessels. These vessels grow along the inside covering of the retina and extend into the vitreous gel, filling the eye.

1.1. Related work

Numerous research endeavors have focused on the challenge of early detection of DR. Initially, researchers explored classical computer vision and machine learning methods to develop viable solutions. For example, Priya and Aruna[6] proposed a computer vision-based approach using color fundus images for DR stage detection. Their methodology involved extracting features from raw images through image processing techniques, which were subsequently fed into a support vector machine for binary classification. They achieved performance results on a testing set of 250 images, with a sensitivity of 98%, a specificity of 96%, and an accuracy of 97.6%. In addition, researchers have also explored other models for multiclass classification. For example, employing principal component analysis on

images and applying decision trees, Naive Bayes, or k-NN,[7] yielded an accuracy of 73.4% and an F-measure of 68.4% on a dataset of 151 images with varying resolutions. These efforts illustrate ongoing advancements in leveraging machine learning for DR detection.

The rise of deep learning approaches has led to the development of various methods for applying CNNs for DR detection. Pratt *et al.*[8] developed a network with a CNN architecture and data augmentation capable of identifying intricate features related to classification tasks such as MAs, exudates, and hemorrhages (HEs) in the retina, providing automated diagnoses without user input. Their model achieved a sensitivity of 95% and an accuracy of 75% on a validation set of 5000 images. Other researchers have also contributed to CNN-based approaches.[9,10] Notably, Asiri *et al.*[11] conducted a comprehensive review of existing methods and datasets, highlighting their pros and cons while emphasizing the challenges in designing efficient and robust deep learning algorithms for diverse problems in DR diagnosis and suggesting directions for future research.

Moreover, researchers have explored transfer learning using CNN architectures. Hagos and Kant[12] attempted to train InceptionNetV3 for 5-class classification with pre-training on the ImageNet dataset, achieving an accuracy of 90.9%. Sarki *et al.*[13] trained ResNet50, Xception Nets, DenseNets, and VGG with ImageNet pre-training, obtaining the best accuracy of 81.3%. Both research teams utilized datasets provided by APTOS and Kaggle.

1.2. Problem statement

1.2.1. Datasets

The image data employed in this study was obtained from diverse datasets. The research encompasses three primary objectives: lesion segmentation, disease/image grading, and treatment recommendations.

1.2.2. Lesion segmentation

Lesion segmentation in the Indian Diabetic Retinopathy Image Dataset (IDRiD) uses fundus images captured by a retinal specialist at an eye clinic in Nanded, Maharashtra, India. From the 100 of examinations available, we have extracted 516 images to form our dataset, as shown in Figure 1A and B.[12]

Experts verified that all images are of adequate quality and clinically relevant, that no image is duplicated, and that a reasonable mixture of disease stratifications representative of DR and diabetic macular edema is present.[14] The fundus images were acquired using a Kowa VX-10 alpha digital fundus camera with a 50° field of view, centered near the macula, and a resolution of 4288 × 2848 pixels stored in jpg file format, with each image approximately 800 KB in size.[15]

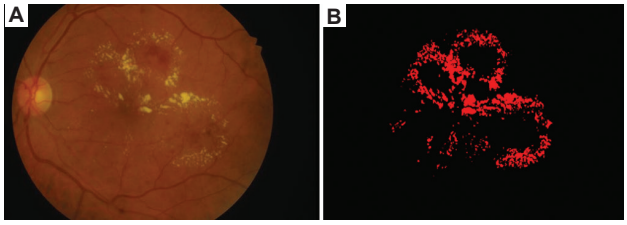


Figure 1. The original fundus image and its corresponding mask from the IDRiD. This dataset, consisting of 516 images, captures lesion segmentation in fundus images taken by a retinal specialist in Nanded, Maharashtra, India. (A) Original fundus image and (B) the corresponding mask. Images obtained from the IDRiD.
Abbreviation: IDRiD: Indian diabetic retinopathy image dataset.

For assessing lesion segmentation techniques related to DR, pixel-level annotated data includes binary masks for distinct abnormalities, including MAs, hard exudates (EXs), HEs, and soft exudates (SEs). The dataset comprises color fundus images in.jpg format, along with corresponding binary masks in.tiff files. The dataset contains 81 images with binary masks for MAs, 81 for EXs, 80 for HEs, and 40 for SEs, accommodating images with multiple lesions and enhancing robust research and performance evaluation of lesion segmentation techniques in DR.[36]

1.2.3. Image grading

Our research utilized image data sourced from multiple datasets, primarily focusing on an open dataset obtained from the Kaggle DR Detection Challenge 2015[16] for pre-training CNNs.[17] This dataset is widely recognized as the largest publicly available, comprising 35,126 fundus photographs capturing both the left and right eyes of American citizens. The images are labeled with stages of DR, ranging from no DR (label 0) to proliferative DR (label 4), as illustrated in Figure 2.[18]

In addition to the Kaggle dataset, we incorporated other smaller datasets, including the IDRiD,[19] as shown in Figure 3, from which we utilized 413 fundus photographs, and the methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology (MESSIDOR)[20] dataset, contributing 1200 fundus photographs. To ensure consistency, we used a version of the MESSIDOR dataset that had been relabeled to standard grading by a panel of ophthalmologists.[21]

Evaluation of our models was conducted on the Kaggle APTOS 2019 Blindness Detection[22] dataset, with access limited to the training portion. The full APTOS 2019 dataset comprises 18,590 fundus photographs divided into

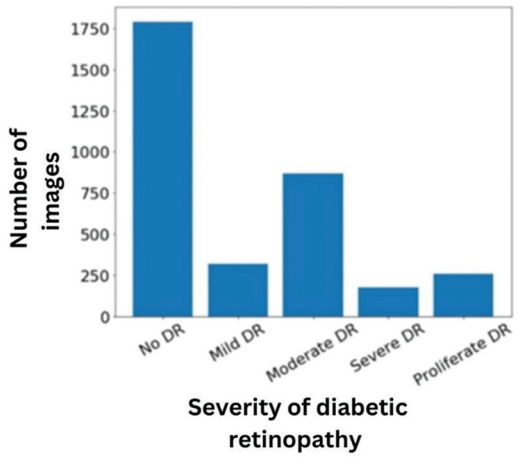


Figure 2. Class distribution in the APTOS 2019 dataset. Image generated using VS code
Abbreviation: DR: Diabetic retinopathy.

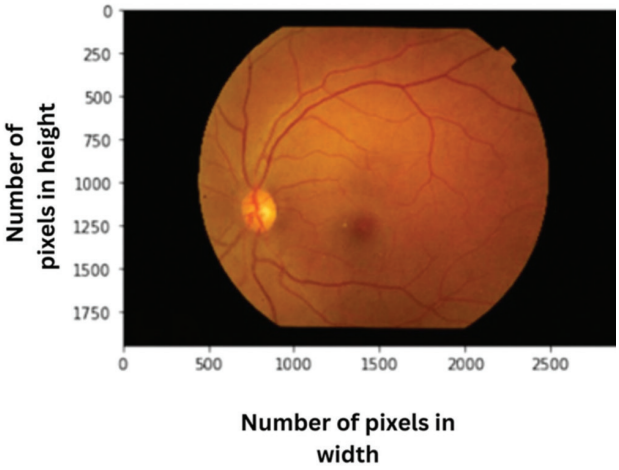


Figure 3. Sample of fundus photograph from the dataset. Image is a screenshot from VS code

3662 training, 1928 validation, and 13,000 testing images as organized by the Kaggle competition organizers. All datasets exhibit similar class distributions, as illustrated in Figure 1 for APTOS 2019. We maintained the original distribution of the datasets without any modifications, such as undersampling or oversampling. The smallest native size among all datasets is 640 × 480. A sample image from APTOS 2019 is presented in Figure 3.[23]

1.2.4. Large language models (LLMs)

In the dataset section, the generation of test/treatment recommendations involves the integration of pre-trained LLMs, with a comprehensive range of inputs derived from segmented images. These inputs encompass binary indicators for various lesions, including blood vessel segmentation, HE segmentation, EX segmentation, MA segmentation, optical disc segmentation, and SE segmentation. Each lesion is represented as either true (present) or false (absent) in the binary inputs. In addition, string inputs are generated from a classification or image grading model, offering insights into the DR stage classified as classes 0 through 4. The amalgamation of binary and string inputs forms a robust dataset that is processed by ChatGPT, a pre-trained LLM. ChatGPT interprets and synthesizes this diverse information to generate nuanced test/treatment recommendations, contributing to a sophisticated decision-support system that factors in both the detailed visual segmentation features and the clinical classifications of DR severity.[24]

1.3. Research gap

1.3.1. Treatment recommendations

While significant strides have been made in the realm of early DR detection, the existing research landscape reveals a distinct gap when comparing traditional methodologies with emerging approaches, particularly those involving pre-trained LLMs integrated with segmented image inputs for generating test/treatment recommendations. Classical methods, as evidenced by Priya and Aruna,[6] have predominantly employed computer vision and machine learning techniques for DR stage detection using color fundus images. Similarly, the advent of deep learning, particularly CNNs, has demonstrated promising results in intricate feature identification for classification tasks related to DR. Noteworthy works by Pratt *et al.*[8] have showcased the effectiveness of CNN architectures, achieving high sensitivity and accuracy in diagnosing retinal abnormalities.[18]

However, the existing body of literature primarily emphasizes isolated aspects such as lesion segmentation or DR classification, with a limited exploration of the synergies between visual segmentation features and clinical classifications within a decision-support system. This is evident in the literature reviewed, which often overlooks the potential intricacies arising from the amalgamation of binary indicators for various lesions and string inputs representing DR stages. The research gap lies in the absence of comprehensive investigations into the challenges and opportunities associated with the proposed methodology's integration of diverse data inputs. While previous studies have contributed valuable insights and benchmarking using classical methods and deep learning architectures, there is a need for focused research that bridges the gap between visual segmentation and clinical classifications to refine the efficacy of decision-support systems in DR management. Exploring this gap will contribute to advancing the field by providing a holistic

understanding of the challenges and opportunities presented by the integration of pre-trained language models with segmented image data.[25]

1.4. Multi-model integration

While the integration of various models, encompassing lesion segmentations (blood vessel, HE, EX, MA, optical disc, and SE), disease classification/image grading, and an LLM for test/treatment recommendations represents a noteworthy advancement in DR research, there exists a research gap in understanding the dynamic interactions and synergies among these integrated components within the context of a web application. The current literature often focuses on individual models or components separately, providing limited insights into the intricacies and challenges encountered when these models collaborate in real time. The integration of lesion segmentations, disease classification, and LLMs in a web application suggests a complex interplay of data flow and feedback mechanisms. Addressing this research gap is crucial for comprehensively understanding how these models collectively enhance accuracy and provide more accurate inputs. Exploring the dynamics of multi-model integration in a web environment will contribute to advancing the field by providing insights into the real-time interactions, potential bottlenecks, and opportunities for optimizing the collaborative functionality of diverse models within a unified interface.[26]

1.5. Objectives

This research paper aims to address the challenges associated with the detection of DR, a severe complication of diabetes leading to potential blindness. The primary objectives include:

i. Comprehensive DR detection methodology: This involves creating a novel approach that leverages transfer learning to automatically detect the stage of DR using a single fundus photograph.
ii. Integration of various models: This phase includes exploring and implementing the integration of diverse models, including lesion segmentations (blood vessel, HE, EX, MA, optical disc, and SE), disease classification/image grading, and LLM for test/treatment recommendations. The emphasis is placed on understanding the dynamic interactions and synergies among these integrated components within a web application context.
iii. Performance evaluation: The effectiveness of the proposed approach will be assessed by achieving a commendable ranking in the APTOS 2019 Blindness Detection Competition, demonstrating its capability with a high quadratic weighted kappa score of 0.92546.
iv. Contribution to detection methodologies: This research aims to significantly advance DR detection methodologies, particularly in the context of automated systems. It addresses the critical need for early diagnosis and intervention in the fight against DR.
v. Gap identification and exploration: This research identifies and explores research gaps in existing methodologies, specifically focusing on the integration of pre-trained LLMs for generating test/treatment recommendations and the dynamic interactions among integrated models in a web application. This will provide insights into the challenges and opportunities associated with these approaches. The overarching goal is to enhance the precision, personalization, and efficiency of DR detection, ultimately contributing to improved patient outcomes through timely diagnosis and intervention.

1.6. Scope

This research paper extends its scope to revolutionize DR detection methodologies by integrating cutting-edge technologies. By focusing on transfer learning, the proposed approach aims to overcome the limitations of traditional diagnostic methods, providing an automated and efficient solution for early-stage DR detection using single fundus photographs. The integration of various models, including lesion segmentations, disease classification, and LLMs, within a web application further widens the scope. This integration not only enhances accuracy but also addresses real-time challenges, offering a holistic and dynamic decision-support system. The paper's scope encompasses an in-depth exploration of research gaps in existing methodologies, emphasizing the need for comprehensive investigations into the challenges and opportunities associated with the proposed integration of diverse data inputs. By achieving a commendable ranking in the APTOS 2019 Blindness Detection Competition, the proposed methodology's effectiveness is demonstrated, contributing to the advancement of DR detection methodologies. The outcomes of this research hold promise for the wider field of medical imaging and automated diagnostics, potentially influencing the development of more precise and personalized solutions for various medical conditions.

2. Methods

2.1. Data pre-processing

The data pre-processing stage incorporates a custom PyTorch dataset class, namely "DriveDataset," tailored for handling the DRIVE dataset. This class serves as a crucial bridge between raw data and the U-Net model, streamlining the integration process. The "init" method initializes the dataset by storing the paths to the fundus images and their

corresponding masks, along with calculating the total number of samples. The "getitem" method is responsible for reading and pre-processing each sample, where the fundus image undergoes normalization and transposition to align its dimensions appropriately for the subsequent U-Net input.[18] Simultaneously, the binary mask is read and expanded to accommodate the model's requirements. Both the pre-processed image and mask are converted to PyTorch tensors before being returned as a tuple.[27]

This dataset class provides a seamless interface for interacting with the DRIVE dataset, offering a standardized and efficient means of loading and preparing data for training and evaluation. The "len" method ensures that the total number of samples can be easily accessed, facilitating iterative processes during model training and validation.[28] Overall, the data pre-processing workflow is encapsulated within this dataset class, contributing to the robustness and adaptability of the U-Net architecture for semantic segmentation tasks on retinal fundus images.[29]

2.2. Data augmentation

The data augmentation process plays a pivotal role in enhancing the robustness and diversity of the dataset used for training machine learning models, particularly in the domain of medical image segmentation. In the provided code, a comprehensive data augmentation pipeline is implemented to augment retinal fundus images and their corresponding masks. The primary goal is to introduce variability in the dataset by applying horizontal flips, vertical flips, and rotation to the original images and masks. The augmentation process aims to simulate different orientations and perspectives that may be encountered in real-world scenarios, thereby enriching the dataset and improving the generalization capability of the subsequent U-Net model.[9]

The "augment_data" function iterates through the training dataset, applying various augmentations to each image-mask pair. The augmentations include horizontal flips, vertical flips, and rotations, with the associated masks adjusted accordingly. The resulting augmented images and masks are resized to a standardized dimension of (512 × 512). The augmented data is then saved in a separate directory structure, creating distinct folders for augmented images and masks. The use of data augmentation is particularly valuable when the dataset size is limited, as it introduces diversity that aids in preventing overfitting and improves the model's ability to handle variations in real-world data.

It is important to note that the data augmentation pipeline is designed to be flexible, allowing for the option to enable or disable augmentation based on the "augment" parameter. This flexibility caters to different experimental setups, enabling researchers to assess the impact of data augmentation on model performance. The implementation adheres to best practices in data augmentation for medical image analysis, contributing to the overall reliability and generalization of the U-Net model for retinal fundus image segmentation tasks.[21]

2.3. Network architecture

The neural network architecture presented in the code is a U-Net, a popular architecture widely used in image segmentation tasks. The U-Net architecture consists of an encoder-decoder structure with skip connections, allowing the model to capture both high-level and low-level features effectively. The encoder portion of the network employs convolutional blocks to extract hierarchical features from the input image.[30] Specifically, it comprises four encoder blocks, each consisting of two convolutional layers with batch normalization and rectified linear unit activation functions, followed by max-pooling layers for downsampling, as presented in Figure 4.[10]

The bottleneck layer acts as a feature representation for the entire input image, condensing the learned features.[34] It consists of a convolutional block with the same structure as the encoder blocks. The decoder portion of the network utilizes transposed convolutions for up-sampling and concatenates the features from the corresponding encoder block through skip connections.[28] This enables the decoder to recover spatial information lost during the down-sampling process. The decoder also incorporates convolutional blocks for feature refinement.[11]

The classifier at the end of the network is a 1 × 1 convolutional layer, mapping the features to a single channel output, which is suitable for binary segmentation tasks.[24] The entire architecture is designed for semantic segmentation, particularly for tasks where precise delineation of object boundaries is crucial. In summary, this U-Net architecture facilitates robust feature extraction, effective information fusion through skip connections, and accurate segmentation outputs.[3]

2.4. Training process

2.4.1. Pre-training

In the pre-training phase, the U-Net model is meticulously configured with various hyperparameters to ensure optimal performance. The image dimensions (H [height] × W [weight]), batch size, number of epochs, learning rate, and checkpoint path for model saving are carefully set. The dataset is then loaded using custom data loaders, and the training and validation sets are meticulously prepared from augmented retinal fundus images along

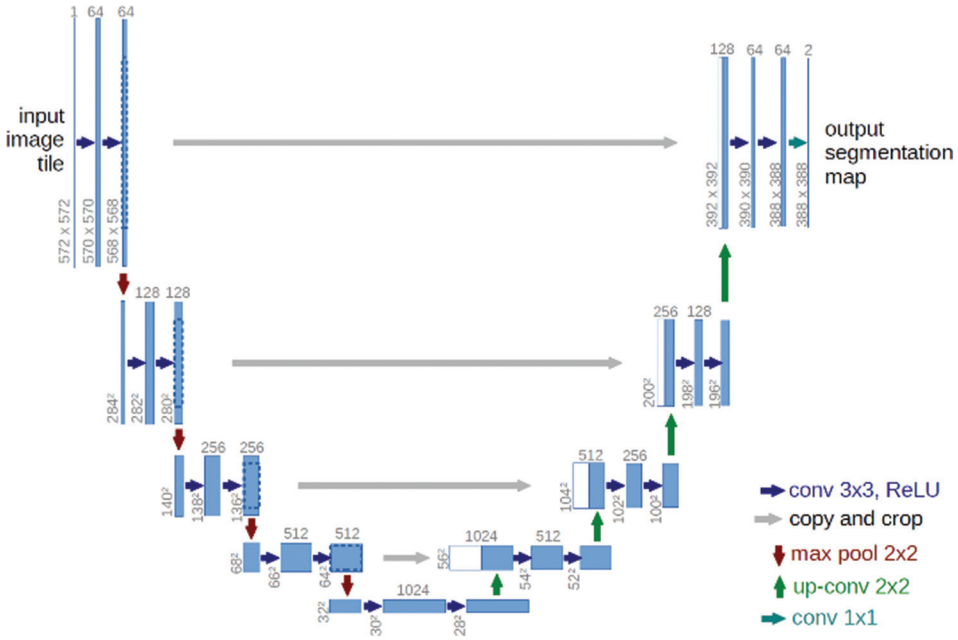


Figure 4. The picture illustrates the architecture of U-Net layers used for segmentation. Image reprinted with permission, Copyright © 2015, Springer International Publishing Switzerland.[38]
Abbreviations: Conv: 3 ×x 3; ReLU: Rectified linear unit.

with their corresponding masks. To offer a comprehensive overview of the experimental setup, key hyperparameters and insightful statistics about the dataset are presented. The choice of computation device, which is based on the availability of a Compute Unified Device Architecture-enabled graphic processing unit, is disclosed, and the U-Net model is adeptly moved to the selected device. The initialization of the Adam optimizer, the learning rate scheduler, and the combination of dice loss and binary cross-entropy provide a robust foundation for the subsequent training phases.[31]

2.4.2. Main training

The main training phase unfolds over a specified number of epochs, wherein the U-Net model undergoes iterative training.[33] Within each epoch, the model is rigorously trained using the prepared training dataset, and the optimizer diligently works to minimize the loss, computed through a combination of the dice loss and binary cross entropy. Simultaneously, the model's proficiency is rigorously evaluated on the validation dataset to monitor its generalization capabilities.[18] The training process is intricately monitored with detailed output, including epoch-wise loss values and elapsed time, fostering a nuanced understanding of the model's convergence patterns. The strategic selection of the best model checkpoint ensures that the model attains optimal performance. This checkpoint, capturing the model's state with the lowest validation loss, serves as a pivotal asset for future deployment and analysis.[33]

2.4.3. Post-training

The post-training phase marks the culmination of the training experiment, where the training outcomes are meticulously summarized, and the final model is prepared for deployment or further analysis. The best-performing model is meticulously selected based on the lowest validation loss achieved during training and is safeguarded as a checkpoint for subsequent use. Key metrics, such as training loss, validation loss, and epoch-wise training times, are presented to provide a holistic evaluation of the model's performance.[18] Furthermore, this phase allows for insights into the training process, including potential improvements or challenges faced, fostering a deeper understanding of the model's behavior in the context of DR detection (Figure 5). The post-training phase thus solidifies the experiment's completion, with the trained U-Net model ready for deployment in practical applications or further investigative studies.[34]

2.5. Testing

In the testing phase, the trained U-Net model is rigorously evaluated on a separate dataset to assess its performance in semantic segmentation of retinal fundus images for DR detection. The experiment involves loading the preprocessed test dataset, consisting of retinal fundus

images and their corresponding ground truth masks. Utilizing the U-Net model previously trained on augmented data, the model's predictive capabilities are scrutinized for pixel-level segmentation accuracy. The model's state is restored using the best-performing checkpoint achieved during the training phase, ensuring the evaluation is based on the most optimized configuration.

For each test sample, the retinal fundus image is pre-processed by normalizing pixel values and transposing the channels to match the model's input requirements. Similarly, the ground truth mask undergoes pre-processing to facilitate direct comparison with the model predictions. The evaluation metrics, including Jaccard index, F1 Score, recall, precision, and accuracy, are computed for each test image. These metrics quantify the model's ability to accurately delineate DR-related regions in the retinal fundus images.[35]

The computational efficiency of the model is also assessed through the calculation of frames per second (FPS) during the inference process. This metric provides insights into the real-time processing capabilities of the model, offering valuable information for potential deployment in clinical or real-world scenarios. Visual representations of the model's predictions, alongside the original retinal fundus images and ground truth masks, are saved for qualitative analysis and comparison.

3. Results

3.1. Segmentation results

UNet achieved the highest performance with a dice coefficient of 0.95 and an intersection over union (IoU) of 0.92. Following closely, fully convolutional network exhibited a pixel accuracy of 0.92 and an IoU of 0.70. DeepLab attained a mean IoU (mIoU) of 0.80, while SegNet and LinkNet demonstrated pixel accuracies of 0.88 and 0.85, respectively, with IoUs of 0.65 and 0.68. PSPNet yielded a mIoU of 0.78. Finally, Mask R-CNN, evaluated by average precision, achieved a performance of 0.65. These results provided insights into the efficacy of different segmentation methods, highlighting UNet as a particularly promising approach for accurate image segmentation tasks in Tables 1 and 2.

In this study, we employed a U-Net-based model for the segmentation of blood vessels in retinal images. The implemented model was evaluated on a test dataset comprising retinal images and corresponding ground truth masks. The testing process involved loading images and masks, pre-processing the data, and utilizing a pre-trained U-Net model for predictions. The model's performance was assessed using several metrics, including Jaccard similarity, F1 score, recall, precision, and accuracy.

Table 1. Segmentation results

Segmentation method	Metric(s)	Performance
UNet	Dice coefficient; IoU	0.95 dice; 0.92 IoU
DeepLab	Dice coefficient; IoU	0.80 mIoU
FCN	Pixel accuracy; IoU	0.92-pixel accuracy; 0.70 IoU
SegNet	Pixel accuracy; IoU	0.88-pixel accuracy; 0.65 IoU
PSPNet	Mean intersection over union (mIoU)	0.78 mIoU
LinkNet	Pixel accuracy; IoU	0.85-pixel accuracy; 0.68 IoU
Mask R-CNN	AP	0.65 AP

Abbreviations: IoU: Intersection over union; mIoU: Mean intersection over union; AP: Average precision; FCN: Fully Convolutional Network.

Table 2. Performance metrics for lesion segmentation in the Indian diabetic retinopathy image dataset

Different segmentations	Jaccard score	F1 score	Recall score	Precision	Accuracy
Blood vessel	0.6634	0.7974	0.7771	0.8240	0.9922
Hard exudate	0.6663	0.7953	0.7634	0.8252	0.9986
Soft exudate	0.6679	0.7874	0.7738	0.8340	0.9981
Hemorrhage	0.6551	0.7874	0.7852	0.8161	0.9958
Microaneurysms	0.6761	0.8061	0.7731	0.8279	0.9967
Optical disc	0.6638	0.7956	0.7671	0.8238	0.9989

Notably, the model demonstrated efficient segmentation with an average accuracy of 0.9986 and an impressive FPS rate of 361.188719052. Visual results were generated for each test image, illustrating the original image, the ground truth mask, and the predicted segmentation mask. Overall, these findings highlight the effectiveness and computational efficiency of the proposed blood vessel segmentation model, showcasing its potential for applications in DR diagnosis and treatment planning, as visualized in Figure 6.

In this research, we conducted a comprehensive evaluation of a segmentation model across various retinal structures, including blood vessels,[15] EXs,[14] SEs, HEs, MAs, and the optical disc.[37] The model's performance was quantitatively assessed using key metrics, revealing high segmentation accuracy across all structures. The Jaccard scores ranged from 0.6551 to 0.6761, indicating substantial overlap between the predicted and ground truth masks. The model achieved notable F1 scores, demonstrating a harmonious balance between precision and recall, ranging from 0.7874 to 0.8061. Particularly, commendable was the recall scores, signifying the model's ability to correctly identify relevant instances, with values ranging from 0.7634 to 0.7852. Precision scores, representing the

accuracy of positive predictions, ranged from 0.8161 to 0.8340. The overall accuracy of the model was consistently high across all structures, with values ranging from 0.9922 to 0.9989, as presented in Table 2. These results collectively underscored the efficacy of the segmentation model in accurately delineating retinal structures, showcasing its potential for enhancing diagnostic capabilities in the context of retinal pathology assessment.

3.2. Image grading results

In our investigation of the APTOS dataset, we thoroughly assessed the performance of our proposed method in DR grading. Our method exhibited exemplary results, achieving the highest accuracy (ACC) and kappa scores among all evaluated methods. Specifically, our approach attained an ACC of 89.1% and a kappa score of 93.4%, surpassing the performance of MIL-VT, which achieved an AUC of 97.9%. These results underscore the robustness and effectiveness of our proposed method for accurately grading DR on the APTOS dataset, as presented in Table 3. Notably, our method demonstrated competitive performance compared to state-of-the-art models, highlighting its potential as a reliable tool for precise DR grading in clinical settings.[38]

3.3. Application integration results

The proposed interface for DR detection and classification is designed to offer a comprehensive diagnostic platform by incorporating multiple segmentation techniques with deep learning models for fundus image analysis. Users can interact with various features, including buttons for triggering segmentation techniques, allowing the identification of specific regions of interest related to potential DR indicators. The interface also provides the flexibility to toggle between grayscale and normal views, facilitating a detailed examination of fundus images in different visual representations.[34] The DR classification

Table 3. DR and DME grading on the IDRiD dataset[37]

Methods	AUS	ACC	F1	Kappa
DLI	-	82.5	80.3	89.0
CANet	-	83.2	81.3	90.0
GREEN-ResNet50	-	84.4	83.6	90.8
GREEN-SE-ResNext50	-	85.7	85.2	91.2
MIL-VT	97.9*	85.5	85.3	92.0
VT	97.9*	89.1*	88.9*	93.4*

Note: * Have the best performance across multiple metrics (ACC, F1, Kappa) compared to the other listed methods.
Abbreviations: ACC: Accuracy; AUC: Area under the curve; DME: Diabetic macular edema; DR: Diabetic retinopathy; IDRiD: Indian diabetic retinopathy image dataset.

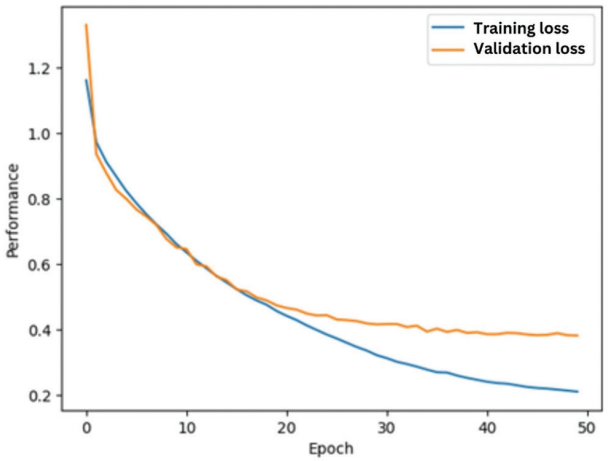


Figure 5. Training and validation accuracy. Image generated using VS code

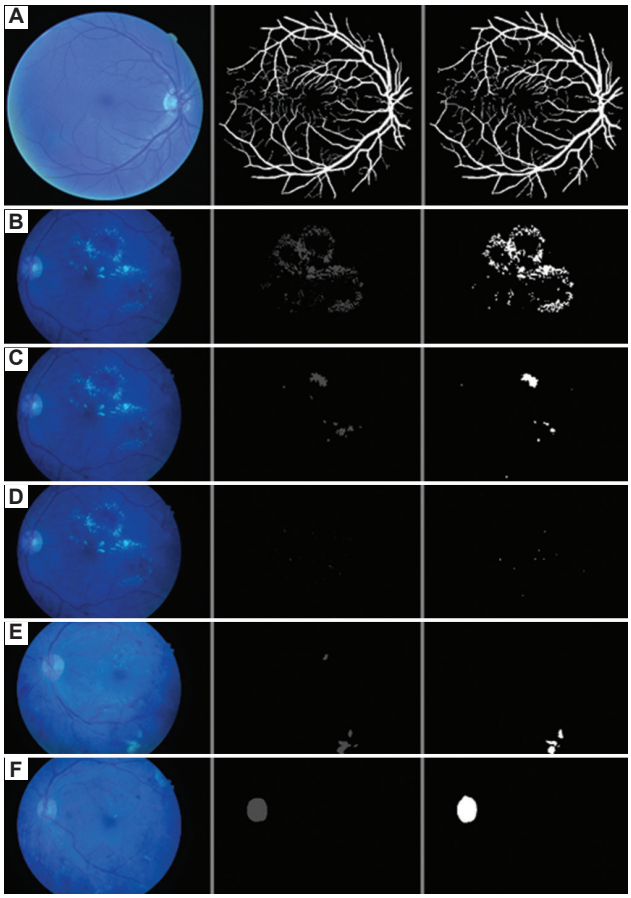


Figure 6. Segmentation results for different types of lesions segmentation on the Indian Diabetic Retinopathy Image Dataset (IDRiD) showing the original image, its corresponding mask, and the predicted segmented image. These results were generated using VS code during the testing phase. (A) Blood vessel segmentation; (B) hard exudate segmentation; (C) hemorrhage segmentation; (D) microaneurysm segmentation; (E) soft exudate segmentation; (F) optical disc segmentation. Images generated using VS code.

G
H

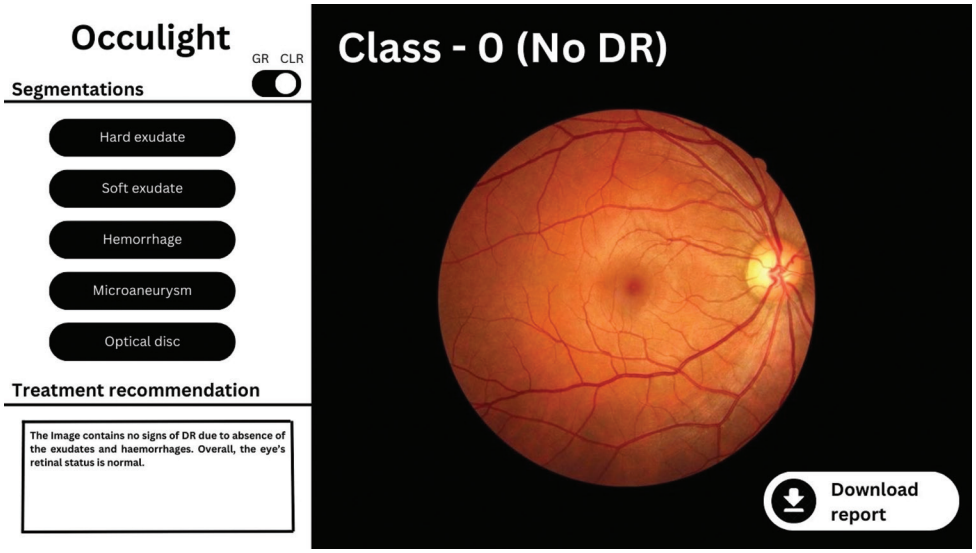


Figure 7. Diabetic retinopathy (DR) detection and grading system interface. Image generated using Canva.com.

display at the top of the image promptly communicates the severity of the detected condition with a numerical value (0, 1, 2, 3, or 4). Following segmentation results, users are prompted to input relevant information for further analysis, with the data processed by a pre-trained LLM to recommend a suitable treatment strategy for the identified DR class. The interface culminates in the generation of a downloadable PDF report, consolidating segmented images, normal and grayscale views, classification results, and the recommended treatment, offering a comprehensive summary of the analysis performed by the system (Figure 7). This integrated approach enhances the diagnostic capabilities, making the interface a user-friendly and holistic solution for DR detection and classification.[26]

3.4. LLM results

In the domain of medical diagnosis and treatment prediction, the utilization of pre-trained LLMs has emerged as a promising avenue for enhancing clinical decision-making processes. Leveraging the flexibility and adaptability of LLMs, our research explored their application in predicting tests and treatment strategies based on the stage predicted by image grading algorithms. Using the output from image grading as input for LLMs, we capitalized on the comprehensive information extracted from retinal images to inform subsequent medical decisions. The inherent capacity of LLMs to comprehend and contextualize complex medical data enabled them to provide nuanced predictions tailored to individual patient profiles. This integration of image grading results with LLM-based prediction models facilitated a holistic approach to patient care, allowing for more accurate and personalized test recommendations and treatment plans. Furthermore, the dynamic nature of LLMs enables continuous learning and refinement, ensuring adaptability to evolving clinical scenarios and enhancing the efficacy of predictive analytics in health care. Overall, our findings underscore the potential of leveraging LLMs in conjunction with image grading techniques to optimize clinical workflows and improve patient outcomes in the field of ophthalmology.[32]

4. Conclusion

In this research paper, we address the challenges associated with the early detection of DR, a severe complication of diabetes that can lead to blindness. The escalating prevalence of DR underscores the critical need for accurate and timely diagnosis. Traditional diagnostic methods often face inefficiencies and disagreements among clinicians, prompting the development of algorithms, particularly deep learning approaches, to enhance DR detection. Our proposed approach employs transfer learning, leveraging a single fundus photograph for automatic DR stage detection. Notably, it achieved a commendable ranking in the APTOS 2019 Blindness Detection Competition, emphasizing its effectiveness with a high quadratic weighted kappa score of 0.92546. The research aims to contribute significantly to DR detection methodologies, particularly in the context of automated systems, addressing the crucial requirement for early diagnosis and intervention.

The study reviews related work, tracing the evolution from classical computer vision approaches to the rise of deep learning, where CNNs have demonstrated prowess in DR classification. Transfer learning with CNN architectures is explored, highlighting promising results achieved by various research teams. The problem

statement emphasizes the challenges in existing diagnostic methods and introduces datasets encompassing lesion segmentation, disease/image grading, and LLMs for test/treatment recommendations.

Identifying research gaps, the paper underscores the need for exploring the integration of pre-trained LLMs with segmented image data, emphasizing the potential synergies between visual segmentation features and clinical classifications within a decision-support system. Another research gap pertains to the dynamics of multi-model integration, particularly in a web application context where lesion segmentations, disease classification, and LLMs collaborate. The objectives outline a comprehensive DR detection methodology, the integration of various models, performance evaluation, contribution to detection methodologies, and the identification and exploration of research gaps.

The research scope aims to revolutionize DR detection methodologies by integrating cutting-edge technologies and contributing to the wider field of medical imaging and automated diagnostics. The methods encompass data pre-processing, data augmentation, network architecture detailing the U-Net model, and the training and testing processes. The results showcase the effectiveness of the segmentation model across various retinal structures, with high Jaccard scores, F1 scores, recall scores, precision scores, and overall accuracy, underscoring its potential for enhancing diagnostic capabilities in retinal pathology assessment.

Acknowledgments

We would like to express our gratitude to our guide, Dr. Narender M., Assistant Professor, Department of Computer Science and Engineering at the National Institute of Engineering (NIE), for his invaluable guidance and support throughout the course of this research. His expertise and insightful feedback significantly contributed to the development of this work.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Manoj Saligrama Harisha, Arya Arun Bhosale
Investigation: All authors
Methodology: Manoj Saligrama Harisha
Writing-original draft: Manoj Saligrama Harisha, Arya Arun Bhosale
Writing-review & editing: Manoj Saligrama Harisha, Arya Arun Bhosale

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data for this study are sourced from two primary datasets: the APTOS 2019 Blindness Detection dataset, available on Kaggle.com, and the IDRiD, accessible through Grad-Challenge.org. In addition, the code repository for the DR Detection and Classification System used in this research can be found at https://github.com/Manoj-Sh-AI/Diabetic-Retinopathy-Detection-and-Clasification-System.

Further disclosure

This paper has been uploaded and made publicly available on a preprint server (arXiv:2401.02759), available at: https://arxiv.org/abs/2401.02759

References

1. NCHS. *Eye Disorders and Vision Loss among U.S. Adults Aged 45 and Over with Diagnosed Diabetes*; 2019. Available from: https://stacks.cdc.gov/view/cdc/80081 [Last accessed on 2024 Jul 11].
2. Pezzullo L, Streatfeild J, Simkiss P, Shickle D. The economic impact of sight loss and blindness in the UK adult population. *BMC Health Serv Res*. 2018;18:63.
   doi: 10.1186/s12913-018-2836-0.
3. Hann CE, Chase JG, Revie JA, Hewett D, Shaw GM. Diabetic retinopathy screening using computer vision. *IFAC Proc Vol*. 2009;42:298-303.
   doi: 10.3182/20090812-3-DK-2006.0086
4. Rohan TE, Frost CD, Wald NJ. Prevention of blindness by screening for diabetic retinopathy: A quantitative assessment. *BMJ*. 1989;299:1198-1201.
   doi: 10.1136/bmj.299.6709.1198
5. SNEC. *Singapore's Eye Health*; 2019. Available from: https://www.snec.com.sg [Last accessed on 2024 Jul 11].
6. Priya R, Aruna P. *SVM and Neural Network Based Diagnosis of Diabetic Retinopathy*; 2012. Available from: https://www.researchgate.net/publication/261177114_svm_and_neural_network_based_diagnosis_of_diabetic_retinopathy [Last accessed on 2024 Jul 11].

7. Conde P, De la Calleja J, Medina M, Benitez-Ruiz AB. *Application of Machine Learning to Classify Diabetic Retinopathy*; 2012. Available from: https://www.researchgate.net/publication/259251279_application_of_machine_learning_to_classify_diabetic_retinopathy [Last accessed on 2024 Jul 11].

8. Pratt H, Coenen F, Broadbent DM, Harding SP, Zheng Y. Convolutional neural networks for diabetic retinopathy. *Procedia Comput Sci*. 2016;90:200-205.

   doi: 10.1109/IJCNN.2016.7727515

9. Lam C, Yi D, Guo M, Lindsey T. Automated detection of diabetic retinopathy using deep learning. *AMIA Jt Summits Transl Sci Proc*. 2018;2017:147-155.

10. Li YH, Yeh NN, Chen SJ, Chung YC. Computer-assisted diagnosis for diabetic retinopathy based on fundus images using deep convolutional neural network. *Mobile Inf Syst*. 2019;2019:1-14.

    doi: 10.3390/app9030448

11. Asiri N, Hussain M, Aboalsamh HA. Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey. *Artif Intell Med*. 2019;99:101701.

    doi: 10.1016/j.artmed.2019.07.009

12. Hagos MT, Kant S. *Transfer Learning-based Detection of Diabetic Retinopathy from Small Dataset* [Preprint]; 2019.

    doi: 10.48550/arXiv.1905.07203

13. Sarki R, Michalska S, Ahmed KE, Wang H, Zhang Y. *Convolutional Neural Networks for Mild Diabetic Retinopathy Detection: An Experimental Study EyePACs (2015)*; 2019. Available from: https://www.eyepacs.org [Last accessed on 2024 Jul 11].

14. Fu Y, Zhang G, Lu X, Wu H, Zhang D. U-net: Hard Exudate segmentation for retinal fundus images, expert systems with applications. *Expert Syst Appl*. 2023;234:120987.

    doi: 10.1016/j.eswa.2023.120987

15. Fu Y, Zhang G, Li J, Pan D, Wang Y, Zhang D. Fovea localization by blood vessel vector in abnormal fundus images. *Pattern Recognit*. 2022;129:108711.

    doi: 10.1016/j.patcog.2022.108711

16. Sahasrabuddhe V, Porwal P, Meriaudeau F, *et al*. *Indian Diabetic Retinopathy Image Dataset* (*IDRID*); 2018. Available from: https://idrid.grand-challenge.org [Last accessed on 2024 Jul 11].

17. Devries T, Taylor GW. *Improved Regularization of Convolutional Neural Networks with Cutout* [Preprint]; 2017.

    doi: 10.48550/arXiv.1708.04552

18. Architecture for Computer Vision. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Available from: https://www.cv-foundation.org [Last accessed on 2024 Jul 11].

19. Google Brain. *Messidor-2 Diabetic Retinopathy Grades*; 2019; Available from: https://research.google/tools/datasets/messidor2

20. APTOS. *APTOS 2019 Blindness Detection*; 2019. Available from: https://www.kaggle.com/c/aptos2019-blindness-detection [Last accessed on 2024 Jul 11].

21. Buslaev A, Parinov A, Khvedchenya E, Iglovikov VI, Kalinin AA. *Albumentations: Fast and Flexible Image Augmentations* [Preprint]; 2020.

    doi: 10.48550/arXiv.1809.06839

22. Cheng J. *A Neural Network Approach to Ordinal Regression* [Preprint]; 2007.

    doi: 10.48550/arXiv.0704.1028

23. Abràmoff MD, Reinhardt JM, Russell SR, *et al*. Automated early detection of diabetic retinopathy. *Ophthalmology*. 2010;117:1147-1154.

    doi: 10.1007/978-3-642-18551-8_17

24. Naveed H, Khan AU, Qiu S, *et al*. *A Comprehensive Overview of Large Language Models* [Preprint]; 2023.

    doi: 10.48550/arXiv.2303.02171

25. Lundberg SM, Lee SI. *A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems*; 2017. Available from: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-abstract.html

26. Silberman N, Ahlrich K, Fergus R, Subramanian L. Case for Automated Detection of Diabetic Retinopathy. In: *Artificial Intelligence for Development - Papers from the AAAI Spring Symposium, Technical Report*; 2010.

    doi: 10.1109/CVPR.2010.5540189

27. Paszke A, Gross S, Chintala S, *et al*. *Automatic Differentiation in PyTorch. In NIPS Autodiff Workshop*; 2017. Available from: https://openreview.net/forum?id=bjjsrmfcz [Last accessed on 2024 Jul 11].

28. Wan S, Liang Y, Zhang Y. Deep convolutional neural networks for diabetic retinopathy detection by image classification. *Comput Electr Eng*. 2018;72(10):274-282.

    doi: 10.1109/ACCESS.2018.2883722

29. Tan M, Le QV. EfficientNet: *Rethinking Model Scaling for Convolutional Neural Networks* [Preprint]; 2019.

    doi: 10.48550/arXiv.1905.11946

30. Hu J, Shen L, Sun G. *Squeeze-and-Excitation Networks* [Preprint]; 2017.

    doi: 10.48550/arXiv.1709.01507

31. Loshchilov I, Hutter F. *SGDR: Stochastic Gradient Descent with Restarts* [Preprint]; 2016.

    doi: 10.48550/arXiv.1608.03983

32. Nichol A, Achiam J, Schulman J. *On First-order Meta-learning Algorithms* [Preprint]; 2018.

doi: 10.48550/arXiv.1803.02999

33. Yoon B. *A Machine Learning Approach for Efficient Multi-dimensional Integration. Scientific Reports*; 2021. Available from: https://www.nature.com/articles/s41598-021-81994-8

34. Krogh A, Hertz JA. *A Simple Weight Decay can Improve Generalization. In: Advances in Neural Information Processing Systems*; 1992. Available from: https://papers.nips.cc/paper/1992/file/8eefcfdf5990e441f0fb6f3fad709e21-paper.pdf [Last accessed on 2024 Jul 11].

35. Krause J, Gulshan V, Rahimy E, *et al. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy* [Preprint]; 2017.

doi: 10.48550/arXiv.1710.01711

36. Fu Y, Chen J, Li J, Pan D, Yue X, Zhu Y. Optic disc segmentation by U-Net and probability bubble in abnormal fundus images. *Pattern Recognit*. 2021;117(12):107971.

doi: 10.1016/j.patcog.2021.107978

37. Zhu W, Qiu P, Chen X, *et al. nnMobile Net: Rethinking CNN for Retinopathy Research* [Preprint]; 2024.

doi: 10.48550/arXiv.2306.01289

38. Ronneberger O, Fischer P, Brox T. *U-Net: Convolutional Networks for Biomedical Image Segmentation. Papers*; 2015.

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

A multi-adaptive neuro-fuzzy inference system with variable thresholds for heartbeat classification

Roghayeh Rafieisangari and Nabiollah Shiri*

Department of Electrical Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran

Abstract

Various heart disorders are non-invasively diagnosed using electrocardiograms (ECGs). An ECG records a variety of waveforms, including P, QRS, and T waves, which represent the electrical activity of the human heart. Cardiovascular diseases are diagnosed by examining the length, form, and spacing of these waveforms. This research develops a multi-adaptive, neuro-fuzzy inference system (MANFIS), enhanced by a variable threshold approach, to enhance heartbeat classification accuracy. The MIT-BIH arrhythmia database was utilized, and seven features were extracted from each record. A subtractive clustering method was employed to prepare the inputs for the MANFIS, enabling heartbeat classification. By applying a variable threshold to the MANFIS outputs, classification accuracy was further enhanced. The proposed method, termed variable-threshold MANFIS, can separately detect normal sinus rhythm, left bundle branch block, right bundle branch block, premature ventricular contraction, atrial premature condition, and paced beat. This is achieved using six different ANFIS classifiers, each with its own threshold. The system was evaluated, achieving an accuracy of 98.33%, a sensitivity of 93.12%, a specificity of 99.66%, a precision of 98.33, and an F_1-score of 95.44. A distinct feature of this machine-learning-based model is its controllable threshold, which delivers promising results across all training, testing, and validation datasets. The proposed diagnostic system is applicable in new automated medical instrumentation and serves as a valuable tool in cardiology.

Keywords: Electrocardiograms signals; Feature extraction; Classification; Adaptive neuro-fuzzy inference system; Subtractive clustering; Variable threshold

***Corresponding author:**
Nabiollah Shiri
(na.shiri@iau.ac.ir)

Citation: Rafieisangari 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

Received: April 4, 2024

Accepted: June 26, 2024

Published Online: October 24, 2024

1. Introduction

The World Health Organization reports that the number one cause of death worldwide is cardiovascular diseases (CVDs). In recent years, various programs have been implemented in increasingly diverse communities to reduce the incidence of both initial and recurrent cardiovascular events. To accomplish this, the electrocardiogram (ECG) has emerged as the most widely utilized bio-signal for the early detection of CVDs. The ECG graphically displays the electrical activity of the heart and is used to diagnose a range of heart conditions and anomalies. The ECG signals have been used

by medical professionals for more than 70 years to identify heart conditions, such as arrhythmias.[1] The heart's rhythm and bioelectrical activity are expressed in an ECG. Adults typically have a resting heart rate between 60 and 100 beats/min. A lower resting heart rate typically indicates improved cardiovascular fitness and more effective cardiac function. An athlete who has received proper training, for instance, may typically have a resting heart rate of about 40 beats/min. In contrast, heart disease can alter the shape and characteristics of the heartbeats seen in an ECG, resulting in significant deviations from the normal pattern.

Cardiac arrhythmias are common in CVDs, and their accurate classification is crucial, as successful treatment relies on early detection. A typical ECG waveform is made up of the P, QRS, and T waves in a single period (Figure 1). Each wavelet and segment of waveform carries distinct energy and physiological significance. Notably, the QRS is higher in energy and amplitude compared to the P and T wave groups.[2] The ability of wearable or implantable remote monitoring devices to continuously monitor cardiac activity allows for more effective healthcare for patients with periodic heart arrhythmias. However, these devices generate a significant amount of ECG data that medical professionals must interpret. Consequently, there is a growing need for reliable techniques of automatic ECG interpretation to assist doctors. The simultaneous capture of copious amounts of ECG data requires efficient and accurate interpretation. Doctors, nurses, and other medical professionals are responsible for this task. This extra workload compounds the fatigue that medical staff already experience, raising the risk of medical errors. Furthermore, a large percentage of the received ECG recordings are often false alarms, as remote monitoring devices are highly sensitive to ECG abnormalities and may not effectively filter out significant cardiovascular events. As a result, helping doctors interpret ECGs has become increasingly important.

Every year, millions of ECG recordings are taken globally, with most being automatically processed and deciphered by computers. This places pressure on the ECG interpretation techniques to be patient- and device-independent, as well as quick and accurate. Deep learning (DL) techniques, which can process massive volumes of raw data and the widespread digitization of ECG data have opened up new avenues for enhancing automated ECG interpretation. As a result, supporting doctors in interpreting ECG recordings is becoming increasingly important.[3] DL is the study of knowledge extraction, intelligent decision-making, and prediction, or the process of identifying complex patterns from a corpus of primary sentences or training data. There have been several DL models proposed in recent years to increase the accuracy of various learning tasks.[4] This paper contributes to the design of a variable-threshold multi-adaptive neuro-fuzzy inference system (VTMA) for classifying different heartbeats. To improve the accuracy and speed of heart

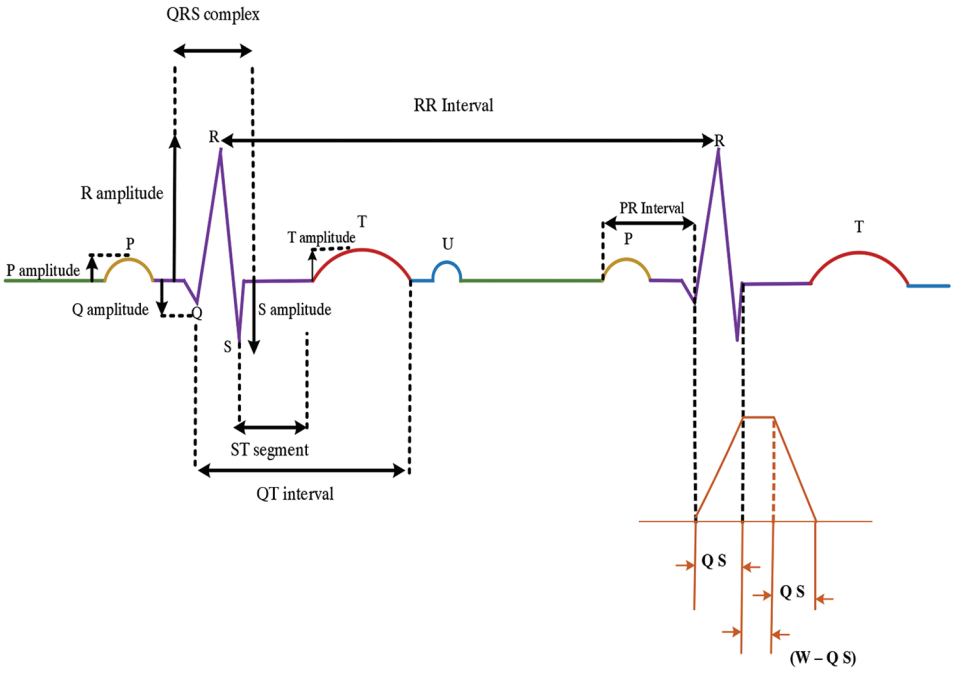


Figure 1. Labeled features of the electrocardiograms[1]

condition detection, a variable threshold is applied to the adaptive neuro-fuzzy inference system (ANFIS) output, and the parameters are adjusted accurately. The proposed system has three parts. First, the input ECGs undergoes preprocessing to eliminate noise. In the second step, feature extraction prepares the inputs for the ANFIS, while heart conditions are labeled using the subtractive clustering method to train the ANFIS.

ANFIS is a binary classifier, so to classify six classes; six separate ANFIS systems are required. This mechanism, as the third stage, can be known as a multi-binary classifier. Across this process, classification is achieved; however, to enhance accuracy, six thresholds are applied to the output of each ANFIS. By tuning the threshold values, a variable-threshold system is created that yields optimal results. The categorization of ECG arrhythmias is crucial for quick identification and diagnosis of cardiovascular disorders. A more accurate diagnosis allows for timely and suitable interventions when needed. These diagnostic techniques can be integrated into electronic devices used by individuals with cardiac problems, and in the event of an emergency, appropriate alerts can be delivered to medical facilities, such as hospitals or physicians. In addition, hospitals can regularly employ these methods to diagnose cardiac problems early, reducing the need for human intervention. With the use of ECG classification approaches, individuals with established cardiac diseases can be monitored continuously. As the disease progresses or as a patient responds to treatment, changes in the ECG pattern over time can help doctors make well-informed judgments about modifying treatment plans. With significant accuracy and precision, the suggested method can make a significant contribution to this field by classifying six classes of heart diseases. Since biomedical signals are very sensitive, even a small improvement in speed, accuracy, and precision can greatly impact individuals' health and well-being. Related works are explored in Section 2, while Section 3 discusses the data and methods employed in this research. The results are discussed in Section 4, and the conclusion is provided in Section 5.

2. Related works

Researchers aim to achieve accurate and high-speed classification while keeping computational costs low. Robust automated diagnostic approaches require preprocessing of the ECG, enhancing the signal, extracting features, and classifying the data. Different techniques have been explored in the literature for detecting diseases through ECG analysis. To enhance ECG analysis, an algorithm is put forth that makes use of the fractional Fourier transform (FrFT) and two-event-related moving averages (TERMA) algorithms. While the FrFT rotates ECG signals in the time-frequency plane to display the locations of different peaks, the TERMA algorithm designates specific areas of interest to locate the desired peak.[1] In the study, a DL-based system is presented; using convolutional neural networks (CNNs) for ECG classification with the PhysioNet MIT-BIH Arrhythmia database.[1] The suggested system[1] uses a 1-D convolutional deep residual neural network (ResNet) model, which uses the input heartbeats directly to extract features. To handle the class imbalance in the training dataset and effectively classify five heartbeat types in the test dataset, the synthetic minority oversampling technique was employed.[2] In addition, raw ECG recordings are classified using deep CNNs.[3] However, these CNNs require extensive annotated samples for effective training, which can be costly to obtain. To mitigate this issue, transfer learning is utilized.[3] Using the largest available collection of continuous raw ECG signals, the first CNNs were pre-trained. Next, the networks were refined for the most common cardiac arrhythmia and atrial fibrillation using a small data set. An artificial NN approach was presented for the automatic identification and categorization of ECG.[4] To thoroughly mine the hierarchical and time-sensitive features of ECG data, a dense heart rhythm network has been developed that combines a 24-layer deep CNN and bidirectional long short-term memory. The original ECG is filtered using a combination of wavelet transform and median filtering to remove the influence of noise on the signal. In addition, three different sizes of convolution kernels (32, 64, and 128) are used to mine the detailed features from the ECG signal.

The symlet wavelet transform was presented to detect the QRS complex and reduce the error.[5] Using values of RR intervals, amplitude, and Hjorth parameters, some features of the ECG were extracted for heartbeat classification.[6] Techniques, including variation mode decomposition, phase space reconstruction, euclidean distance, and Shannon energy envelope were employed to detect myocardial dysfunction.[7] The Hilbert-Huang transform was used for feature selection, which includes a set of essential features.[8] The ANFIS employed Lyapunov exponents for the ECG classification.[9] A reliable beat classification was performed using the wavelet transform and principal component analysis-independent component analysis.[10] An extreme learning machine was applied to the MIT/BIH database, and feature selection was performed using the variances of the wavelet transform and parameters of the autoregressive model.[11] Furthermore, the ECG classification was done by an automatic, reliable, two-stage hybrid hierarchical approach.[12] The ANFIS, along with a fuzzy rule-based model classifier, was used to identify premature ventricular contraction (PVC) beats with a high readability-accuracy trade-off.[13]

Neural networks provide efficient ways for ECG classification without requiring pre-processing,[14] such as multilayer perceptron (MLP).[15] Based on the conjoint use of the MLP that was trained by an enhanced particle swarm optimization algorithm, the ECG arrhythmias were classified.[16] Four types of heart rates were classified by identifying QRS features that were extracted from multi-resolution wavelet transform.[17] In addition, a quality-aware mechanism was employed to classify ECG beats, reducing false alarms and ensuring accuracy.[18] A multi-module neural network system was developed to classify ECGs, specifically addressing the issue of heartbeat imbalance.[19] A research architecture employs ANFIS to learn fuzzy logic, using inputs that are preprocessed with the subtractive clustering method. Five morphological and five statistical ECG features were used to classify the patient's heartbeats based on whether they were irregular or normal.[20] Six various heart conditions, including normal sinus rhythm (NSR), PVC, atrial premature condition (APC), left bundle branch block (LBBB), right bundle branch block (RBBB), and paced beat (PB), were detected by an ANFIS.[21] A weight assignment method based on multi-label ECGs, combined with an ensemble classifier, was applied for classification.[22] The neuro-fuzzy system has proven helpful in disease diagnosis,[23] while some research has employed eigenvalues and DL for ECG classification.[24] Considering previous research, some challenges remain, such as accuracy improvement, complexity reduction, speed increment, and power reduction. There is a growing interest in computer-aided identification and diagnosis of cardiac illness using ECG data. Some researchers are turning to neural networks to overcome the drawbacks of manual feature selection methods. However, it is still difficult to build and choose a high-performing diagnostic model that is appropriate for clinical implications.[25,26]

3. Data and methods

To improve the speed and accuracy of ECG classification, a modified ANFIS structure is proposed (Figure 2). The advantages of neural networks and fuzzy logic systems are combined in ANFIS. Its hybrid approach allows it to learn and comprehend complex patterns in the data in an adaptive way, which greatly increases its versatility for classification jobs. ANFIS can effectively handle this uncertainty because of its fuzzy logic component, which allows for approximate reasoning and decision-making in the face of ambiguity and vagueness. Over time, ANFIS models can adapt to changes in the input data distribution or environment. This adaptability is especially useful for recognizing heartbeats in real-world scenarios, where data features may vary due to factors, such as patient condition, activity level, or sensor positioning. ANFIS eliminates the need for manual feature engineering by automatically extracting relevant features from the input data. This capability can save time and effort during the preprocessing stage, enabling the model to effectively capture minute patterns in heartbeat signals that might not be immediately apparent to human observers (Equation I).

$$H(z)=\frac{\left(1-z^{-6}\right)^2}{\left(1-z^{-1}\right)^2} \quad \text{(I)}$$

The amplitude response is given by Equation II, where T is the period of sampling.

$$\left|H(WT)\right|=\frac{\sin^2(3wT)}{\sin^2(wT/2)} \quad \text{(II)}$$

The low-pass filter from Equation I is represented by the difference equation shown in Equation III.

$$y(nT) = 2y(nT-T)-y(nT-2T)+X(nT)-2X(nT-6T)+X(nT-12T) \quad \text{(III)}$$

For the low-pass filter, the cutoff frequency and gain are set at 11 Hz and 36, respectively, with a processing delay of six samples. A high-pass filter has also been designed, with its transfer function represented in Equation IV. The amplitude response is given in Equation V, and the difference equation is provided in Equation VI. Here, the low cutoff frequency is 5 Hz, the gain is 32, and the delay is 16 samples.

$$H(z)=\frac{\left(-1+32z^{-16}+z^{-32}\right)}{\left(1+z^{-1}\right)} \quad \text{(IV)}$$

$$\left|H(WT)\right|=\frac{\left[26+\sin^2(16wT)\right]^{\frac{1}{2}}}{\cos\left(\frac{wT}{2}\right)} \quad \text{(V)}$$

$$y(nT) = 32x(nT-16T)-[y(nT-T)+x(nT)-x(nT-32T)] \quad \text{(VI)}$$

After filtering, the five-point derivative, along with the transfer function (Equation VII) and the amplitude response (Equation VIII), is applied to differentiate the signal. Equation VII indicates the derivative operator. The derivative procedure gives a large gain to the high-frequency components resulting from the high slopes of the QRS complex while suppressing the low-frequency components of the P and T waves. The difference equation in Equation IX results in a nearly linear frequency response between DC and 30 Hz. Next, the signal is squared point by point, as indicated in Equation X. This squaring operation suppresses the small differences from the P and

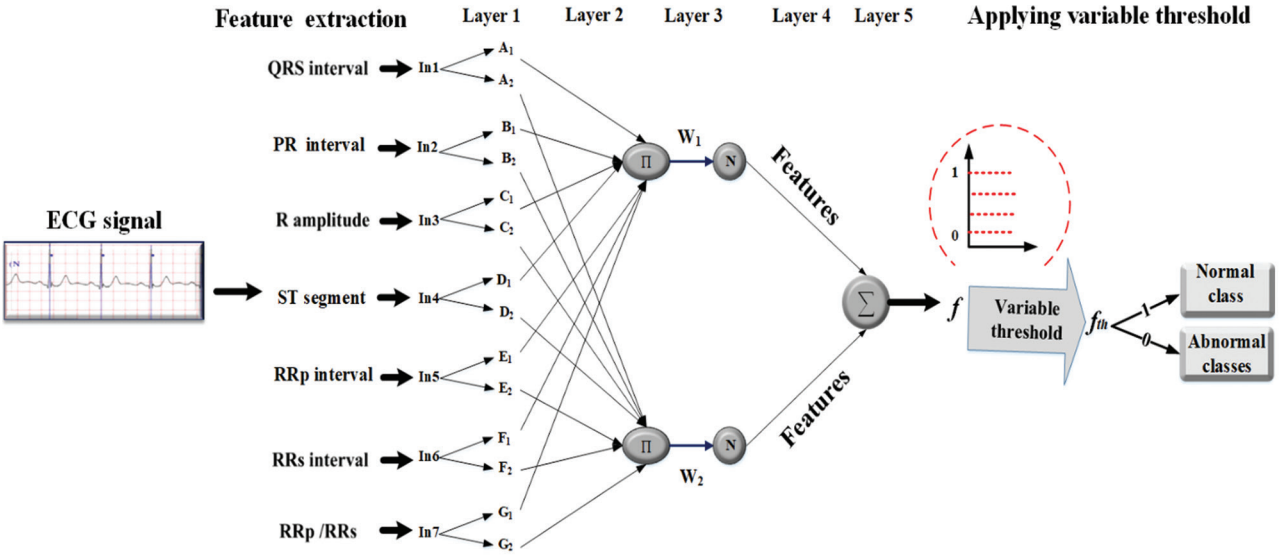


Figure 2. Proposed variable-threshold ANFIS for heartbeat classification (illustration by the authors)
Abbreviations: ANFIS: Adaptive neuro-fuzzy inference system; ECG: Electrocardiogram.

T waves, making the results positive and accentuating the larger differences from the QRS complexes. Consequently, the high-frequency components associated with the QRS complex are further amplified. This non-linear transformation involves squaring the signal samples individually.

$$H(z)=\left(\frac{1}{8}T\right)\left(-z^{-2}-2z^{-1}+2z^{-1}+z^{2}\right) \quad \text{(VII)}$$

$$\left|H(WT)\right|=\left(\frac{1}{4}T\left[\sin(2wT)+2\sin(wT)\right]\right) \quad \text{(VIII)}$$

$$y(nT)=\left(\frac{1}{8}T\right)\begin{bmatrix}-x(nT-2T)-2x(nT-T)\\+2x(nT+T)+x(nT+2T)\end{bmatrix} \quad \text{(IX)}$$

$$y(nT)=[x(nT)]^{2} \quad \text{(X)}$$

The moving-window integration described in Equation XI is used to capture the waveform feature information and the slope of the R wave:

$$y(nT)=\left(\frac{1}{N}\right)\begin{bmatrix}x(nT-(N-1)T)+\\x(nT-(N-2)T)+\ldots+x(nT)\end{bmatrix} \quad \text{(XI)}$$

Here, N represents the number of samples in the integration window, and Equation XI indicates the integration process. The moving window integrator is passed through by the squared waveform. This integrator advances one sample interval, integrates the new pre-defined interval window, and sums the area under the squared waveform over an appropriate interval. To capture the duration of extended abnormal QRS complexes, the half-width of the window has been set in a way that is sufficiently short to avoid overlapping a T-wave and a QRS complex at the same time. In addition to the R wave's slope, features are extracted using the moving average (MA) filter. The difference Equation XI is used to implement it. Usually, the width of the integration window is approximately equal to the widest possible QRS complex. In a wide integration window, the integration waveform merges the QRS and T complexes. Furthermore, in a narrow integration window, several peaks are produced in the integration waveform by some of the QRS complexes. In these conditions, the next QRS detection is difficult. The width of the window is specified by observation and experience. In the proposed algorithm, the sample rate is considered 200 samples/s, while the width of the window is 30 samples (150 ms). At first, the signal is analyzed by the high values of the two thresholds. The low thresholds are used for the case of no QRS detection, here, a search-back technique is used to refer to the previous time for the QRS complex.[27] Usually, the width of the integration window is approximately equal to the widest QRS complex. In a wide integration window, the integration waveform merges the QRS and T complexes. On the other hand, in a narrow integration window, several peaks are produced in the integration waveform by some of the QRS complexes. In these conditions, the next QRS detection is difficult. The width of the window is specified by observation and experience. In the proposed algorithm, the sample rate is considered 200 samples/s, while the width of the window is 30 samples

(150 ms). At first, the signal is analyzed using high values of the two thresholds. The low thresholds are used for the case of no QRS detection; here, a search-back technique is used to refer to the previous time for the QRS complex (Figures 3 and 4).[13,28]

Once the features are extracted, ANFIS is applied. The fuzzy rules between a set of inputs and outputs are generated by the Sugeno fuzzy model. The proposed ANFIS operates like a fuzzy inference system (FIS) and produces fuzzy rules and membership functions (MFs), with its output described by Equation XII:

$$\gamma = \sum_{i=1}^{L} \left\{ \frac{\left(\prod_{j=1}^{n} MF\left(x_j\right)\right)\left(z^i\right)}{\sum_{i=1}^{L}\left(\prod_{j=1}^{n} MF\left(x_j\right)\right)} \right\} \quad \text{(XII)}$$

where the membership function is denoted by MF, the j^{th} input is x_j (j = 1, 2,…, n), and the output of the j^{th} fuzzy rule is z^i.

Using neural networks and customized learning rules to reduce error, ANFIS modifies the Sugeno inference system's settings. One potential set of rules is shown in Equations XIII and XIV for a fuzzy system with two inputs (x and y), one output (z), and a fuzzy inference Sugeno model.

Rule 1: If x is equal to A_1, y is equal to B_1 then, $f_{1= p1\ x+q1\ y+r1}$ (XIII)

Rule 2: If x is equal to A_2, y is equal toB_2 then, $f2=p_2\ x+q_2\ y+r_2$ (XIV)

The fuzzification, rule, normalization, defuzzification, and summing layers make up the five layers of the Sugeno ANFIS network model. The degree of membership is determined for each input and the adaptive node represents the first layer (Equation XV).

$$O_{1,i=} = \mu_{A_i}\left(x\right) = \exp\left\{-\left(\frac{x - c_i}{a_i}\right)^2\right\} \quad \text{(XV)}$$

Figure 3. Block diagram of pan-Tompkins algorithm[27]
Abbreviation: ECG: Electrocardiogram.

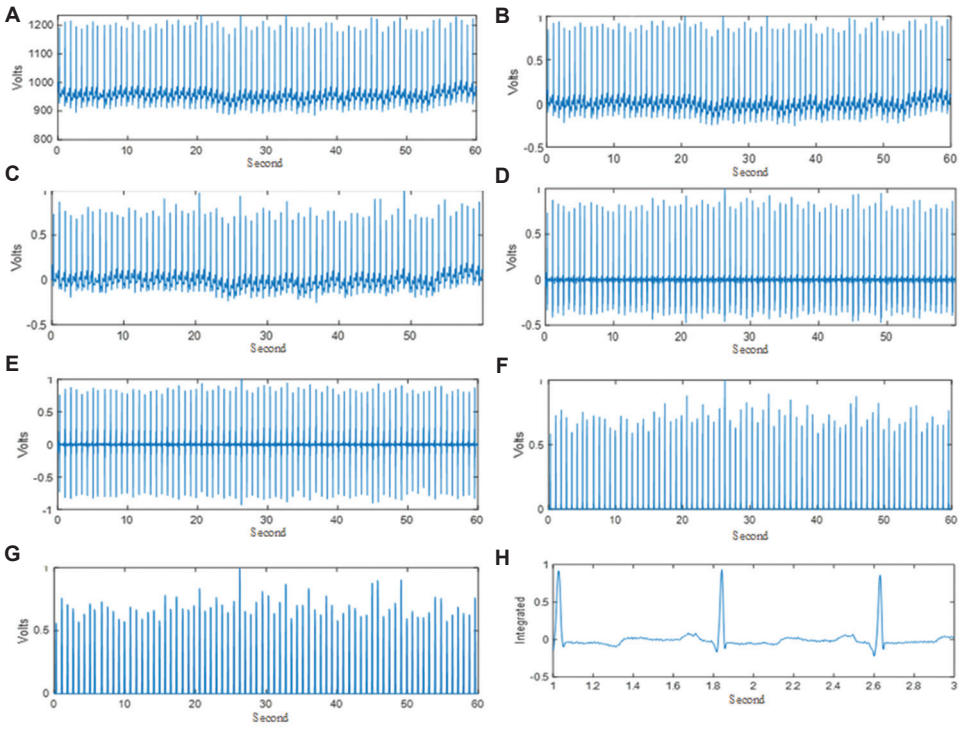


Figure 4. Pan-Tompkins algorithm steps. (A) Raw signal, (B) DC drift and normalization, (C) low-pass filter, (D) high-pass filter, (E) derivative filter, (F) squaring, (G) averaging, and (H) moving-window integration[27]

where $O_{1,i}$ is the membership function of A_i, x is the input to node i, A_i is the linguistic label associated with this node function, and it specifies the degree to which the given x satisfies the quantifier A_i. μ_{Ai} (x) is chosen to be bell-shaped, with a maximum equal to 1 and a minimum equal to 0, and the parameter set is $\{a_i, b_i, c_i\}$. Layer 2 (rule layer) consists of circle nodes labeled with numbers that multiply incoming signals and send the product out (Equation XVI).

$$O_{2,I} = w_i = \mu_{Ai}(x) \cdot \mu_{Bi}(y), i = 1, 2 \quad \text{(XVI)}$$

Every node in the normalization layer is represented by the circle node N. The outputs of this layer are sometimes known as normalized firing strengths. The i_{th} node determines the firing strength rate of rule i_{th} to the sum of all firing strength of rules, given by Equation XVII.

$$O_{3,i} = \overline{w_i} = \frac{w_i}{w_1 + w_2}, i = 1,2 \quad \text{(XVII)}$$

Each node of i in the defuzzification layer is a square node with a node function (Equation XVIII).

$$O_{4,i} = \overline{w}_i f_i = \overline{w}_i \left(p_i x + q_i y + r_i\right) \quad \text{(XVIII)}$$

Where $\{p_i, q_i, r_i\}$ is the set of parameters (consequence parameters) and $\overline{w}_i$ is the layer's output. The ANFIS summation layer comprises a single fixed node, denoted as Σ, which is responsible for preparing the final output, which is the summing of all signals (Equation XIX).

$$\text{Final output} = O_{5,1} = \sum_i \overline{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \quad \text{(XIX)}$$

ANFIS uses a back-propagation mechanism for the input membership function parameters to train a fuzzy system. In addition, the parameters and the output membership function are connected using the least mean squares (LMS) technique. The output of the nodes is sent to the defuzzification layer in the hybrid training algorithm's forward step. From there, the least-squares approach is used to identify the resultant parameters. The gradient descent method adjusts the initial parameters during the backward phase, which also sees the propagation of the error signal backward.

To optimize the fuzzy system and fuzzy rules, subtractive clustering is applied, which helps the ECG pattern detection. This method divides the data into clusters and creates a FIS with a minimum number of rules; then, the fuzzy qualities are related to the respective cluster. Indeed, extracted features are given to the ANFIS in a subtractive clustering way, in which there are no particular restrictions; one only needs to pay attention to the influence radius because of the main element in determining the number of clusters.

The scale of the issue is not significant when differential clustering is used, where processing is proportionate to the quantity of data points. An M-dimensional space containing n data points $\{x_1, \ldots, x_n\}$ is examined, with each point normalized to a hypercube. Since every data point has the potential to be the cluster center, Equation XX determines the density measurement at every point x_i.

$$D_i = \sum_{j=1}^{n} \exp\left(-\frac{\|x_i - x_j\|^2}{\left(r_a/2\right)^2}\right) \quad \text{(XX)}$$

The neighborhood is defined by the radius r_a, a positive constant and the points outside of the neighborhood have relatively little impact on the density measurement D_i. The point with the highest density is chosen to be the first cluster's center once the D_i has been computed for each of the points. The measured density for each point is updated in accordance with Equation XXI, where x_{C1} is the selected point, and D_{C1} is the density value.

$$D_i = D_i - D_{C1} \exp\left(-\frac{\|x_i - x_{C1}\|^2}{\left(r_b/2\right)^2}\right) \quad \text{(XXI)}$$

The density of each point is reviewed, and the subsequent center x_{C2} is chosen; then, all of the density measures of the points are revised again. This mechanism is repeated to attain an adequate number of clusters. When employing the subtractive clustering technique for a collection of input-output data, each cluster center represents a prototype that exhibits certain characteristics of the modeled system. These centers are used as centers of the premises of the fuzzy rules during a zero-order Sugeno model.[29,30]

Consequently, ECGs are processed through classifiers, and the proposed VTMA ultimately classifies normal and abnormal beats (Figures 5 and 6). As ANFIS functions as a binary classifier, six ANFISs are implemented and then trained, validated, and tested. Outputs are divided into six categories, including NSR, LBBB, RBBB, PVC, APC, and PB. For example, the RBBB heart rate is defined as "1," and the other five types of heart rate are defined as "0." The VTMA is capable of classifying six types of heartbeats using fuzzy logic and neural learning. In this method, seven features, as mentioned before, are used as the system inputs. The values of these characteristics for each type of heartbeat are adapted from reliable medical sources and used in the implementation process. As the results of the determined classes are not accurate enough using just

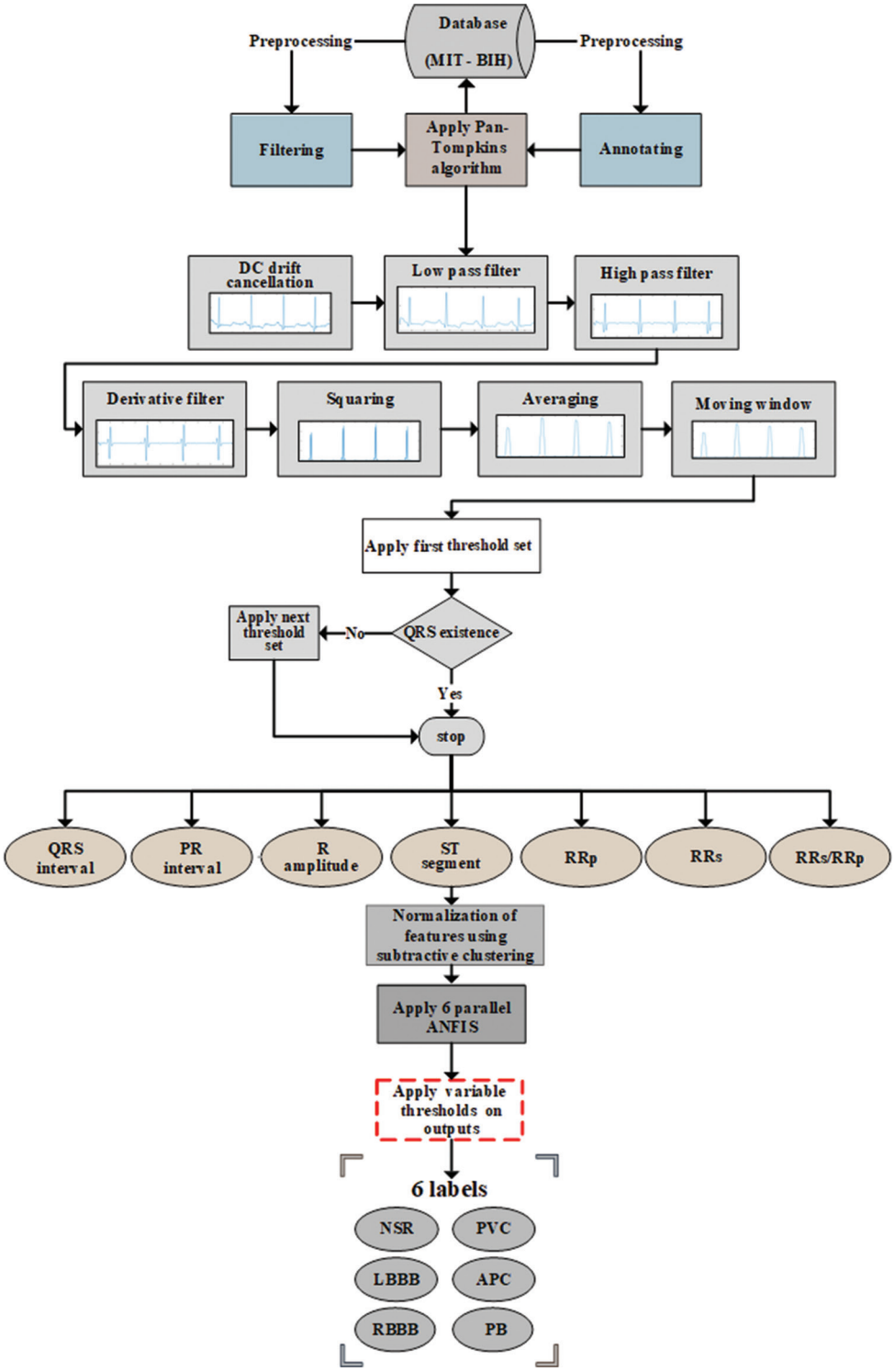


Figure 5. Workflow supporting the proposed methodology (illustration by the authors)
Abbreviations: ANFIS: Adaptive neuro-fuzzy inference system; APC: Atrial premature condition; LBBB: Left bundle branch block; NSR: Normal sinus rhythm; PB: Paced beat; PVC: Premature ventricular contraction; RBBB: Right bundle branch block.

ANFIS, a new technique (VTMA) is added to the ANFIS to attain an accurate ANFIS classifier. A variable threshold (not constant as in previous works) is added at each ANFIS output to classify the heart rate as a specific heart rate.

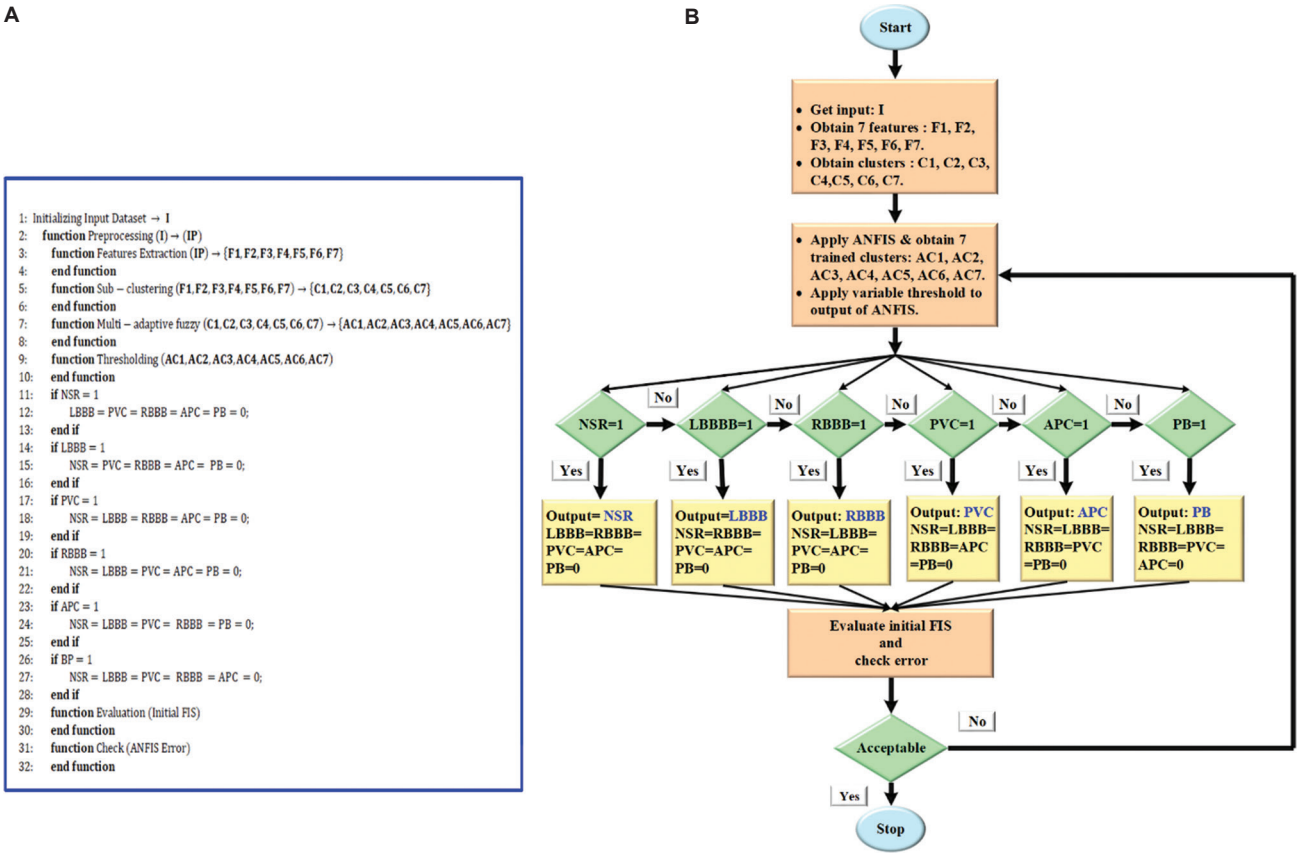


Figure 6. The classification mechanism of the proposed VTMA**.** (A) Algorithmic structure and (B) schematic representation of the proposed methodology (illustration by the authors)
Abbreviations: ANFIS: Adaptive neuro-fuzzy inference system; APC: Atrial premature condition; LBBB: Left bundle branch block; NSR: Normal sinus rhythm; PB: Paced beat; PVC: Premature ventricular contraction; VTMA: Variable-threshold multi-adaptive neuro-fuzzy system.

The VTMA is applied to all parallel ANFIS structures. A specific heart rate is denoted by "1," and the other heart rates are denoted by "0." The threshold of each ANFIS determines whether the specific heart rate is the target, "1," or not. The "0" output of the VTMA shows no target has happened. The V_{th} is a variable threshold determined by trial and error and given by Equation XXII; "f" shows the ANFIS output, and f_{th} indicates the output value after the threshold, "0" or "1." The range of changing threshold is [0 1, and the best results are seen in the interval of [0.4 0.6].

$$f_{th} = \begin{Bmatrix} 0 & if & f < V_{th} \\ 1 & if & f \geq V_{th} \end{Bmatrix} \quad \text{(XXII)}$$

The threshold mechanism of the proposed ANFIS is used to remove inaccuracies. It makes the multi-ANFIS classifier more certain to be labeled correctly. Accuracy is a criterion to evaluate the classification process, and in some cases adding a threshold increases the accuracy. The threshold maps all values into two classes. The used variable threshold solves the problems of severe imbalance classes, where the default threshold causes poor classification. On the other hand, the variable threshold in the ANFIS tunes imbalanced classification problems and maps probabilities to class labels. In this research, by using a variable fuzzy level threshold technique, the optimal threshold is set in such a way that probabilities are converted to class labels, imbalanced classification is performed with high accuracy, and the optimal receiver operating characteristic (ROC) and precision-recall curves result.

4. Results and discussion

The MIT-BIH arrhythmia database is used for training and performance evaluation of the proposed VTMA classifier. This database contains 48 half-hour excerpts from dual-channel outpatient ECG records of 47 individuals. The sampling frequency of the recordings is 360 samples per second, with 11 bits resolution, and the amplitude range is 10 mV. Each record of the database may contain only one or more specific types of beats, and it does not necessarily have all types of beats because the recordings belong to different people over 30 min. On the other hand,

selecting which records to use in the algorithm is based on the number of specific beats in that record. Table 1 shows the input properties by input numbers that are important for discussing the FIS membership functions. For diminutive clusters, for each input, a cluster (Gaussian membership functions) is created and the diminutive clustering algorithm normalizes the input properties. The normalization layer of the ANFIS normalizes the network weights. In Table 1, the left column shows different heartbeats, NSR, LBBB, RBBB, PVC, APC, and PB. The first row lists extracted features. NF means no information is available about that case. Six different heart conditions have their specific values in terms of temporal and amplitude characteristics. Amplitude features are in millivolts, and temporal features are in both seconds and milliseconds.

For NSR, both RRs and RRp intervals are the same, so their ratio is equal to 1; here, the PR interval is longer than the QRS interval. In LBBB, RBBB, PVC, and paced, the QRS interval is the same and it is more than 120 ms. There is no specific information about the PR interval of LBBB, RBBB, PVC, and APC. APC and PB are in common in terms of R amplitude. It seems that the ST segment does not play much significant role in classification because, in four heart conditions, there is no sign of this feature; however, it improves the simulated results. RR features (RRs, RRp, and RRs/RRp) are not prominent factors in PB. ECG properties are extracted in 30 min for about 646400 samples. Figure 7 shows the results of feature extraction algorithms on the recordings of the ECG on which the heart condition is to be diagnosed. These signals can be seen from a modified limb lead II with 360 samples per second of approximately 646,400 samples. Six types of heartbeats are labeled as "N," "L," "R," "V," "A," and "/" to represent the various conditions. At this stage, all types of beats are identified, along with features such as the P-wave, QRS complex, and T-wave, including their onset and offset. The duration of this record is long and the detail of these annotations is not clear enough. Hence, it is better to separate one random heartbeat as below. Histograms of NSR characteristics are plotted to verify whether the features of Table 1 are extracted correctly from the annotation function or not. Here, histograms of the other five types of heartbeats are avoided, and only diagrams subject to the normal heart rate of one of the records are given in Figure 8. Histograms are helpful to see the data distribution and to show the differences in the outputs. Histograms of Figure 8E-K are approximately normal, while Figure 8A-D, and L are nearly right-skewed distributions. As the features are different in terms of measurement units, the horizontal axis differs in these histograms. For the histogram of the extracted QRS feature, the center of data is located in 0.1 s, and most bins are devoted to this histogram. ST intervals, QT intervals, and ST segments have a peak of about 50 bins. The peaks of P amplitudes, S offset amplitudes, and RR ratios are about 40 bins. PR intervals, P wave intervals, and Q onset amplitudes have approximately the same height (about 100 bins). PR segments and R amplitudes reach 60 bins. The centers of histograms in Figure 8B, C, I, J, and K are between 0 and 0.5 s. Q onset and S offset amplitudes have negative amplitude in the duration of [−1 0] millivolt. However, S offset amplitudes are higher than Q onset amplitudes. Among these extracted features, seven of them are more efficient to go through ANFIS as inputs for classification effectively. The selected features for this research are PR segment, P wave interval, P amplitude, Q onset amplitude, S offset amplitude, T wave interval, QT interval, and ST interval. The subtractive clustering generates an initial FIS for each ANFIS. Then, the central parameters of the Gaussian function and standard deviation are adapted by the ANFIS. The number of fuzzy membership functions affects the ANFIS model; fewer membership functions cause less complexity and a lesser run-time. In Figure 9, the initial FIS for NSR ANFIS is shown.

When Gaussian membership functions are used, the transitions between membership values are smooth and continuous. The model can detect minute variations since the input data are smooth, which is particularly useful for detecting variations in heartbeat signals that can result from noise or heart rate variability, among other factors.

Table 1. The input feature range of an electrocardiogram signal[21]

ANFIS	QRS interval (ms)	PR interval (ms)	R amplitude (mV)	ST segment (ms)	RRp interval (s)	RRs interval (s)	RRs/RRp
NSR	80 – 100	120 – 200	1.5 – 2	80 – 120	0.6 – 1.2	0.6 – 1.2	1
LBBB	>120	NF	NF	NF	NF	NF	NF
RBBB	>120	NF	NF	>120	NF	NF	NF
PVC	>120	NF	<2	NF	<0.6	>1.2	>1
APC	<80	NF	>2	NF	<0.6	>1.2	>1
PB	>120	>280	>2	NF	NF	NF	NF

Abbreviations: ANFIS: Adaptive neuro-fuzzy inference system; APC: Atrial premature condition; FN: False negative; FP: False positive; LBBB: Left bundle branch block; NF: No feature; NSR: Normal sinus rhythm; PB: Paced beat; PVC: Premature ventricular contraction.

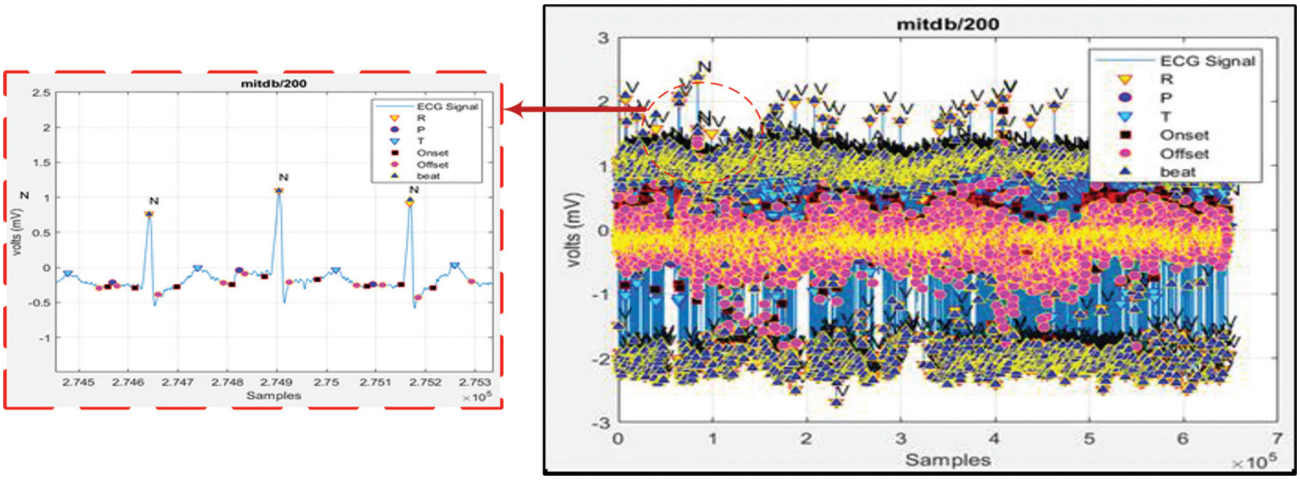


Figure 7. Feature annotating on the electrocardiogram, including feature extraction over 30 min and a closer-up view (illustration by the authors)

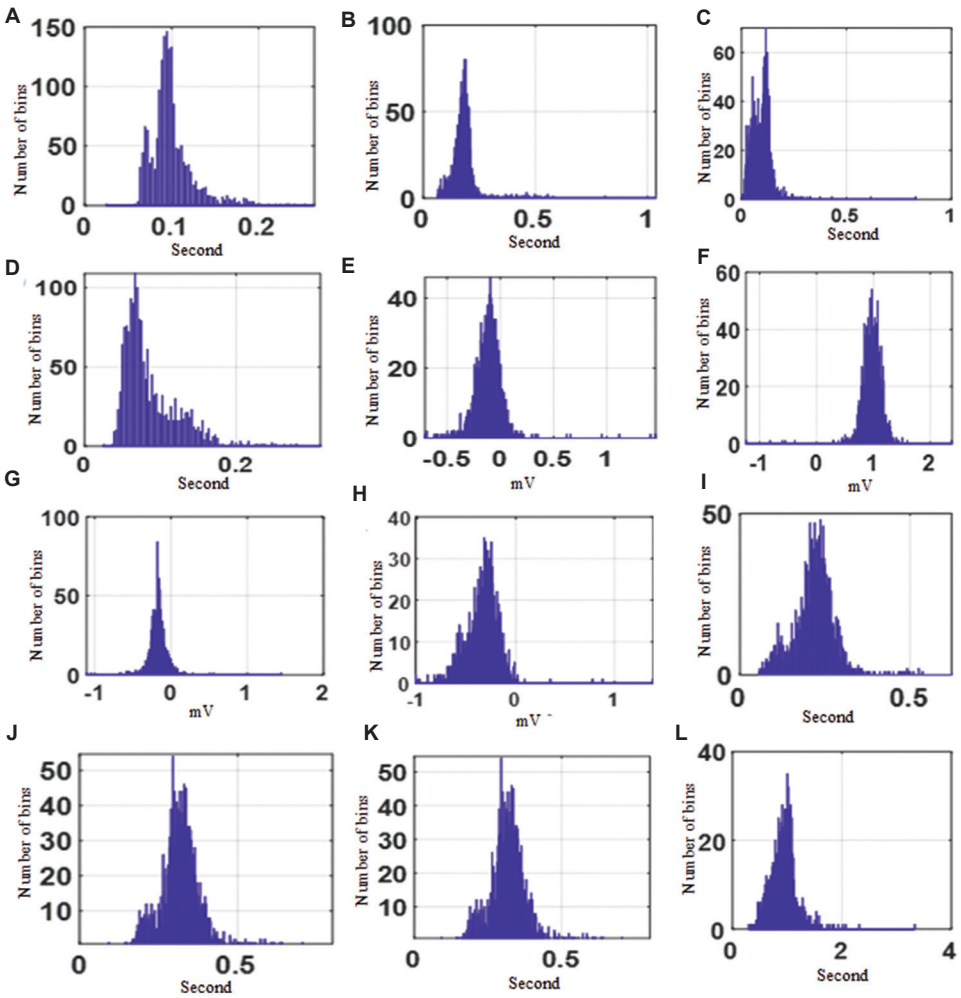


Figure 8. Features histograms of NSR ECG record (200). (A) QRS intervals. (B) PR intervals. (C) PR segments. (D) P wave intervals. (E) P amplitudes. (F) R amplitudes. (G) Q Onset amplitudes. (H) S offset amplitudes. (I) ST intervals. (J) QT intervals. (K) ST segments. (L) RR ratio (illustration by the authors)

Abbreviations: ECG: Electrocardiogram; NSR: Normal sinus rhythm.

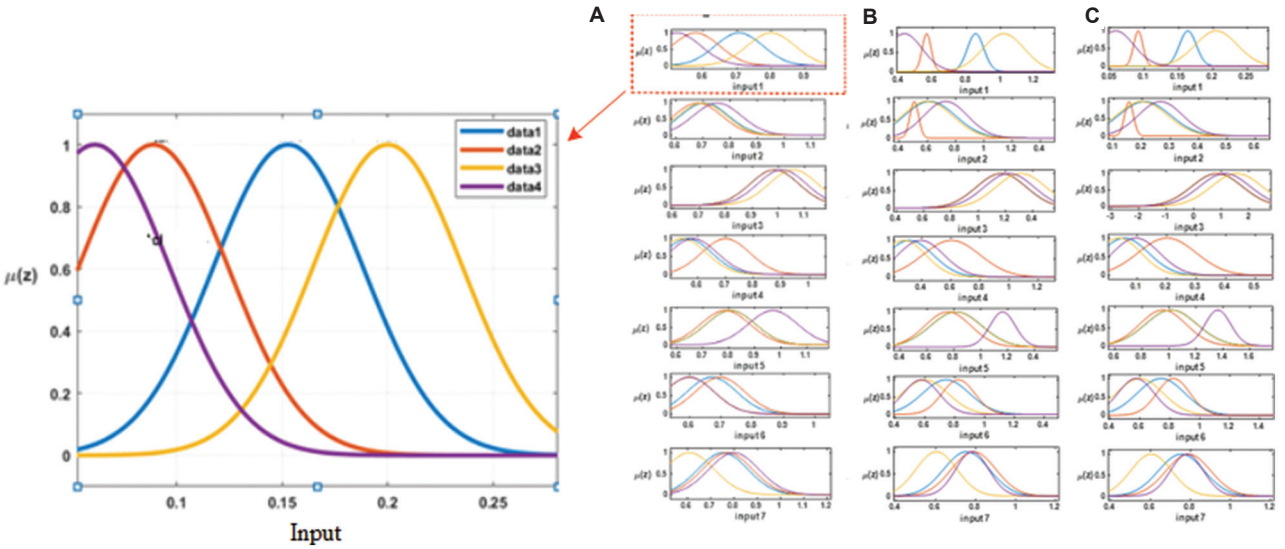


Figure 9. Membership functions for NSR ANFIS, (A) initial membership functions, (B) adapted FIS, and (C) checking of FIS membership functions (illustration by the authors)
Abbreviations: ANFIS: Adaptive neuro-fuzzy inference system; FIS: Fuzzy inference system; NSR: Normal sinus rhythm.

Because Gaussian membership functions are limited around their centers, they can focus on specific regions of the input space.

In this study, for each input, there are four clusters, so there are four membership functions for each input. These clusters are created by supported influence radius through the subtractive clustering method.

Throughout the research, all efforts are applied to stay determined four clusters, so tuning this radius is at the top of this research and defined by the user. In the subtractive method, the user-specified radius (range of influence) determines the number of membership functions, so it is necessary to make an intelligent decision for this parameter. Since the number of iterations is 2000, the initial FIS generated by subtracting clustering is generated 2000 times in the ANFIS. Six hundred samples are used for the evaluation of the proposed system. So for each type of output (heart condition), 100 samples are chosen randomly, 55% of them are devoted to training data, 35% is allocated to testing evaluation, and the rest of the data (10%) is for checking or validation. Therefore, for each specific heartbeat type, 330 samples are used for the training process divided into two parts: normal (specific) and abnormal (non-specific). Normal beat conveys the meaning of those data, which are the desired ones; for example, for PVC detection of training, 55 samples are specific, and the others 275 are non-specific. These all are true for both checking and testing data. Here, normal and abnormal heartbeats can also be called "beat" and "not-beat," respectively. As a case study, Figure 10 shows the step curve, which records the step size during the first VTMA training. This step size index acts as a reference for setting the initial step size and the amount of increase and decrease of the corresponding step size. The index is usually a curved step size that first increases, reaches a maximum, and then decreases for the rest of the training. Figure 11 illustrates the decision surface obtained from the fourth ANFIS (PVC) as a case study. This level is created between the first input (QRS interval) and the second input (PR interval). Twenty-one different levels can be viewed because there are seven inputs. This is one of the advantages of ANFIS because the user can interpret this figuration as ANFIS-based input mapping.

For the training data, the root-mean-square error (RMSE) is calculated for each sample. Decreasing RMSE or convergence is desirable in increasing the number of iterations. Figure 12 expresses the RMSE for six clusters. RMSE compares the desired output value (y) with the actual FIS output ($\check{y}$). "t" indicates the number of heartbeats. RMSE can be expressed as Equation XXIII.

$$\mathrm{RMSE} = \sqrt{\frac{1}{t}\sum_{i=1}^{t}\left(\mathrm{y_i} - \widehat{\mathrm{y_i}}\right)^2} \qquad \text{(XXIII)}$$

When comparing a model's FIS output to the observed data, RMSE provides a more accurate measure of error. The RMSE is a desirable value because of statistical features, such as variance and standard deviation. When increasing the number of iterations, a decreasing RMSE or convergence is preferred. Testing of the data is evaluated using trained FIS in ANFIS evaluation and the classification is performed.

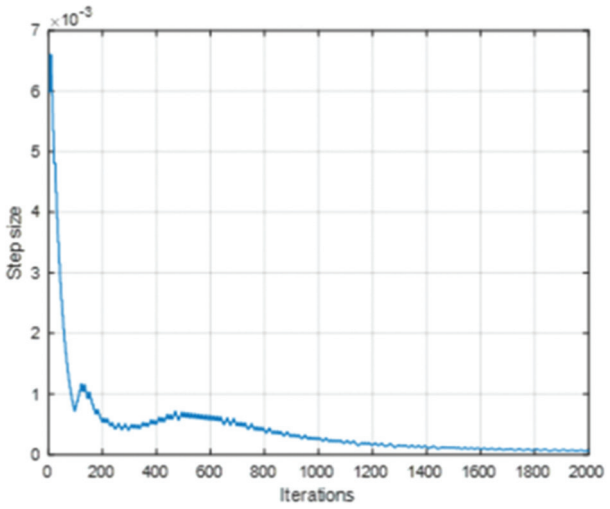


Figure 10. Step size curve of NSR for the proposed VTMA (illustration by the authors)
Abbreviations: NSR: Normal sinus rhythm; VTMA: Variable-threshold multi-adaptive neuro-fuzzy system.

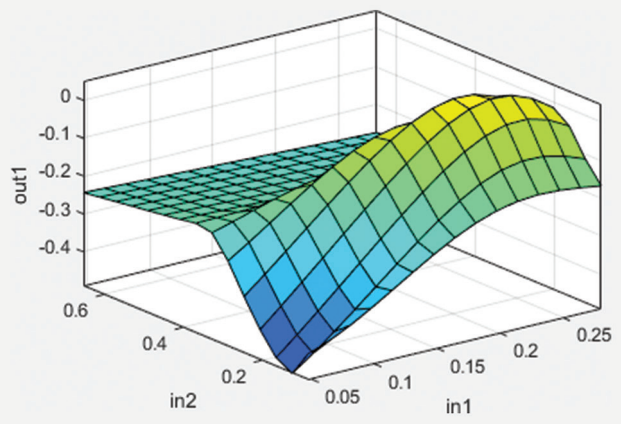


Figure 11. Fuzzy surface representation between the first and second inputs and the output of the system (illustration by the authors)

The train, check, and test data›s output vector is subjected to a threshold, which is covered in the following sections, changing the fractional numerical assignments for each heartbeat to "0" or "1."

After applying ANFIS to data, classification is done, as shown in Figure 13. At first, the threshold has not been applied, so the classification is not performed accurately, and then by using a variable threshold on the obtained classes, more accurate classifications result. The data aligned with the number one ("1") in the vertical axis represents the normal classified data, and the zero-point data ("0") from the vertical axis represents the correctly classified abnormal data. For example, if APC is the desired beat, all other beats (NSR, LBBB, RBBB, PVC, and PB) are defined as abnormal beats. In Figure 13, the sub-figures are depicted with different markings to show the performance clearly. Blue circular shows the actual data, however, they are covered by classified data. Red cross points represent the classified data (accurate or inaccurate) before applying the variable thresholds. At the same time, black star points demonstrate checking classification after implementing the variable threshold. From Figure 13**,** it is concluded that the application of the threshold in the NSR and RBBB systems is insignificant because they are accurate enough due to their distinguished characteristics. The LBBB and PVC beats are similar in terms of some characteristics, so it is demanding for the ANFIS to distinguish these beats 100% accurately thus, the accuracy of classification is somewhat lower than other beats; this fact is more obvious in PVC beat classification, and it can be seen in simulation results. The APC classification is done efficiently, and applying variable threshold plays a prominent role in the classification of these kinds of beats. In PB beats, the effective role of applying threshold is illustrated, and there is no irregularity in classification. As a result, PVC, APC, LBBB, and PB classification have improved dramatically and these changes convey the meaning of the strong impact of putting a variable threshold on the six ANFIS outputs.

The effectiveness of classification is assessed using performance parameters such as accuracy, sensitivity (recall), specificity, precision, and F_1-score (dice), which are represented by **Equations** XXIV–XXVIII, respectively. Heart rate serves as a proxy for true positive (TP), true negative (TN), false positive (FP), and false negative (FN) in all five measurements.

$$\text{Accuracy}(\%) = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \times 100\% \quad \text{(XXIV)}$$

$$\text{Sensitivity}(\%) = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100\% \quad \text{(XXV)}$$

$$\text{Specificity}(\%) = \frac{\text{TN}}{\text{TN}+\text{FP}} \times 100\% \quad \text{(XXVI)}$$

$$\text{Precision}(\%) = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100\% \quad \text{(XXVII)}$$

$$F_1 - \text{score}(\%) = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} \times 100\% \quad \text{(XXVIII)}$$

Based on Tables 2-4**,** the most common metrics—accuracy, sensitivity, specificity, precision, and F_1-score—demonstrate the high performance of the proposed classification method across all training, testing, and

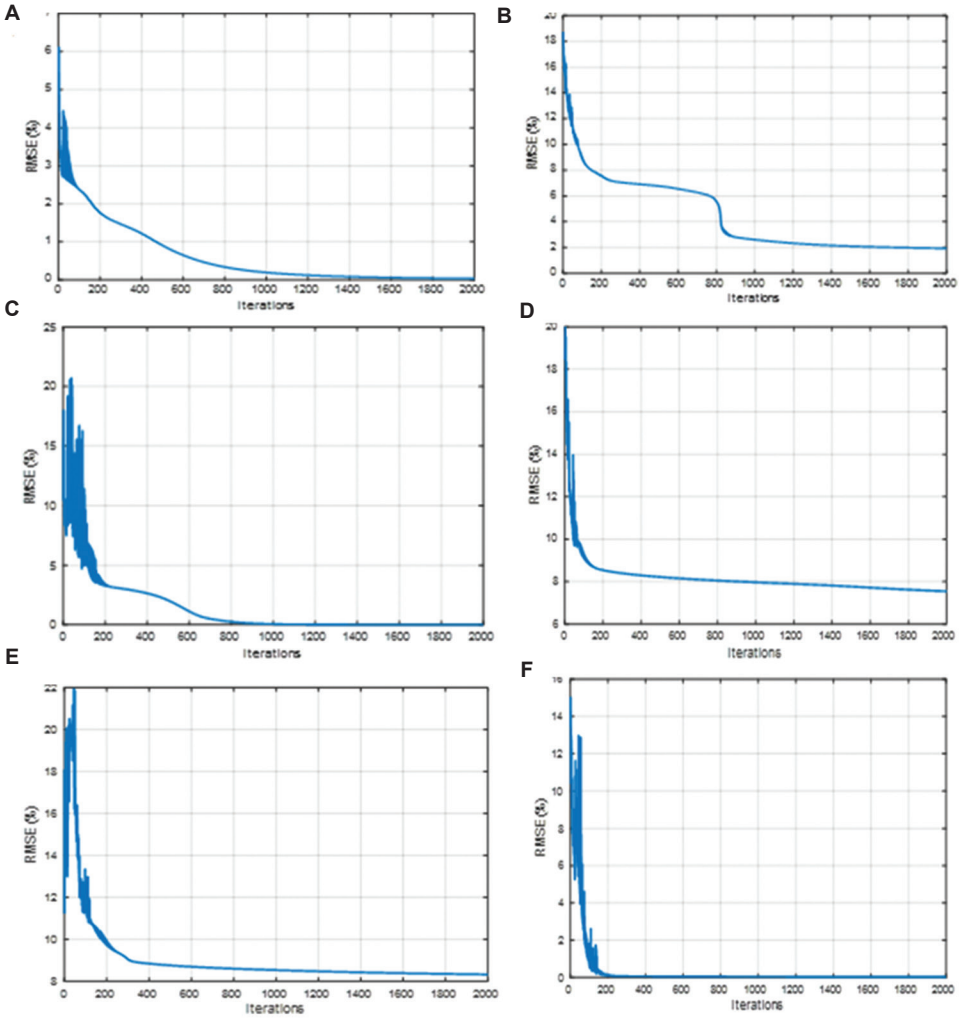


Figure 12. RMSE for six ANFISs. (A) NSR. (B) LBBB. (C) RBBB. (E) PVC. (E) APC. (F) PB (illustration by the authors)
Abbreviations: ANFIS: Adaptive neuro-fuzzy inference system; RMSE: Root-mean-square error; APC: Atrial premature condition; LBBB: Left bundle branch block; NSR: Normal sinus rhythm; PB: Paced beat; PVC: Premature ventricular contraction.

validation datasets. From Table 2, we observe that training for NSR using VTMA achieved 100% accuracy, sensitivity, specificity, precision, and F_1-score. For LBBB, the metrics were 99.39% accuracy, 100% sensitivity, 99.28% specificity, 96.36% precision, and 98.15% F_1-score. The RBBB achieved 100% accuracy, sensitivity, specificity, precision, and F_1-score, while the PVC achieved accuracy of 98.79%, sensitivity of 100%, specificity of 98.57%, precision of 92.73%, and F_1-score of 96.23%. The APC achieved an accuracy of 99.39%, sensitivity of 98.18%, specificity of 99.64%, precision of 98.18%, and F_1-score of 98.18%, while the PB using VTMA achieved 100% accuracy, sensitivity, specificity, precision, and F_1-score. It is evident that NSR, RBBB, and PB in the VTMA achieved perfect performance with 100% accuracy across all types of beats and datasets.

To prevent repetition, we have not included detailed explanations of Tables 3 (testing data) and 4 (validation data), as they present similar results in a different context, as illustrated in the corresponding tables. The NSR, RBBB, APC, and PB are classified with 100% accuracy. However, the similarity between PVC heartbeats and LBBB makes it challenging for the diagnostic system to distinguish between them, resulting in slightly lower accuracy for these two classifications. Numerous studies have explored ECG classification techniques. Table 5 compares some recent publications [1], [2], [4], [20], [21], [25], and [26] with the proposed method, showing that the VTMA technique outperforms these counterparts. In terms of accuracy, our method outperforms those in [1], [4], [20], [25], and [26]. Regarding sensitivity, this method performs better than [1], [2], and [26], while references [4] and [21] did not

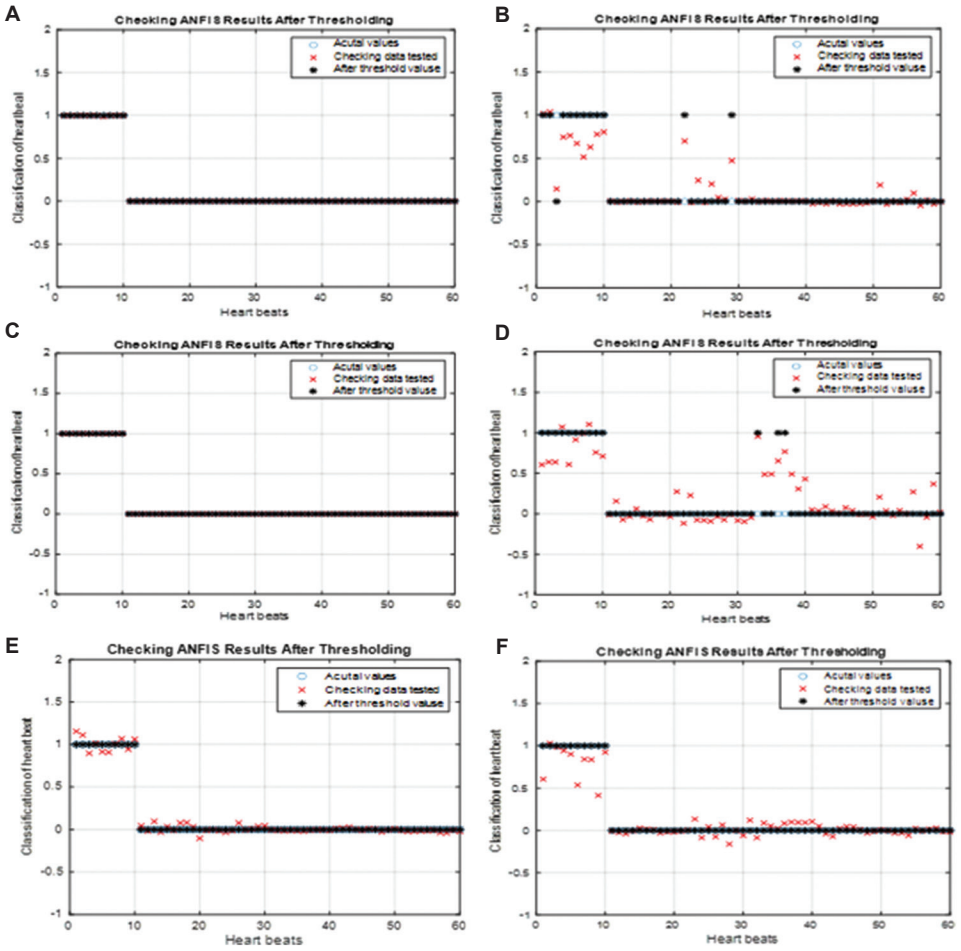


Figure 13. Simulation results of six systems under subtractive clustering before/after applying the variable threshold for check data classifications (illustration by the authors). (A) NSR. (B) LBBB. (C) RBBB. (E) PVC. (E) APC. (F) PB

Abbreviations: ANFIS: Adaptive neuro-fuzzy inference system; APC: Atrial premature condition; LBBB: Left bundle branch block; NSR: Normal sinus rhythm; PB: Paced beat; PVC: Premature ventricular contraction.

Table 2. Performance evaluation results of training data using the VTMA classification method

ANFIS	TP	TN	FP	FN	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F_1-score (%)
NSR	55	275	0	0	100	100	100	100	100
LBBB	53	275	2	0	99.39	100	99.28	96.36	98.15
RBBB	55	275	0	0	100	100	100	100	100
PVC	51	275	4	0	98.79	100	98.57	92.73	96.23
APC	54	274	1	1	99.39	98.18	99.64	98.18	98.18
PB	55	275	0	0	100	100	100	100	100

Abbreviations: ANFIS: Adaptive neuro-fuzzy inference system; APC: Atrial premature condition; FN: False negative; FP: False positive; LBBB: Left bundle branch block; NSR: Normal sinus rhythm; PB: Paced beat; PVC: Premature ventricular contraction; TN: True negative; TP: True positive.

provide sensitivity evaluations. The specificity of the VTMA method is more than that of [2], [20], [25], and [26], matching the performance of [21]. There was no specificity assessment in [1], [4], [25], and [26]. Represented methodology is more precise than [1] and [2]. However, [4], [20], [21], [25], and [26] did not report on precision. The F_1-score metric of this paper is higher than those reported in [1], [2], and [4], while [20] and [21] did not access the F_1-score. The classes covered by the mentioned method outnumber those in [2], [4], [20], [25], and [26], with papers [1] and [21]

Table 3. Performance evaluation results of testing data using the VTMA classification method

ANFIS	TP	TN	FP	FN	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	*F1*-score (%)
NSR	35	175	0	0	100	100	100	100	100
LBBB	34	172	1	3	98.09	91.89	99.42	97.14	94.44
RBBB	35	175	0	0	100	100	100	100	100
PVC	33	172	2	3	97.62	91.67	98.85	94.29	92.96
APC	35	175	0	0	100	100	100	100	100
PB	55	275	0	0	100	100	100	100	100

Abbreviations: ANFIS: Adaptive neuro-fuzzy inference system; APC: Atrial premature condition; FN: False negative; FP: False positive; LBBB: Left bundle branch block; NSR: Normal sinus rhythm; PVC: Premature ventricular contraction; TN: True negative; TP: True positive.

Table 4. Performance evaluation results of checking data using the VTMA classification method

ANFIS	TP	TN	FP	FN	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision(%)	F_1-score (%)
NSR	10	50	0	0	100	100	100	100	100
LBBB	9	48	1	2	95	81.82	97.96	90	85.71
RBBB	10	50	0	0	100	100	100	100	100
PVC	10	47	0	3	95	76.92	100	100	86.96
APC	10	50	0	0	100	100	100	100	100
PB	10	50	0	0	100	100	100	100	100

Abbreviations: ANFIS: Adaptive neuro-fuzzy inference system; APC: Atrial premature condition; FN: False negative; FP: False positive; LBBB: Left bundle branch block; NSR: Normal sinus rhythm; PB: Paced beat; PVC: Premature ventricular contraction; TN: True negative; TP: True positive.

Table 5. Comparison of the proposed methods and related works

Performance evaluation	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F_1-score (%)	Number of classes
[1]	TERMA- FrFT	82.2	84.25	-	89.03	80.23	6
[2]	DCNN	98.63	92.41	99.06	92.86	92.63	5
[4]	CNN	89.3	-	-	-	89.10	1
[20]	ANFIS	97.75	97.75	99.25	-	-	4
[21]	ANFIS	98.39	-	99.67	-	-	6
[25]	CNN-RNN	95.90	95.90	96.34	-	-	5
[26]	2-D CNN	97.3	89.3	98	-	-	2
This work	VTMA	98.33	93.12	99.66	98.33	95.44	6

Abbreviations: ANFIS: Adaptive neuro-fuzzy inference system; CNN: Convolutional neural network; DCNN: Deep convolutional neural network; RNN: Recurrent neural network; TERMA- FrFT: Two event-related moving averages-fractional Fourier transform; VTMA: Variable-threshold multi-adaptive neuro-fuzzy system.

also covering six classes, the same as our proposed method. Overall, the proposed work outperforms the mentioned studies, as presented in Table 5.

5. Conclusion

This research presents a VTMA for detecting ECG abnormalities. The process begins with selecting ECG records for classification. Next, the ECG signals undergo normalization using a subtractive clustering algorithm. To pre-process the data, low-pass and high-pass filters are applied to eliminate noise. Following this, input properties are extracted and prepared for the neural-fuzzy system through subtractive clustering. Each member of the output vector is assigned a specific heart rate, with the output pulses divided into six categories. An input data matrix of properties is generated and joined to an output vector, resulting in a system that classifies six types of heartbeats using fuzzy logic and neural learning.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: All authors
Investigation: Roghayeh Rafieisangari
Methodology: Nabiollah Shiri
Writing – original draft: Roghayeh Rafieisangari
Writing – review & editing: Nabiollah Shiri

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The dataset used in this study, the MIT-BIH Arrhythmia Database, was expanded on February 24, 2005, and is now freely available on PhysioNet. The dataset can be accessed through the following link: MIT-BIH Arrhythmia Database.

References

1. Aziz S, Ahmed S, Alouini MA. ECGbased machinelearning algorithms for heartbeat classification. *Sci Rep.* 2021;11:18738.

   doi: 10.1038/s41598-021-97118-5

2. Khan F, Yu X, Yuan Z, Rehman A. ECG classification using 1-D convolutional deep residual neural network. *PLOS One*. 2023;18:e0284791.

   doi: 10.1371/journal.pone.0284791

3. Weimann K, Conrad TO. Transfer learning for ECG classification. *Sci Rep.* 2021;11:5251.

   doi: 10.1038/s41598-021-84374-8

4. Cheng J, Zou Q, Zhao Y. ECG signal classification based on deep CNN and BiLSTM. *BMC Med Inform Decis Mak*. 2021;21:365.

   doi: 10.1186/s12911-021-01736-y

5. Sumathi S, Lilly Beaulah H, Vanithamani R. A wavelet transform based feature extraction and classification of cardiac disorder. *J Med Syst*. 2014;38:98.

   doi: 10.1007/s10916-014-0098-x

6. Leite JP, Moreno RL. Heartbeat classification with low computational cost using Hjorth parameters. *IET* Signal *Process.* 2018;12(4):431-438.

   doi: 10.1049/iet-spr.2017.0296

7. Varatharajan R, Manogaran G, Priyan MK. A big data classification approach using LDA with an enhanced SVM method for ECG signals in cloud computing. *Multimedia Tools Applic*. 2018;77:10195-10215.

   doi: 10.1007/s11042-017-5318-1

8. Sharma P, Ray KC. Efficient methodology for electrocardiogram beat classification. *IET* Signal *Process*. 2016;10(7):825-832.

   doi: 10.1049/iet-spr.2015.0274

9. Übeyli ED. Adaptive neuro-fuzzy inference system for classification of ECG signals using Lyapunov exponents. *Comput Methods Programs Biomed*. 2009;93:313-321.

   doi: 10.1016/j.cmpb.2008.10.012

10. Zhang L, Peng H, Yu C. An approach for ECG Classification Based on Wavelet Feature Extraction and Decision Tree. In: 2010 International Conference on Wireless Communications and Signal Processing. 2010. p. 1-4.

    doi: 10.1109/WCSP.2010.5633782

11. Karpagachelvi S, Arthanari M, Sivakumar M. Classification of electrocardiogram signals with support vector machines and extreme learning machine. *Neural Comput Applic*. 2012;21:1331-1339.

    doi: 10.1007/s00521-011-0572-z

12. El-Saadawy H, Tantawi M, Shedeed HA, Tolba MF. Hybrid hierarchical method for electrocardiogram heartbeat classification. *IET Signal Process.* 2018;12(4):506-513.

    doi: 10.1049/iet-spr.2017.0108

13. Chikh MA, Ammar M, Marouf R. A neuro-fuzzy identification of ECG beats. *J Med Syst*. 2012;36:903-914.

    doi: 10.1007/s10916-010-9554-4

14. Gupta A, Thomas B, Kumar P, Kumar S, Kumar Y. Neural Network based Indicative ECG Classification. In: 2014 5th International Conference - Confluence The Next Generation Information Technology Summit. 2014. p. 277-279.

    doi: 10.1109/CONFLUENCE.2014.6949262

15. Banerjee S, Mitra M. Application of cross wavelet transform for ECG pattern analysis and classification. *IEEE Trans Instrument Measure*. 2014;63(2):326-333.

    doi: 10.1109/TIM.2013.2279001

16. Bouaziz F, Oulhadj H, Boutana D, Siarry P. Automatic ECG arrhythmias classification scheme based on the conjoint use of the multi-layer perceptron neural network and a new improved metaheuristic approach. *IET Signal Process*. 2019;13(8):726-735.

    doi: 10.1049/iet-spr.2018.5465

17. Sahoo S, Kanungo B, Behera S, Sabut S. Multiresolution wavelet transform based feature extraction and ECG classification to detect cardiac abnormalities. *Measurement*. 2017;108:55-66.

doi: 10.1016/j.measurement.2017.05.022

18. Satija U, Ramkumar B, Manikandan MS. A new automated signal quality-aware ECG beat classification method for unsupervised ECG diagnosis environments. IEEE Sensors J. 2019;19(1):277-286.

    doi: 10.1109/JSEN.2018.2877055

19. Jing J, Huaifeng Z, Dechang P, Chenglong D. A novel multi-module neural network system for imbalanced heartbeats classification. *Expert Syst Applic X*. 2019;1:100003.

    doi: 10.1016/j.eswax.2019.100003

20. Tandale S, Barhatte AS, Ghongade R, Dale M. Arrhythmia Classification Using Neuro Fuzzy Approach. In: 2017 3rd International Conference on Advances in Computing, Communication and Automation. 2017. p. 1-4.

    doi: 10.1109/ICACCAF.2017.8344712

21. Rivera J, Rodriguez K, Yu XH. Cardiovascular Conditions Classification Using Adaptive Neuro-Fuzzy Inference System. In: 2019 IEEE International Conference on Fuzzy Systems. 2019. p. 1-6.

    doi: 10.1109/FUZZ-IEEE.2019.8858896

22. Sun Z, Wang C, Zhao Y, Yan C. Multi-label ECG signal classification based on ensemble classifier. IEEE Access. 2020;8:117986-117996.

    doi: 10.1109/ACCESS.2020.3004908

23. Kour H, Manhas J, Sharma V. Usage and implementation of neurofuzzy systems for classification and prediction in the diagnosis of different types of medical disorders. *Artif Intellig Rev*. 2020;53:1-56.

    doi: 10.1007/s10462-020-09804-x

24. Hanbay K. Deep neural network based approach for ECG classification using hybrid differential features and active learning. *IET Signal Process*. 2019;13(2):165-175.

    doi: 10.1049/iet-spr.2018.5103

25. Xu X, Jeong S, Li J. Interpretation of electrocardiogram (ECG) rhythm by combined CNN and BiLSTM. *IEEE Access*. 2020;8:125380-125388.

    doi: 10.1109/ACCESS.2020.3006707

26. Zhai X, Tin C. Automated ECG classification using dual heartbeat coupling based on convolutional neural network. *IEEE Access*. 2018;6:27465-27472.

    doi: 10.1109/ACCESS.2018.2833841.

27. Pan J, Tompkins JW. A real-time QRS detection algorithm. *IEEE Trans Biomed Eng*. 1985;32(3):230-236.

    doi: 10.1109/TBME.1985.325532

28. Kuhn M, Johnson K, editors. Classification models. In: *Applied Predictive Modeling*. New York: Springer; 2013.

    doi: 10.1007/978-1-4614-6849-3

29. Balbinot A, Favieiro G. A Neuro-Fuzzy system for characterization of arm movements. *Sensors*. 2013;13(2):2613-2630.

    doi: 10.3390/s130202613

30. Jang JS. Adaptive-network-based Fuzzy inference system. *IEEE Trans Syst Man Cybernet*. 1993;23(3):665-685.

    doi: 10.1109/21.256541Et omnim eveliquod que et, vollabo.

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Heartbeat classification using various machine learning models: A comparative study

Marc Nshimiyimana[1], Jovial Niyogisubizo[2]*, and Jean de Dieu Ninteretse[3]

[1]Department of Bridge, Tunnel and Underground Engineering, School of Civil Engineering, Southeast University, Nanjing, China

[2]Shenzhen Key Laboratory of Intelligent Bioinformatics and Center for High-Performance Computing, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

[3]Department of Construction and Real Estate, School of Civil Engineering, Southeast University, Nanjing, China

Abstract

Cardiac arrhythmias, known as irregular heartbeats, pose a notable health threat that necessitates prompt diagnosis, as untreated arrhythmias can lead to severe heart complications. Among the various methods for arrhythmia detection, electrocardiography is the most prevalent due to its non-invasive monitoring of heart activity. However, manual electrocardiogram (ECG) analysis is inefficient and prone to errors, prompting the exploration of machine learning (ML) models for ECG feature recognition. Integrating ML models with ECG analysis can revolutionize cardiac diagnostics by improving healthcare efficiency and outcomes by enhancing the accuracy and consistency of existing approaches as well as their processing speed for large datasets. Unfortunately, current ML methods encounter two key limitations: prolonged training times and the need for manual feature selection. To address these issues, we propose using ML models enhanced with innovative techniques such as the Fourier transform (FT) and Gaussian noise injection for improved cardiac health assessment. To validate this approach, we utilized statistical tools, including Pearson correlation and p-values, to uncover relationships within the data. In addition, we employed the FT technique to extract and analyze frequency-domain features. Our comparative study of different ML models relied on metrics such as accuracy, precision, recall, F1 score, and receiver operating characteristic area under the receiver operating characteristic curve, demonstrating XGBoost's impressive average recall of 0.956 with 99.96% overall accuracy. An average precision of 0.956 further underscored the accuracy of XGBoost's predictions, indicating its high level of reliability in distinguishing various cardiac conditions. These results highlight the considerable potential of ML techniques for precise ECG-based clinical diagnoses, helping healthcare professionals make more accurate and timely decisions in patient care.

Keywords: Electrocardiogram; Fourier transform; Gaussian noise; Heartbeat classification; Machine learning; Pearson correlation

***Corresponding author:**
Jovial Niyogisubizo
(jovial@siat.ac.cn)

Citation: Nshimiyimana M, Niyogisubizo J, Ninteretse JDD. Heartbeat classification using various machine learning models: A comparative study. *Artif Intell Health*. 2024;1(4):61-72. doi: 10.36922/aih.3543

Received: April 30, 2024

Accepted: September 3, 2024

Published Online: October 14, 2024

1. Introduction

Electrocardiography generates an electrocardiogram (ECG or EKG), a record of the heart's electrical activity, which can be monitored using both traditional medical devices

and smart healthcare systems. An EKG is a graphical representation of the heart's electrical activity over time, obtained by affixing electrodes on the skin surface. The human body generates various biomedical signals, and electrocardiography sensors record the heart's electrical activity by generating three-lead EKGs for heart monitoring and surface electromyography for muscle contractions.[1] ECGs are crucial for diagnosing cardiovascular conditions and aid in the early detection of heart attacks and treatment to prevent them.[2] However, identifying and classifying arrhythmias are challenging due to the requirement of extensive data analysis, potential human errors, and signal variability. Furthermore, developing an automated detection system is difficult due to data complexity and noise; however, ECGs are indispensable for diagnosing heart diseases. Anomalous deviations from a typical ECG pattern are observed in various cardiac disorders, including irregular heart rhythms such as atrial fibrillation and ventricular tachycardia;[3] diminished blood flow through the coronary arteries, which is observed in cases such as myocardial ischemia and myocardial infarction; and disruptions in the electrolyte balance such as in hypokalemia and hyperkalemia.[4]

Since the 20th century, ECG analysis has been vital for diagnosing cardiovascular diseases by monitoring the heart's electrical activity. The conventional 12-lead ECG, recorded by affixing electrodes to the chest and limbs, is crucial for detecting arrhythmias. Early diagnosis is essential for effective treatment, sometimes requiring over 24 h of continuous monitoring.[5] Advances in the digital industry have improved devices, data acquisition techniques, and computer-assisted diagnostics. Although ECGs are widely used by cardiologists to monitor heart health, a manual analysis of the associated signals is time-consuming and error-prone. Therefore, an accurate diagnosis of cardiovascular conditions is crucial as these conditions contribute to approximately one-third of global mortality. The prevalence of irregular heart rhythms highlights the need for precise and cost-effective methods to diagnose arrhythmic heartbeats.[6]

The classification of ECG signals is challenging due to individual variability and the lack of standardization in feature extraction, often leading to low diagnostic accuracy.[7-9] To address the limitations of manual ECG signal analysis, numerous studies have applied machine learning (ML) techniques for accurate anomaly detection.[10-12] Traditional methods for this analysis involve signal preprocessing, handcrafted feature extraction, and the use of ML and deep learning algorithms. However, deep learning requires extensive datasets due to the large number of the associated parameters. The traditional 12-lead ECG analysis in clinical settings depends on the expertise and recording quality of clinicians, yielding high diagnostic accuracy but remaining prone to human errors.[5] ML models, trained on large annotated datasets and enhanced with preprocessing techniques such as Fourier transform (FT) and Gaussian noise injection, can achieve similar or even superior accuracy at 90 – 99% in detecting cardiac abnormalities.[13]

This study integrates the use of the FT and Gaussian noise injection techniques to enhance the capability of ML models trained and tested on the datasets provided by the University of Chinese Academy of Sciences, which will be presented in subsequent sections. The FT technique is essential for preprocessing heartbeat signals in ML models, converting time-domain signals into frequency-domain ones to filter noise and improve signal quality, thereby enhancing model performance to a level equal to or greater than that obtained with the 12-lead ECG analysis. Unlike state-of-the-art algorithms such as convolutional neural networks (CNNs), which filter noise implicitly, the FT technique explicitly reduces computational demands and improves training efficiency. Traditional ML models benefit greatly from noise reduction, leading to improved classification accuracy. In addition, Gaussian noise injection adds controlled noise during training, enhancing model robustness and generalization capability, which is particularly useful for heartbeat classification, thereby improving performance on unseen data and reducing the risk of overfitting. In practical applications, these methods enhance accuracy and reliability, making them suitable for applications in clinical settings and resource-constrained environments such as wearable devices and mobile health applications.

This study aims to classify heartbeats using data from heartbeat sequences recorded by patient heartbeat signal sensors. The key contributions of this study are summarized as follows:

- Introduce the innovative application of the FT technique to preprocess ECG signals by converting time-domain data into frequency-domain representations, allowing for the effective noise reduction and enhancement of signal quality.
- Employ Gaussian noise injection as a novel approach to simulate real-world noise conditions by training datasets to be more robust against various types of noise and artifacts commonly observed in ECG signals.
- Conduct a comparative analysis of various ML models by assessing their performance on the preprocessed ECG data. Emphasize the influence of the FT and Gaussian noise integration techniques on key metrics such as accuracy, precision, recall, and F1 score for each algorithm.

- Evaluate the scaling of the ML models with varying dataset sizes, emphasizing the manner in which preprocessing methods and algorithm adaptations help maintain efficiency and effectiveness across different data volumes.

The organization of this paper, illustrated in Figure 1, is as follows: Section 1 introduces the topic, explaining ECG classification using various models. Section 2 outlines the data and methods used, covering all aspects from dataset collection to signal classification. Section 3 focuses on interpreting the results, and Section 4 discusses the proposed approach.

2. Data and methods

2.1. ECG datasets

Previous studies have achieved promising results in classifying heartbeat segments based on arrhythmia classes using the MIT-BIH Arrhythmia Database.[14-16] However, class imbalance has remained a notable issue in electronic health (eHealth), where abnormal samples are much fewer than normal ones. This imbalance can bias the model toward the dominant class, leading to the poor or average classification of the minority class, which negatively impacts classification accuracy and other performance metrics.[17] Instead of the MIT-BIH Arrhythmia Database, which is widely known for ECG classifications, this study uses a dataset provided by the University of Chinese Academy of Sciences, which is available on request. This dataset includes four categories: Normal (N), supraventricular (S), ventricular ectopic (V), and fusion (F), as indicated in Table 1.

The adopted dataset consists of ECG recordings from 205 heartbeat signals. To protect the personal information of the patients, appropriate measures are undertaken to ensure fairness in the evaluation process, which utilizes a training set comprising 80,000 samples for model construction and validation and a test set comprising 20,000 samples. The primary task is to predict the ECG heartbeat signal category of the dataset signals, which are provided by a platform that records ECG data by capturing only one column of the heartbeat signal sequence. Each sample within this sequence is sampled at the same frequency and is of equal length to ensure consistency across the dataset. Annotations in this dataset are used to create four different beat categories, and this categorization follows the standards set by the Association for the Advancement of Medical Instrumentation EC57.[18] Table 1 summarizes the mappings between beat annotations in each category.

2.2. Data preprocessing

This section details the N, S, V, and F categories used in this study (Table 1). The training and test set distributions are illustrated in Figures 2 and 3, respectively, which depict the class imbalance phenomenon. On training with different ML models, class weights are assigned to address this class imbalance. Figure 4 presents a normal and an abnormal heartbeat, with its x-axis denoting the time frame ranging from 0.0 ms to 1.6 ms and its y-axis representing the normalized amplitudes of heartbeat signals. The methodology section further describes the associated methods.

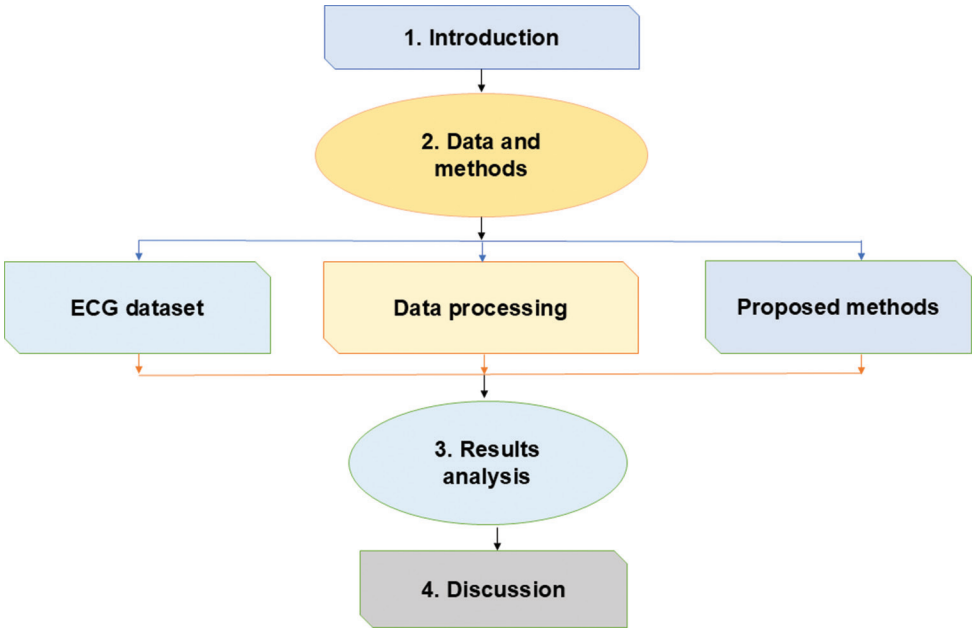


Figure 1. Organization of the paper

Table 1. Summary of mappings between beat annotations and Association for the Advancement of Medical EC57 categories[18]

Label	Category	Annotation
0	N	• Normal • Left/right bundle branch block • Atrial escape • Nodal escape
1	S	• Atrial premature • Aberrant atrial premature • Nodal premature • Supraventricular premature
2	V	• Premature ventricular contraction • Ventricular escape
3	F	• Fusion of ventricular and normal

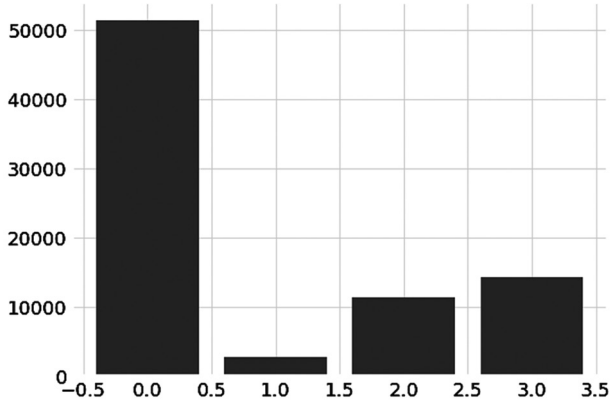


Figure 2. Distribution of the training set

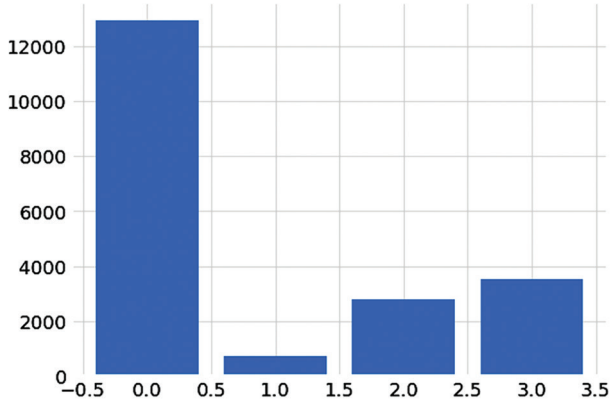


Figure 3. Distribution of the test set

2.3. Methods used

Enhancing the classification prediction accuracy can aid in the early diagnosis of cardiovascular diseases. However, a review of current state-of-the-art research indicates that metrics and prediction rates often fall short. Literature highlights several challenges, including (I) handling missing and imbalanced data, (II) selecting a robust classification algorithm, (III) managing ECG signal complexity, (IV) addressing computational demands, and (V) recognizing methodological limitations. To achieve the objectives of this study, seven classification algorithms were used to classify heartbeat categories, and their performance was evaluated. Herein, ML models-nearest neighbors (KNN), naive Bayes (NB) classifier, random forest (RF) classifier, logistic regression (LR), eXtreme gradient boosting (XGBoost) classifier, support vector machines (SVMs), and decision trees (DTs) were employed, along with the incorporation of the FT and Gaussian noise injection techniques. The overall design of the implementation method is presented in Figure 5. In this study, we utilized the advantages of Pearson correlation and the associated p-values, which serve as statistical tools, to unveil meaningful relationships and dependencies within our data. In addition, we introduced controlled noise to enhance the robustness of our models, allowing them to better adapt to real-world variations. Furthermore, the FT technique was leveraged to extract essential frequency-domain features from our data. This combination enabled our models to make more accurate predictions and better understand complex data patterns, offering valuable insights for various applications, comparable to state-of-art algorithms.

The ML models enhanced through the proposed approach can exhibit improvements in terms of key metrics such as accuracy and F1 score. KNN and NB are favored for their simplicity and real-time efficiency, while RF and XGBoost excel in handling complex interactions and large datasets, offering robustness and feature importance. LR is valued for its interpretability in binary classification, SVMs are known for managing high-dimensional data and noise, and DTs are favored for their clear and interpretable results. These algorithms were selected for their effectiveness in handling complex ECG data, computational efficiency, adaptability, and noise robustness, ensuring reliable performance and scalability. Notably, the ability of ML models to handle motion artifacts in ECG signals varies. RF and XGBoost are particularly robust under noisy environments due to their ensemble nature,[19] while SVMs effectively maintain accuracy using kernel methods.[20] KNN and NB face challenges with respect to noise sensitivity; however, preprocessing techniques such as FT can help in this regard. LR and DT require feature engineering and noise reduction for exhibiting better performance,[21,22] with ensemble methods further boosting their robustness.

As detailed in a previous study,[23] Pearson's correlation test is a statistical method used to assess the relationship between two continuous variables. It yields a coefficient

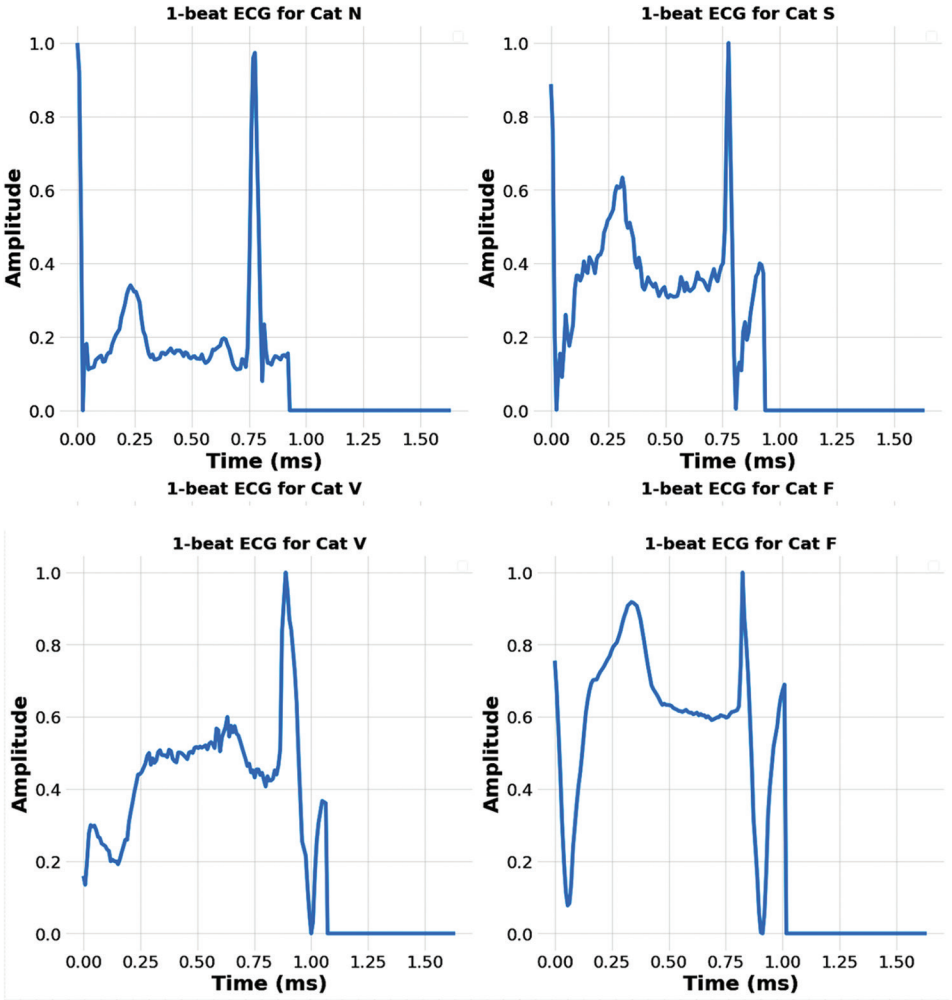


Figure 4. Categories of heartbeats in the given datasets

denoted as "r," whose value falls within the range of −1 to 1. A coefficient value of 0 implies no correlation, 1 represents a perfect positive linear relationship, and −1 indicates a perfect negative linear relationship. This test quantifies the strength of the linear relationship, as explained in previous studies.[24,25] The accompanying p-value gauges the probability of obtaining the observed result assuming no correlation (the null hypothesis). A p-value below 0.05 signifies statistical significance. Furthermore, the 95% confidence interval specifies a range within which the true correlation coefficient is likely to be found with 95% confidence.[26] The Pearson correlation is governed by (I).

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}} \quad \text{(I)}$$

where x_i denotes the samples of the x variable, $\overline{x}$ represents the mean of the x variable, y_i denotes the samples of the y variable, and $\overline{y}$ represents the mean of the y variable.[27]

Regularization in ML often involves adding noise during training, similar to techniques such as dropout. This practice enhances model robustness and minimizes overfitting risk by making it difficult for the model to perfectly fit data. Noise can be introduced at various stages, such as in inputs, weights, gradients, and activation functions, offering flexibility in its use in regularization. Introducing noise during training improves resilience and reduces generalization errors. Typically, this noise is applied to the input data; however, it can also be incorporated into weights, gradients, and activation functions as alternative strategies.[28] The input vectors $\{x_1, x_2,\ldots,x_n\}$[18] yield $\{y_1, y_2,\ldots,y_n\}$ associated with it, and the noise calculation is detailed in (II).

$$y = f(x) + \varepsilon \quad \text{(II)}$$

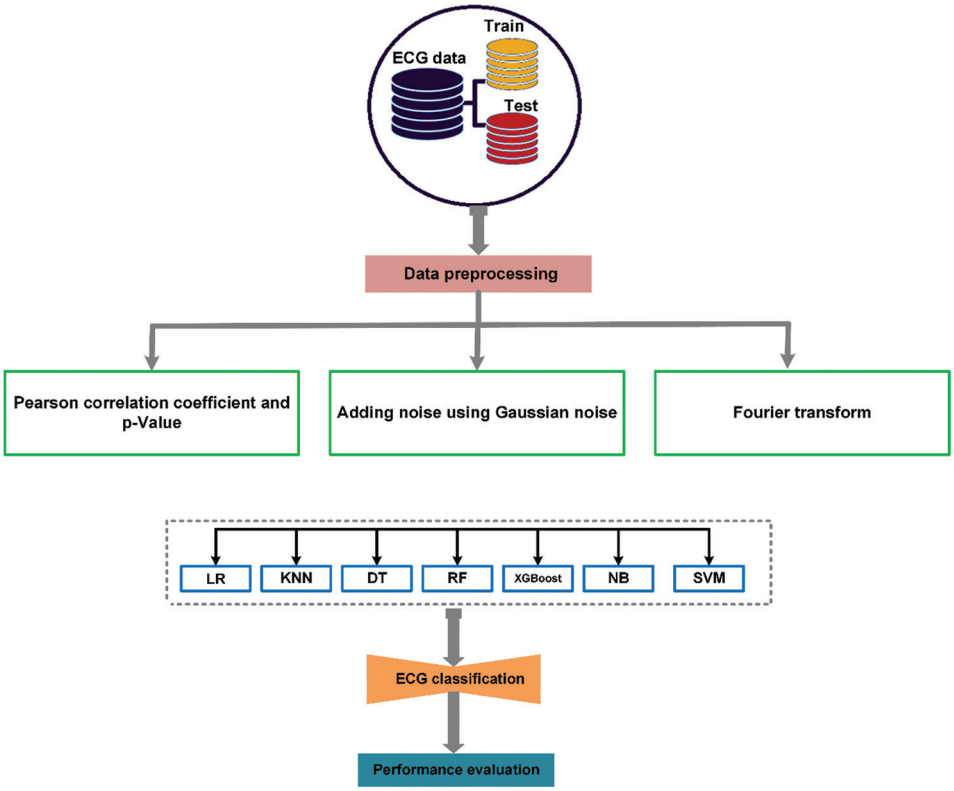


Figure 5. Proposed methodology

where $f(x)$ represents the relationship between the input features x and outcome y and ϵ accounts for noise in the data. A Fourier series is used to represent periodic signals by decomposing them into their frequency components. The fast Fourier-transform (FFT) algorithm converts a digital signal from the time domain to the frequency domain, which is useful for analyzing signals with intriguing frequency characteristics. In fields such as image processing and ML models like CNNs, the FFT algorithm simplifies convolution operations by converting images and kernels into the frequency domain, enabling straightforward element-wise multiplications. However, employing FFT for this purpose introduces an additional computational overhead, an aspect detailed in a previous study.[29] In the FFT algorithm, the real and imaginary components for image processing[30] are governed by (III) and (IV).

$$F(x,y)=\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}f(m,n)e^{-(i\times 2\times\pi\left(x\frac{m}{M}+y\frac{n}{N}\right))} \quad \text{(III)}$$

$$F(x,y)=\frac{1}{M.N}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}F(m,n)e^{(i\times 2\times\pi\left(x\frac{m}{M}+y\frac{n}{N}\right))} \quad \text{(IV)}$$

where $F(m,n)$ denotes the pixel located at the coordinates (m,n), and $F(x,y)$ is the function describing the image in the frequency domain at coordinates x and y. The image dimensions are represented by $M \times N$, and i refers to the square root of −1.

2.3.1. KNN

The KNN algorithm, introduced by statisticians Richard Cover and Peter Hart,[31] is a fundamental technique for pattern recognition, and ML is used primarily for classification tasks. The algorithm predicts the class of an observation by identifying "K" nearest data points in the feature space based on a distance metric, following which it assigns the class through a majority vote among these neighbors. The key step in the KNN algorithm involves accurately calculating distances between the target data point and other data points in the dataset to assess their similarity.[32] The KNN algorithm makes predictions by identifying the closest data points, and its output is dependent on the majority outcome of the neighbors. The model calculates the distance between the target data points and nearest K neighbors using metrics such as the Euclidean distance, which is utilized to quantify similarity. The number of neighbors (K) is selected through cross-validation to optimize prediction accuracy. Using the Euclidean distance in the KNN algorithm ensures a straightforward and intuitive assessment of proximity by

leveraging the geometric distance in a multidimensional space. The Euclidean distance between two observations is expressed in (V).

$$d_{ij} = \sqrt[2]{\left(x_i - x_j\right)^2 + \left(y_i + y_j\right)^2} \quad \text{(V)}$$

where d_{ij} represents the Euclidean distance between points i and j. Point i is represented as $x_i=[x_{i1},x_{i2},x_{i3},\ldots,x_{iM}],y_i$. Meanwhile, point j is represented as $x_j=[x_{j1},x_{j2},x_{j3},\ldots,x_{jM}],y_j$.

2.3.2. LR

LR is a supervised ML algorithm widely used in multivariate statistical methods to predict binary outcomes based on a set of observed independent variables. It specifically handles categorical response variables that represent binary events, rather than continuous parameters. The detailed process of constructing an LR model can be found elsewhere.[33] The final output of LR is a probability, ranging from 0 to 1, that represents the occurrence likelihood of an event. This probability is mathematically expressed in (VI):

$$P_x = \frac{1}{1+e^{-C_0+C_1x}} \quad \text{(VI)}$$

where P_x represents the occurrence probability of the event, and e denotes the base of the natural logarithm. The terms $C_0 + C_1$ are model parameters. The coefficients of the LR model are estimated using the maximum likelihood method, which involves choosing coefficients that maximize the probability of the model correctly predicting the observed outcomes.

2.3.3. RF

RF is an ensemble learning technique employed for classification and regression tasks. It builds multiple classifiers and combines classifier outputs during training. The RF model comprises numerous DTs, each contributing to the final prediction by providing a vote for the most common class. This method enhances prediction performance by employing uncorrelated trees created through bootstrap aggregation, wherein training subsets are generated through replacement, allowing data points to be sampled multiple times to create diverse subsets. Cross-validation within the RF model reduces estimation and out-of-bag errors, yielding highly reliable trees and improving prediction accuracy.[34] Moreover, all characteristics are bounded using a stochastic methodology. An inherent benefit of the RF model is its ability to generate an extensive array of trees, which enhances diversity and mitigates bias-related concerns. Following the tree generation, a new observation is classified by averaging its vote across all the DTs.

2.3.4. DTs

DT classifiers are among the most supervised learning algorithms and are widely utilized for classification tasks. These trees are constructed from the provided data using straightforward equations as they employ attribute selection measures such as the gain ratio measure to rank attributes and identify the most important ones. This process enables researchers to determine the most effective attributes for prediction purposes. DT is a prominent data mining technique for creating classification models, and it is highly practical due to its speed,[35] lack of the requirement of domain knowledge or parameter tuning, ability to handle multidimensional data, and capacity to produce easily interpretable classification rules. Generally, DT classifiers offer good accuracy and common examples include ID3, C4.5/C5.0/J48, CART, and Random Tree.[36]

2.3.5. SVMs

SVMs are supervised ML models designed for pattern classification and regression tasks, based on the structural risk minimization theory.[37] These non-parametric approaches employ kernels to tackle non-linear and high-dimensional problems by mapping data into a higher-dimensional space, where linear separation is feasible. SVMs optimize the margin between classes, thus minimizing the overfitting risk and enhancing the model's generalization ability for unseen data. By constructing an optimal hyperplane that maximizes the margin between classes, SVMs ensure effective classification and regression across various fields.

2.3.6. XGBoost

XGBoost, introduced by Tianqi Chen, is a powerful ML algorithm grounded in the principle of gradient boosting. Officially released on March 27, 2014, XGBoost is designed to enhance the performance of DTs. Since its introduction, XGBoost has become a popular choice in data science competitions and applications due to its efficiency, accuracy, and scalability.[38] This model is versatile and can be used to tackle regression and classification problems. In regression tasks, it predicts continuous outcomes based on the input features, while in classification tasks, it categorizes data into distinct classes. Mathematically, the gain from XGBoost is used in a regularized boosting technique, which is defined by equation (VII)

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\beta} + \frac{G_R^2}{H_R+\beta} - \frac{\left(G_R+G_L\right)^2}{H_R+H_L+\beta}\right] - \alpha \quad \text{(VII)}$$

where the first and second terms represent the score of the left child and right child, respectively, and the third term

Table 2. Classification performance of different ML models

Method	Training Accuracy (%)	Test Accuracy (%)	Precision	Recall	F1-Score
LR	79.65	79.30	0.793	0.793	0.793
KNN	98.74	96.57	0.965	0.965	0.965
DT	100	91.32	0.913	0.913	0.913
RF	99.99	95.12	0.951	0.951	0.951
XGBoost	99.96	95.68	0.956`	0.956	0.956
NB	34.77	35.21	0.352	0.352	0.352
SVM	95.55	94.48	0.944	0.944	0.944

represents the score when no split occurs. Furthermore, β and α denote the ridge and lasso regularization coefficients, respectively.[39]

2.3.7. NB

The NB method is an ML approach introduced based on Bayes' theorem. In this method, Bayes' theorem serves as the principal foundation for Bayesian inference, which permits the computation of parameter unpredictability using event probabilities. As detailed in the literature,[40] the probability reflects the evolutionary degree of belief regarding the parameters before data observation and after data inspection during analysis. Detailed procedures for developing the Bayesian method can be found elsewhere.[41,42]

2.4. Performance evaluation

To assess the findings of this study, various classification and performance metrics were utilized, including accuracy (ACC), precision (PR), recall (Rec), the area under the receiver operating characteristic curve (AUC), and the F1 score.[43] The following equations provide a concise overview of each metric adopted during the evaluation process.

$$ACC = \frac{TP + T\,N}{TP + TN + FN + FP} \quad \text{(VIII)}$$

$$PR = \frac{TP}{TP + FP} \quad \text{(IX)}$$

$$Rec = \frac{TP}{TP + FN} \quad \text{(X)}$$

$$F1\ Score = \frac{2*Rec*PR}{Rec + PR} \quad \text{(XI)}$$

In these equations, TP, TN, FP, and FN refer to true positive, true negative, false positive, and false negative, respectively.

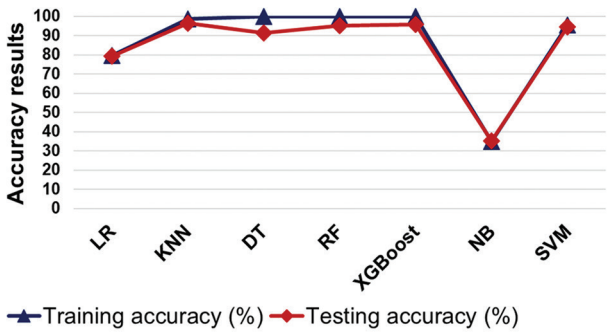


Figure 6. Accuracy performance of all methods

3. Results

3.1. Result analysis

To illustrate the effectiveness of our heartbeat classification approach, we analyze the experimental outcomes obtained from various benchmark ML models. This comparative analysis utilizes training and test results derived from the LR, KNN, DT, RF, XGBoost, NB, and SVM models. As indicated in Table 2 and Figure 6, the XGBoost model demonstrates the best performance, achieving a training accuracy of 99.96% and a test accuracy of 95.68%, thereby outperforming all other models. Conversely, the DT model experiences issues related to overfitting.

Although accuracy is a commonly adopted metric for evaluating individual model performance, relying solely on this metric can be misleading. A model may achieve high accuracy in predicting major classes while struggling with minor ones. To address this limitation, we adopt additional performance indicators, such as precision, recall, and the F1-score. Table 2 details the performance results of all models for both training and test sets. For instance, the XGBoost ensemble achieves an overall average recall of 0.956 on the test set, indicating that nearly 95% of high heartbeat cases are correctly predicted. Similarly, the average precision of 0.956 for the test set implies that our predictions for all heartbeat categories are approximately 95% accurate.

4. Discussion

XGBoost, the best-performing classifier, was applied to the test set to generate a confusion matrix, assessing the effectiveness of the classification model. An overall accuracy of 99.96% demonstrated the model's ability to classify ECG signals more accurately than other MLmodels. Table 3 highlights the superior performance of our method in efficiently and automatically identifying arrhythmias through the classification of heartbeats from the ECG signals. The accuracy of the proposed method with balanced data surpasses that of other state-of-art methods, especially for the N, S, and V categories (Table 1), achieving values of 99.96% and 95.68% for training and test, respectively. These values represent greater performance than obtained with other state-of-the-art methods that use the previously proposed reservoir computing (RC) and next-generation RC approaches.[44] While a previous study[45] achieved an accuracy of 99.6% by introducing a CNN-based model with feature component analysis for multimodal ECG tasks on the MIT-BIH dataset, our model exhibits sound performance by incorporating the FT and Gaussian noise injection techniques. Another study[17] explored heartbeat classification and arrhythmia detection using multimodel deep learning techniques such as a one-dimensional CNN and long short-term memory network, achieving overall accuracies of 99.59% and 99.35%, respectively, which are lower than those achieved in our study, supporting the reliability of our model in heartbeat classification. The proposed approach also requires lesser training time and is more cost-effective than other state-of-the-art algorithms, which are often deemed expensive and time-consuming.

While the results for all models indicate satisfactory performance, the AUC stands out as the key metric for evaluating the overall effectiveness of the top-performing model across all categories. Figure 7 illustrates that the AUC results and the corresponding receiver operating characteristic (ROC) curves for XGBoost across all categories using raw data yield acceptable outcomes. A higher AUC score suggests superior classification performance as points representing model classification better than random guesses are positioned above the diagonal line (Figure 7).

In our ECG heartbeat classification analysis, several ML models were employed to categorize heartbeats from ECG signals based on their distinctive characteristics. These models were trained on labeled datasets and utilized various algorithms to recognize intricate patterns within the ECG data. The performance of these models was collectively evaluated using standard metrics such as accuracy, precision, recall, F1 score, and ROC-AUC (Figure 7). By systematically applying these metrics, we objectively evaluated and identified the most effective model for ECG heartbeat classification, catering to specific clinical requirements and optimizing diagnostic accuracy for various cardiac conditions. Our analysis reveals that XGBoost, enhanced by the FT and Gaussian noise injection techniques, demonstrates reasonable accuracy for clinical analysis.

Table 3. Comparison between the proposed ML model and state-of-art algorithms in terms of their accuracy performance

Approach	Overall accuracy (%)
Proposed approach	99.96
ML models with Optimized RF[46]	97.7
Deep LSTM[47]	95.80
Deep 1D-CNN[48]	97.00
CNN+LSTM[17]	99.35
RC+NG-RC[44]	96.05 and 98.28
Bi-LSTMs[49]	98.70

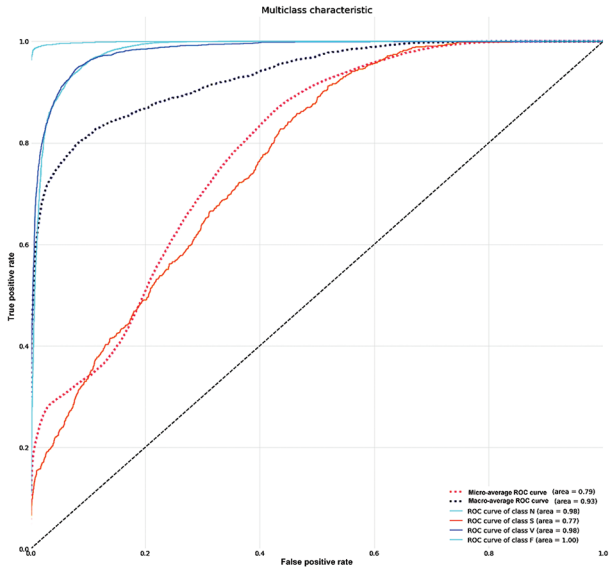


Figure 7. Receiver operating characteristic (ROC) curves and area under the ROC curve results of XGBoost for all heartbeat categories

5. Conclusions

This study examined the performance of ML models designed for heartbeat classification, aiming to help medical specialists in identifying appropriate treatments. It focused on the analysis of four distinct heartbeat signal types, utilizing a substantial ECG dataset comprising 80,000 training and 20,000 test samples. We employed random sampling techniques to address class imbalance problems, ensuring a uniform representation of samples across classes. During our experiments, we implemented seven ML models, incorporating methods such as FT and Gaussian

noise injection, alongside Pearson correlation and p-values, to enhance the classification performance to meet the level of 12-lead ECG analysis, which is considered the most accurate analysis in clinical settings. Notably, the XGBoost method demonstrated exceptional performance, achieving high accuracy in heartbeat classification. Conversely, NB displayed suboptimal classification capabilities among the investigated models. The conclusions of this study are summarized as follows:

1. Introducing FT-based feature extraction and Gaussian noise regularization substantially improved the performance and robustness of ECG heartbeat classification models.
2. The findings provided valuable insights into the comparative performance of various ML algorithms, assisting researchers and clinicians in selecting the most appropriate model for specific healthcare applications.
3. The FT technique enabled effective capture of frequency-domain information that is critical for accurate heartbeat classification, thereby enhancing the models' diagnostic capabilities.
4. Controlled Gaussian noise injection during training proved beneficial for model generalization in real-world scenarios.
5. Our findings demonstrated the potential of ML in advancing cardiac healthcare monitoring and classification, offering practical tools for more accurate and reliable ECG-based diagnoses.

The study's primary limitations are the requirements of advanced deep learning models and larger and diverse datasets to improve ECG heartbeat classification accuracy and robustness. To address these limitations, future research should prioritize the exploration of deep learning architectures and the acquisition of more comprehensive and varied datasets, which will ensure the reliability and real-world applicability of the model in clinical settings.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

Conceptualization: Jovial Niyogisubizo
Formal analysis: Jovial Niyogisubizo, Marc Nshimiyimana
Investigation: Marc Nshimiyimana, Jovial Niyogisubizo
Methodology: Marc Nshimiyimana, Jovial Niyogisubizo
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 code and data related to this study, along with more technical details, can be found here: https://github.com/jovialniyo93/heartbeat-classification-with-machine-models.

Further disclosure

We confirm that this work is the result of a collaborative effort among three researchers, offering diverse perspectives, including AI in health and the application of ML to real-world problems. The contributions have been significant, enhancing expertise in the field and building on our previous publication on predicting red wine quality using novel ML methods. The first author (M.N.) is a Master's student at the Southeast University with expertise in applying AI, particularly ML algorithms to civil and geotechnical engineering fields. The third author, who is a Ph.D. student at the Southeast University specializing in infrastructure resilience, project management, and risk management (J.D.N.), contributed his perspectives to the present work.

References

1. Periyaswamy T, Balasubramanian M. Ambulatory cardiac bio-signals: From mirage to clinical reality through a decade of progress. *Int J Med Inform.* 2019;130:103928.

   doi: 10.1016/j.ijmedinf.2019.07.007

2. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol.* 2021;18(7):465-478.

   doi: 10.1038/s41569-020-00503-2

3. Goudis CA, Konstantinidis AK, Ntalas IV, Korantzopoulos P. Electrocardiographic abnormalities and cardiac arrhythmias in chronic obstructive pulmonary disease. *Int J Cardiol.* 2015;199:264-273.

   doi: 10.1016/j.ijcard.2015.06.096

4. Teymouri N, Mesbah S, Navabian SMH, *et al*. ECG frequency changes in potassium disorders: A narrative review. *Am J Cardiovasc Dis.* 2022;12(3):112.

5. Faruk N, Abdulkarim A, Emmanuel I, *et al*. A comprehensive survey on low-cost ECG acquisition systems: Advances on design specifications, challenges and future direction.

Biocybernetics Biomed Eng. 2021;41(2):474-502.

doi: 10.1016/j.bbe.2021.02.007

6. Acharya UR, Oh SL, Hagiwara Y, *et al*. A deep convolutional neural network model to classify heartbeats. *Comput Biol Med.* 2017;89:389-396.

doi: 10.1016/j.compbiomed.2017.08.022

7. Luz EJS, Schwartz WR, Cámara-Chávez G, Menotti D. ECG-based heartbeat classification for arrhythmia detection: A survey. *Comput Methods Programs Biomed.* 2016;127:144-164.

doi: 10.1016/j.cmpb.2015.12.008

8. Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:170701836*; 2017.

doi: 10.48550/arXiv.1707.01836

9. Krasteva V, Jekova I, Leber R, Schmid R, Abächerli R. Superiority of classification tree versus cluster, fuzzy and discriminant models in a heartbeat classification system. *PLoS One*. 2015;10(10):e0140123.

doi: 10.1371/journal.pone.0140123

10. Syama S, Sweta GS, Kavyasree P, Reddy KJM. Classification of ECG Signal using Machine Learning Techniques. In: *2019 2nd International Conference on Power and Embedded Drive Control (ICPEDC)*: IEEE; 2019. p. 122-128.

doi: 10.1109/ICPEDC47771.2019.9036613

11. Jambukia SH, Dabhi VK, Prajapati HB. Classification of ECG Signals Using Machine Learning Techniques: A Survey. In: *2015 International Conference on Advances in Computer Engineering and Applications*: IEEE; 2015. p. 714-721.

doi: 10.1109/ICACEA.2015.7164783

12. Xue J, Yu L. Applications of machine learning in ambulatory ECG. *Hearts.* 2021;2(4):472-494.

doi: 10.3390/hearts2040037

13. Zulfiqar R, Majeed F, Irfan R, Rauf HT, Benkhelifa E, Belkacem AN. Abnormal respiratory sounds classification using deep CNN through artificial noise addition. *Front Med* (*Lausanne*). 2021;8:714811.

doi: 10.3389/fmed.2021.714811

14. Liu S, Shao J, Kong T, Malekian R. ECG arrhythmia classification using high order spectrum and 2D graph Fourier transform. *Appl Sci.* 2020;10(14):4741.

doi: 10.3390/app10144741

15. Bhattacharyya S, Majumder S, Debnath P, Chanda M. Arrhythmic heartbeat classification using ensemble of random forest and support vector machine algorithm. *IEEE Trans Artif Intell.* 2021;2(3):260-268.

doi: 10.1109/TAI.2021.3083689

16. Marinho LB, Nascimento NMM, Souza JWM, Gurgel MV, Rebouças Filho PP, de Albuquerque VHC. A novel electrocardiogram feature extraction approach for cardiac arrhythmia classification. *Future Generation Comput Syst.* 2019;97:564-577.

doi: 10.1016/j.future.2019.03.025

17. Irfan S, Anjum N, Althobaiti T, Alotaibi AA, Siddiqui AB, Ramzan N. Heartbeat classification and arrhythmia detection using a multi-model deep-learning technique. *Sensors* (*Basel*). 2022;22(15):5606.

doi: 10.3390/s22155606

18. Ahmad Z, Tabassum A, Guan L, Khan NM. ECG heartbeat classification using multimodal fusion. *IEEE Access.* 2021;9:100615-100626.

doi: 10.1109/ACCESS.2021.3097614

19. Wu X, Zheng Y, Chu CH, He Z. Extracting deep features from short ECG signals for early atrial fibrillation detection. *Artif Intell Med.* 2020;109:101896.

doi: 10.1016/j.artmed.2020.101896

20. Müller KR, Mika S, Tsuda K, Schölkopf K. An introduction to kernel-based learning algorithms. In: *Handbook of Neural Network Signal Processing*. United States: CRC Press; 2018. p. 4-1-4-40.

doi: 10.1201/9781315220413-4

21. Rahmani AM, Yousefpoor E, Yousefpoor MS, *et al*. Machine learning (ML) in medicine: Review, applications, and challenges. *Mathematics.* 2021;9(22):2970.

doi: 10.3390/math9222970

22. Zdravevski E, Lameski P, Trajkovik V, *et al*. Improving activity recognition accuracy in ambient-assisted living systems by automated feature engineering. *IEEE Access.* 2017;5:5262-5280.

doi: 10.1109/ACCESS.2017.2684913

23. Mushtaq S, Faizi N, Amin SS, Adil M, Mohtashim M. Impact on quality of life in patients with dermatophytosis. *Australas J Dermatol.* 2020;61(2):e184-e188.

doi: 10.1111/ajd.13191

24. Sari BG, Lúcio ADC, Santana CS, Krysczun DK, Tischler AL, Drebes L. Sample size for estimation of the Pearson correlation coefficient in cherry tomato tests. *Ciên Rural.* 2017;47:e20170116.

doi: 10.1590/0103-8478cr20170116

25. Villavicencio CN, Macrohon JJ, Inbaraj XA, Jeng JH, Hsieh JG. Development of a machine learning based web application for early diagnosis of COVID-19 based on symptoms. *Diagnostics* (*Basel*). 2022;12(4):821.

doi: 10.3390/diagnostics12040821

26. George A, Stead TS, Ganti L. What's the risk: Differentiating risk ratios, odds ratios, and hazard ratios? *Cureus.*

2020;12(8):e10047.

doi: 10.7759/cureus.10047

27. Ramlee N, Ismail N. Analysis COVID-19 death cases in pulau pinang using multiple linear regression. *Proc Sci Math*. 2022;8:102-108.

28. Sabiri B, Asri B El, Rhanoui M. Mechanism of overfitting avoidance techniques for training deep neural networks[J/OL]. *In Proceedings of the 24th International Conference on Enterprise Information Systems*. 2022;1:418-427.

29. Nair V, Chatterjee M, Tavakoli N, Namin AS, Snoeyink C. Fast Fourier transformation for optimizing convolutional neural networks in object recognition. 2020.

doi: 10.48550/arXiv.2010.04257

30. Chughtai BR, Jalal A. Traffic Surveillance System: Robust Multiclass Vehicle Detection and Classification. In: *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*: IEEE; 2024. p. 1-8.

doi: 10.1109/ICACS60934.2024.10473304

31. Pathirana VK. *Nearest Neighbor Foreign Exchange Rate Forecasting with Mahalanobis Distance. Graduate Theses and Dissertations*; 2015.

32. Mucherino A, Papajorgji PJ, Pardalos PM. K-nearest neighbor classification. In: *Data Mining in Agriculture*. Berlin: Springer; 2009. p. 83-106.

doi: 10.1007/978-0-387-88615-2_4

33. James G, Witten D, Hastie T, Tibshirani R. Linear regression. In: *An Introduction to Statistical Learning*. Berlin: Springer; 2013. p. 59-126.

doi: 10.1007/978-1-4614-7138-7_3

34. Fox EW, Hill RA, Leibowitz SG, Olsen AR, Thornbrugh DJ, Weber MH. Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environ Monit Assess.* 2017;189:316.

doi: 10.1007/s10661-017-6025-0

35. Shafiq M, Tian Z, Bashir AK, Jolfaei A, Yu X. Data mining and machine learning methods for sustainable smart cities traffic classification: A survey. *Sustain Cities Soc.* 2020;60:102177.

doi: 10.1016/j.scs.2020.102177

36. Jiawei Han M, Pei J. *Data Mining: Concepts and Techniques*. Amsterdam: Elsevier; 2011.

37. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing.* 2020;408:189-215.

doi: 10.1016/j.neucom.2019.10.118

38. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016. p. 785-794.

doi: 10.1145/2939672.2939785

39. Niyogisubizo J, Liao L, Nziyumva E, Murwanashyaka E, Nshimyumukiza PC. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Comput Educ Artif Intell.* 2022;3:100066.

doi: 10.1016/j.caeai.2022.100066

40. Jin Y, Biscontin G, Gardoni P. A Bayesian definition of "most probable" parameters. *Geotechnical Res.* 2018;5(3):130-142.

doi: 10.1680/jgere.18.00027

41. Houlsby N, Houlsby G. Statistical fitting of undrained strength data. *Géotechnique.* 2013;63(14):1253-1263.

doi: 10.1680/geot.13.P.007

42. Niyogisubizo J, Liao L, Zou F, *et al*. Predicting traffic crash severity using hybrid of balanced bagging classification and light gradient boosting machine. *Intell Data Anal.* 2023;27(1):79-101.

doi: 10.3233/IDA-216398

43. Powers DM. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:201016061*; 2020.

doi: 10.48550/arXiv.2010.16061

44. Arbateni K, Benzaoui A. Enhancing heartbeat classification through cascading next generation and conventional reservoir computing. *Appl Sci.* 2024;14(7):3030.

doi: 10.3390/app14073030

45. Zhou F, Fang D. Multimodal ECG heartbeat classification method based on a convolutional neural network embedded with FCA. *Sci Rep.* 2024;14(1):8804.

doi: 10.1038/s41598-024-59311-0

46. Subba T, Chingtham T. Comparative analysis of machine learning algorithms with advanced feature extraction for ECG signal classification. *IEEE Access*. 2024;12:57727-57740.

doi: 10.1109/ACCESS.2024.3387041

47. Gao J, Zhang H, Lu P, Wang Z. An effective LSTM recurrent network to detect arrhythmia on imbalanced ECG dataset. *J Healthc Eng.* 2019;2019(1):6320651.

doi: 10.1155/2019/6320651

48. Chen TM, Huang CH, Shih ES, Hu YF, Hwang MJ. Detection and classification of cardiac arrhythmias by a challenge-best deep learning neural network model. *iScience.* 2020;23(3):100886.

doi: 10.1016/j.isci.2020.100886

49. Sun L, Wang Y, Qu Z, Xiong NN. BeatClass: A sustainable ECG classification system in IoT-based eHealth. *IEEE Internet Things J.* 2021;9(10):7178-7195.

doi: 10.1109/JIOT.2021.3108792

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Exploring the viability of robotic technology integrated with Vivaldi artificial intelligence for functional assessment in amyotrophic lateral sclerosis

Jacopo Luca Casiraghi[1†], Andrea Lizio[1†], Silvia Bolognini[2], David Tessaro[3], Matteo Xia[3], Giacomo Sommavilla[3], Matteo Cestari[3], Elena Carraro[1], Francesca Gerardi[1], Stefano Regondi[1,2], Raffaele Pugliese[2*], Valeria Ada Sansone[1,4], and Federica Cerri[1]

[1]NEuroMuscular Omnicenter, Milan, Italy

[2]Nemo Lab, ASST GOM Niguarda Cà Granda Hospital, Milan, Italy

[3]Omitech, Padova, Italy

[4]Neurorehabilitation Unit, University of Milan, Milano, Italy

Abstract

In this study, we explore the feasibility and efficacy of leveraging Sanbot Elf – a humanoid intelligent assistive robot – integrated with artificial intelligence (AI), specifically the Vivaldi AI system, for functional assessment in amyotrophic lateral sclerosis (ALS) patients. Our investigation involves evaluating and comparing the performance of the Sanbot Elf in administering the ALS Functional Rating Scale–Revised (ALSFRS-R) to that of human operators, using a structured format where patients respond with either "yes" or "no" answers. This approach is intentionally adopted to minimize ambiguity in patient responses. Patients were given the option to respond either verbally or by utilizing the touchscreen display, particularly beneficial for those experiencing dysarthria or hypophonia. In addition, we examined patient emotional responses to this novel approach. A cohort of 28 ALS patients participated in the study, with a subset undergoing longitudinal follow-up assessments. Our results demonstrate strong agreement between human and robotic administrations of the ALSFRS-R, indicating the potential for AI-enabled robotics to accurately assess ALS functional status. Furthermore, the patients' feedback underscores their acceptability of this technology as a supportive tool in healthcare settings. Our findings also highlight the potential benefits of employing robotic devices with algorithmic capabilities, such as the binary tree method, in hospitals. Moreover, such integration has the potential to alleviate operators' workload. Importantly, this research contributes to the burgeoning field of AI-enabled healthcare operations, highlighting the promising role of robotic systems in enhancing functional assessment and management of ALS.

Keywords: Artificial intelligence; Robotic technology; Functional assessment; Amyotrophic lateral sclerosis; Healthcare operations; Longitudinal study

†These authors contributed equally to this work.

***Corresponding author:**
Raffaele Pugliese
(raffaele.pugliese@nemolab.it)

Citation: 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

Received: May 21, 2024

Accepted: July 29, 2024

Published Online: September 27, 2024

1. Introduction

Amyotrophic lateral sclerosis (ALS) is a devastating neurodegenerative disease, characterized by progressive impairment of motor function, which ultimately leads to paralysis and respiratory failure.[1] In most cases, ALS manifests with either a spinal-onset disease characterized by muscle weakness, cramps, and fasciculations, or a bulbar-onset disease marked by dysarthria and dysphagia. Among neurodegenerative disorders, ALS is notably the most rapidly fatal.[2] The progressive deterioration of symptoms culminates in paralysis and ultimately respiratory failure, typically resulting in death within an average period of 3 – 5 years from symptom onset.[3]

The management of ALS heavily relies on the accurate functional assessment, which is used to monitor disease progression and optimize patient care. At present, one of the most recognized tools for monitoring the progression of functional impairments in ALS is the ALS Functional Rating Scale–Revised (ALSFRS-R).[4,5] It is a multidimensional questionnaire of 12 items, which can be divided into four domains (bulbar, fine and gross motor functions, and respiratory functions). For each question, the score ranges from 0 (complete loss of function) to 4 (normal function).

Regulatory agencies recommend using ALSFRS-R in clinical trials, and a study found that the scale was used as the primary endpoint in 82% of therapeutic trials of ALS.[6] The ALSFRS-R is easy to administer, clinically meaningful, and sensitive – properties that position it as an efficient tool to assess patients' disease progression over time and survival. In a 9-month phase III trial of gabapentin, the scale was proved to wield the most predictive power among the examined outcome measures, such as the forced vital capacity and the maximal voluntary isometric contraction.[7-9]

Despite all these advantages, the scale has some drawbacks. First, specific training related to the administration and to the scoring is required for the proper use of ALSFRS-R. Moreover, when a human operator administers the questionnaire, additional information provided by patients and their requests for clarification with respect to unclear questions may be observed in the clinical practice. The combination of these issues could generally result in longer administration times and increased workload for the operators.

In recent years, there has been a growing interest in leveraging robotic technology integrated with artificial intelligence (AI) to enhance functional assessment in ALS patients.[10-12] These technological advancements hold promise for providing objective, standardized, and efficient evaluations while minimizing the burden on health-care providers. The integration of AI algorithms, such as the Vivaldi AI system, further augments the capabilities of robotic devices, potentially improving the accuracy and reliability of functional assessments.[13]

Despite the potential benefits, the adoption of robotic AI-enabled systems in clinical practice for ALS assessment requires rigorous validation and exploration of patient acceptance and usability. This study aims to address these gaps by evaluating the feasibility and efficacy of a Sanbot Elf robotic device equipped with the Vivaldi AI system in administering the ALSFRS-R and assessing patient emotional reactions to this novel approach. Through a comprehensive assessment of functional status and patient perspectives, we seek to elucidate the role of robotic AI technology in advancing healthcare operations and management for ALS patients.

2. Sanbot Elf robot description and architecture

The Sanbot Elf is a humanoid intelligent service robot, developed by Qihan Technology. With its bipedal configuration, it can effortlessly navigate human environments, ensuring seamless interaction through its ergonomic design. The Sanbot Elf has a total size of 92 × 30 × 40 cm and weighing 19 kg (Figure 1).[13-16] Equipped with a trio of cameras – two within the head and one integrated into the chest tablet – Sanbot boasts an HD Camera and a 3D Camera with resolutions of 8.0 MP and 1.0 MP, respectively. The HD Camera offers functionalities such as photography, videography, audio recording, and live streaming. In addition, Sanbot incorporates a microphone in its head, three head sensors, and a rear-projector boasting a 1920 × 720, 60 Hz, 16:9 resolution.

The chest tablet serves as a versatile interface for displaying and utilizing installed applications, operating on a modified Android OS, and facilitating app development through Android Studio and the Sanbot SDK. Sanbot is further enhanced with a plethora of sensors including infrared, touch, and passive infrared sensor, alongside wireless connectivity options such as WIFI, Bluetooth, and ZigBee (Figure 1). The core components of the assistive application consist of an Intel NUC7i7-BNH, which features an Intel i7-7567U processor clocked at 3.50 GHz, 16GB DDR4 RAM, and a 256 GB SSD. This setup serves as both the applications server and the middleware layer. Inter-device communication, including with the robot, occurs through a Wi-Fi router. The combination of the Intel NUC and the robot forms the backbone of the assistive application, boasting ample processing power and resources to effectively execute all intended tasks

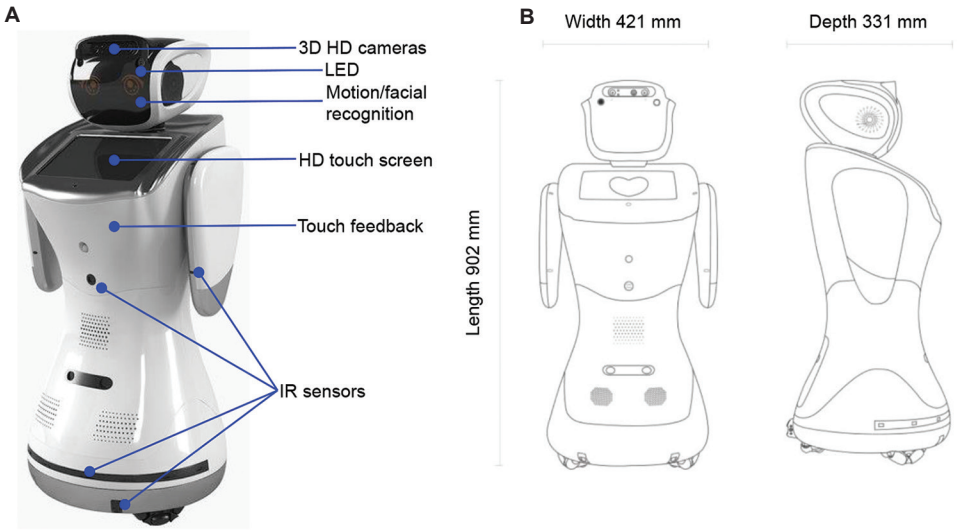


Figure 1. (A and B) Schematic representation and key features of Sanbot Elf. Image created by author.

and activities. Control over Sanbot is facilitated by seven modules or managers accessible through API libraries, encompassing functionalities ranging from voice and hardware control to multimedia and motion control.

In developing the Sanbot application, client-server architecture was adopted.[17] The client component, written in Java for Android, interacts with users and collects data to personalize user-robot interactions. The server, also coded in Java and hosted on a dedicated machine, acts as the central processing unit or "brain" of the robot in the cloud. The architecture comprises various services including emotion and face recognition, user modeling, and speech-to-text functionality.

The server orchestrates communication between different software components, each offering distinct services such as emotion recognition using Affectiva libraries, face recognition with the "FaceRecognition.py" library,[18] and speech-to-text conversion through the Wit service.[13] Given the diverse languages and libraries involved, a central Java server was devised to manage requests and data storage in a central database, facilitating seamless integration and operation. Communication occurs between the client and the Java server, as well as between the Java server and internal and external components.

The Sanbot application commences with a natural language dialogue aimed at user profiling, encompassing factors such as facial recognition, name, gender, age, interests, and mood. This information is utilized by the user modeling component to tailor subsequent interactions, enabling Sanbot to adapt its behavior, offer recommendations, and encourage positive behaviors through persuasive techniques.

With these features, the Sanbot Elf is a versatile service robot designed to fulfill multiple roles, particularly in healthcare settings. At hospitals, it functions adeptly as a receptionist, warmly welcoming and assisting visitors. It excels in providing comprehensive information about the facility, guiding visitors to various departments, and ensuring smooth navigation throughout the premises. Within hospital departments, the Sanbot Elf engages with patients in a compassionate and supportive manner. It enhances patient experience by offering entertainment options such as music and videos and providing companionship during periods of loneliness or anxiety. In addition, the robot enhances security measures by actively monitoring its surroundings and promptly notifying staff of any unusual situations or emergencies that may arise. Moreover, the Sanbot Elf facilitates telemedicine sessions by seamlessly connecting patients with healthcare providers remotely, thereby enhancing access to healthcare services.

3. Related works

Assistive robotics are gaining significant attention for their potential to enhance the quality of life for various patients, especially those with neurological disorders such as Alzheimer's disease, cognitive impairments, neuromuscular diseases, spinal cord injuries, to name a few. These conditions often lead to declines in cognitive and motor functions, severely impacting daily living activities. By improving the management of these robotic systems, it is possible to alleviate the burden on physicians, caregivers, and family members. These robots offer a range of functions, including connecting patients with distant family members, providing companionship, promoting health, and assisting with daily tasks. Furthermore, Gao

et al.[19] outlined how such robotic systems can be useful for the management infectious diseases, including disease prevention and monitoring, clinical care, laboratory automation, and logistics, during the COVID-19 pandemic. However, challenges such as difficult operation, maintenance, reliability, and high costs hinder their widespread adoption in private homes.

Despite these challenges, numerous research efforts have successfully tested assistive robotics in real-world settings, demonstrating their acceptance and usability in personalized daily assistive plans.[20] This progress highlights the potential for these technologies to become integral components of patient care, improving the overall management and quality of life for individuals with neurological disorders.[21]

For instance, Ghafurian *et al.*[22] examined the use of social robots for dementia care, highlighting various levels of robot autonomy and assistive objectives such as connecting patients with distant family members, providing companionship, promoting health, and aiding in daily tasks. The authors underscored the need for robots to be easy to use and maintain, and reliable, as these factors significantly impact their adoption in private homes. Saunders *et al.*[23] reported the use of the Care-O-Bot 3 for complex interactions with elderly people, facilitating independent living and daily tasks. Schroeter *et al.*[24] reported the usability of the social assistive robot, which provides questionnaires and cognitive stimulation games to monitor cognitive decline to be used in a home setting. Instead, Fischinger *et al.*[25] developed a care robot (namely Hobbit robot prototype) for aging by means of fall prevention/detection. The Hobbit project combined research from robotics, gerontology, and human-robot interaction to develop a care robot, which is capable of fall prevention and detection as well as emergency detection and handling. Moreover, to enable daily interaction with the robot, other functions were added, such as bringing objects, offering reminders, and entertaining. Casey *et al.*[26] have demonstrated that robots could promote social connection and reduce loneliness among dementia patients in a long-term residential setting. Finally, Neerincx *et al.*[27] designed a personalized assistive robot for patients with cognitive impairments, taking into account their daily needs and characteristics such as their personal cognitive, emotional, and psychological status as well as their cultural background, which are not always easily modeled according to well-defined classifications.

4. Data and methods

4.1. Participants

This study enrolled a cohort of 28 ALS patients who underwent evaluation at the Nemo Clinical Center in Milan. Participants met the diagnostic criteria for probable or definite ALS as outlined in the revised El Escorial criteria.[28] Inclusion criteria required ALS patients to exhibit a minimum ALSFRS-R score of ≥2 points in both speech and handwriting questions, with a total ALSFRS-R score of ≥18 points. In addition, all participants were capable of verbal communication to answer interview questions. Patients displaying severe cognitive and behavioral impairment, as determined by the Edinburgh Cognitive Assessment Scale at enrollment, were excluded from the study. Before participation, all patients provided informed consent, which was approved by the Local Ethics Committee, Grande Ospedale Metropolitano Niguarda, Milano, Italy (Protocol Number: 404-092019).

4.2. Vivaldi AI system

Sanbot Elf is enhanced by the cutting-edge Vivaldi AI system, which equips it with advanced capabilities such as speech recognition, facial expression analysis, and gesture understanding. Unlike conventional setups reliant on remote cloud servers, this AI engine processes data locally on the robot itself, facilitating faster response times and potentially reducing operational costs by minimizing data transfers. With Vivaldi AI at its core, Sanbot Elf adeptly comprehends human communication, enabling it to engage with users in a natural and intelligent manner. Key functionalities include: (1) speech recognition, facilitating effortless voice interaction with users; (2) natural language processing, enabling the robot to understand and respond to human language nuances; (3) facial expression analysis, allowing Sanbot Elf to discern user sentiment by interpreting facial cues, thereby tailoring its responses accordingly; (4) gesture understanding, enhancing the fluidity of communication between the robot and users; and (5) seamless communication, ensuring smooth and intuitive interaction between Sanbot Elf and its human counterparts.

4.3. Robotic administration of ALSFRS-R questionnaire

For this study, the Sanbot Elf unit is equipped to administer the ALSFRS-R questionnaire using a structured format allowing patients to answer questions with either "yes" or "no." This approach is intentionally adopted to minimize ambiguity in patient responses. Patients have the option to respond either verbally or by utilizing the touchscreen display, particularly beneficial for those experiencing dysarthria or hypophonia. The Vivaldi AI system guides patients through the ALSFRS-R questionnaire, dynamically adjusting the sequence of questions based on the patient's responses. This ensures a personalized experience tailored to each individual's needs. Throughout the questionnaire,

patients have the flexibility to request repetitions of questions or take brief pauses using vocal commands. On completion of all 12 topics, Vivaldi AI system yields a final score for further analysis. Figure 2 illustrates an example of this process, providing a visual representation of the interaction between the patient and the robotic system.

4.4. Clinical assessment

For each participant, a comprehensive set of demographic and clinical characteristics was documented, encompassing age at onset and evaluation, sex, site of onset (bulbar/spinal), and disease duration, calculated as the time interval between onset and evaluation. Furthermore, ALS patients underwent a battery of functional and psychological assessments, including:

(1) ALSFRS-R: Initially administered by the same human operator and subsequently by the robotic operator following a brief orientation in response to the robot.
(2) State-Trait Anxiety Inventory Form Y-1 and Form Y-2 (STAI Y-1 and Y-2): Assessed anxiety levels before and after administration of the ALSFRS-R questionnaire by both human and robotic operators.[29]
(3) Big Five Inventory (BFI): Evaluated the personality dimension "Openness to Experience" of patients before administration of the ALSFRS-R questionnaire by both human and robotic operators.[30]
(4) Observational Grid: Employed by a psychologist during robotic assessment to assess the emotional state of patients at different intervals during the questionnaire administration. This grid also documented patient reactions and behaviors.
(5) Semi-Structured Interview: Designed specifically for this study to explore patients' subjective experiences post-administration of the questionnaire. Thematic areas covered emotions during the robot interaction, perception of the experience, openness toward robotics, and future developments in robotics in hospitals.

In addition, a subgroup of 20 patients underwent longitudinal follow-up, completing the ALSFRS-R questionnaire (administered by both human and robotic operators) during the second time, which was scheduled at 4 – 12 months from the initial evaluation. Sixteen of these patients also underwent psychological assessment at the same timepoint, including the completion of STAI Y-1 and Y-2 questionnaires and the observational grid for assessing emotional states during robot interaction.

4.5. Statistical analysis

Shapiro–Wilk test and Levene's test were used to test the normality of the distribution and the homogeneity of the variance, respectively. Data are reported as median and interquartile range for continuous variables, and number and percentage for categorical ones.

To assess the agreement between the two methods of administering the questionnaire, we used the Bland–Altman plot and the intraclass correlation coefficient (ICC). The

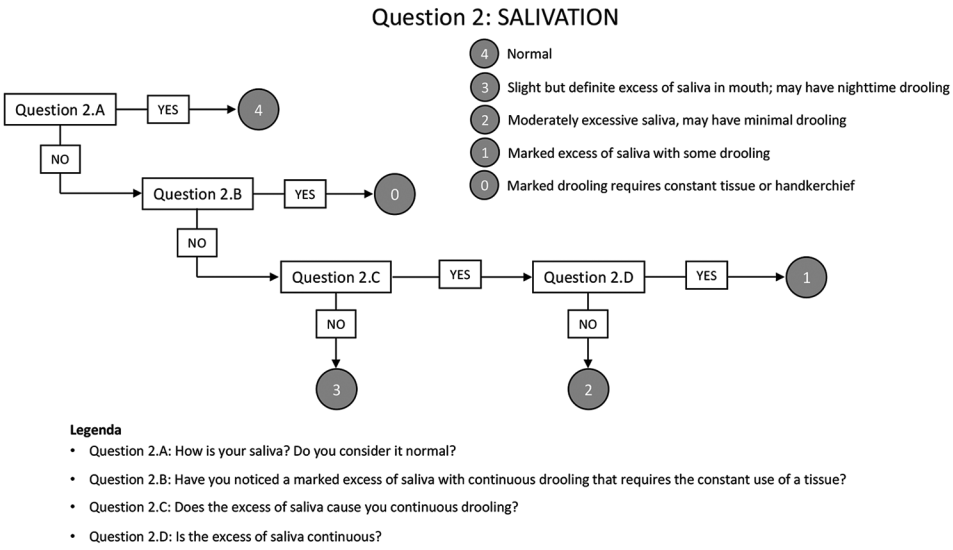


Figure 2. Illustration of the Sanbot Elf unit administering the ALSFRS-R questionnaire with the assistance of the Vivaldi AI system. The structured format allows patients to respond with "yes" or "no" answers, minimizing ambiguity. Patients can interact through verbal responses or touchscreen, with the latter employed to accommodate those with dysarthria or hypophonia. The Vivaldi AI dynamically adjusts question sequences based on responses, offering a personalized experience. Vocal commands enable patients to request repetitions or pauses. On completion, the Vivaldi AI provides a final score for analysis. Abbreviations: ALSFRS-R: Amyotrophic lateral sclerosis functional rating scale-revised; AI: Artificial intelligence.

Bland–Altman plot is a graphical method to compare two measurements techniques,[31,32] also providing a quantitative estimate of how closely the values from two measurements lie.[33] In this context, we reported both the bias (mean of the differences between the two methods) and the 95% limits of agreement (LOA). The ICC provides a single measure of the extent of agreement between the two methods,[33] and we considered the guidelines recommended by Koo and Li for interpreting the results.[34] In detail, basing on the 95% confident interval of the ICC estimate, values <0.5, between 0.5 and 0.75, between 0.75 and 0.9, and >0.90 are indicative of poor, moderate, good, and excellent reliability, respectively.[34] Moreover, to longitudinally quantify the heterogeneity in rates of decline in ALSFRS-R between individuals, we calculated the coefficient of variation (CoV). The CoV was defined as the between-patient standard deviation of slope divided by the mean rate of change, and a lower value indicates both a less variation among patients, which could positively affect sample size calculations, and an increase of the sensitivity in detecting disease progression. Finally, a mixed-effects linear regression with ALSFRS-R (total score and subgroups, separately) as the outcome, and fixed-effects for time, approach (human operator vs. robotic operator), and an interaction between time and approach. In detail, the fixed effect for the interaction between approach and time was used as an assessment of bias in the human-based score compared to the robot-based score. Considering the psychological assessment, a paired *t*-test was used to evaluate the impact of the questionnaire's administration by the robotic operator on patients' anxiety, comparing the state anxiety level pre- and post-administration. Finally, the correlation between the bias in evaluating the ALSFRS-R between the two methods and the "openness to experience" dimension of the BFI was evaluated using Spearman's rank correlation coefficient. Regarding the qualitative analysis collected from the observational grid and the semi-structured interview, frequency and percentage were used to investigate the emotional states and the patients' perception about the experience. Specifically, considering the analysis of the semi-structured interview, a cluster's identification was obtained by analyzing the answers given by patients for the categories of each thematic area. All statistical tests were two-tailed, and $P < 0.05$ was considered statistically significant. All the statistical analyses were performed using SAS 9.3 (SAS Institute, Inc, Cary, NC) software. For the text analysis, the Atlas.ti software was used.

5. Results

5.1. Design and implementation of the ALSFRS-R in Sanbot Elf

The ALSFRS-R questionnaire is an essential tool for evaluating the functional status of ALS patients. In our study, we integrated this questionnaire into the Sanbot Elf robot, leveraging the advanced capabilities of the Vivaldi AI system to create an adaptive and user-friendly interface for patients.

The ALSFRS-R questionnaire, embedded in the chest tablet interface of the Sanbot Elf, comprises 12 items that assess four key domains: bulbar function, fine motor function, gross motor function, and respiratory function. Each item is scored 0 (complete loss of function) to 4 (normal function), enabling a thorough evaluation of the patient's abilities.

In a hospital setting, the Sanbot Elf administered the ALSFRS-R questionnaire through an interactive process tailored to the specific needs of ALS patients, with a typical duration of approximately 10 – 15 min. The Vivaldi AI system guided this administration, ensuring a smooth and adaptive experience. Patients responded to each question using a simple "yes" or "no" format, which minimizes ambiguity and ensures clarity in their responses, and the Vivaldi AI system dynamically adjusted the sequence of questions based on the patient's responses.

Sanbot Elf offers multiple interaction modes to accommodate the diverse needs of ALS patients. Indeed, patients with limited mobility but who can speak were able to interact verbally with the robot. For those experiencing dysarthria or hypophonia, the touchscreen interface provided an alternative method for responding to questions. This dual-mode interaction ensured inclusivity and accessibility for all patients.

In addition, patients could use vocal commands to request repetitions of questions or take brief pauses, a feature facilitated by the Vivaldi AI system (this could cause an increase in the time required to administer the questionnaire, without negatively influencing the operator's burden). This enhanced comfort and ease of use, catering to patients who might need more time or clarification.

On completion of the questionnaire, the Vivaldi AI system automatically calculated the final ALSFRS-R score, promptly providing results. This score is useful for ongoing monitoring and assessment of the patient's functional status and disease progression. The automated scoring process ensures accuracy and consistency, minimizing the risk of human error.

5.2. Agreement analysis between human and robotic ALSFRS-R assessment

A group of 28 ALS patients participated in this study, with a median age at evaluation of 62.37 years (range: 53.88 – 68.43) and a male-to-female ratio of 2.11. Descriptive characteristics of the ALS cohort, including demographic and clinical features, are summarized in Table 1.

In terms of ALSFRS-R total score agreement, the Bland–Altman plot revealed a bias (mean difference) of −0.18 points, with 95% LOA ranging from −4.35 to 3.99 points (Figure 3A). In addition, the ICC of 0.95 (95% CI: 0.90 – 0.98) indicated good to excellent agreement between the robotic- and human-administered questionnaires (Table 2), consistent with the classification by Koo and Li.[34]

For ALSFRS-R subscores, separate Bland–Altman plots were generated for bulbar (Figure 3B), motor (Figure 3C), and respiratory (Figure 3D) domains. The bulbar subscore exhibited a bias of 0.14 points, with 95% LOA from −1.24 to 1.53 points. Similarly, the motor subscore showed a bias of −0.57 points, with 95% LOA from −4.54 to 3.40 points, while the respiratory subscore had a bias of 0.25 points, with 95% LOA from −1.02 to 1.52 points.

Table 1. Descriptive characteristics of ALS patients (*n*=28)

Parameter	Data
Age at evaluation	62.37 (53.88 – 68.43)
Sex, *n* (%)	
Male	19 (67.86)
Female	9 (32.14)
Age at onset	59.97 (50.18 – 67.70)
Site of onset, *n* (%)	
Bulbar	0 (0.00)
Spinal	28 (100.00)
Disease duration*	24.80 (16.87 – 33.60)
Diagnostic delay	10.13 (7.10 – 16.20)
ALSFRS-R** total score	38.50 (35.00 – 41.00)
ALSFRS-R bulbar subscore	12.00 (10.00 – 12.00)
ALSFRS-R motor subscore	16.00 (13.00 – 17.00)
ALSFRS-R respiratory subscore	12.00 (11.00 – 12.00)

Notes: *Time between onset and evaluation; **Measured by the human operator; all the values are expressed as median (interquartile range) unless otherwise indicated.

Table 2. Agreement between operator and robotic device in the ALSFRS-R (total and subscores) administration

	ICC (95% CI)	Bias	95% LOA
ALSFRS-R total score	0.95 (0.90 – 0.98)	−0.18	−4.35 – 3.99
ALSFRS-R bulbar subscore	0.77 (0.57 – 0.88)	0.14	−1.24 – 1.53
ALSFRS-R motor subscore	0.92 (0.84 – 0.96)	−0.57	−4.54 – 3.40
ALSFRS-R AASS	0.89 (0.78 – 0.95)	−0.07	−2.84 – 2.70
ALSFRS-R AAII	0.89 (0.78 – 0.95)	−0.50	−2.97 – 1.97
ALSFRS-R respiratory subscore	0.97 (0.94 – 0.99)	0.25	−1.02 – 1.52

Abbreviations: ALSFRS-R: Amyotrophic lateral sclerosis Functional Rating Scale–Revised; CI: Confidence interval; LOA: Limits of agreement.

Furthermore, ICC values of 0.77 (95% CI: 0.57 – 0.88) for the bulbar domain, 0.92 (95% CI: 0.84 – 0.96) for the motor domain, and 0.97 (95% CI: 0.94 – 0.99) for the respiratory domain indicated moderate to excellent agreement between the robotic- and human-administered questionnaires (Table 2).

Regarding longitudinal analysis, Table 3 displays monthly rates of change for ALSFRS-R total scores and subscores, along with CoVs. While all outcomes demonstrated significant declining trends over time in both administration methods, no significant bias between human and robot-based administration was observed, as indicated by the mixed-effects linear regression interaction terms ($P > 0.05$ for all domains).

5.3. Impact of robotic device usage on patients' anxiety and openness to experience dimension

Our findings revealed a statistically significant decrease in state anxiety before and after the robotic administration of the ALSFRS-R questionnaire (35.50 [30.00 – 40.00] vs. 31.00 [28.00 – 36.00]; P = 0.0218). Despite this decrease, both pre- and post-administration levels of anxiety remained below the clinical threshold of 41, indicating a general absence of clinical anxiety even before robotic administration. Specifically, before the questionnaire's administration by the robot, 17 ALS patients (60.71%) showed no clinical state anxiety, while the remaining 11 patients (39.29%) reported slightly elevated anxiety levels. After robotic administration, five out of these 11 patients (17.86%) experienced a decrease in anxiety levels below the clinical threshold.

In a longitudinal analysis of 16 patients who underwent psychological evaluation at both baseline and follow-up, no significant differences in state anxiety before and after robotic administration were observed, neither at baseline nor at the end of the follow-up period. However, a trend of anxiety reduction was more pronounced during the initial evaluation (effect size = 0.42) compared to the final assessment (effect size = 0.36). Furthermore, we found no significant correlations between the BFI score for openness to experience and the magnitude of difference in ALSFRS-R scores measured by human operators versus robotic devices, for both total scores and individual domains.

In the observational grid analysis, notable patterns emerged during various stages of the interaction between patients and the robotic device. During the initial training phase (T1), the majority of patients (67.85%) displayed curiosity and interest, indicating a high level of engagement with the administration process. Conversely, 39.28% of patients expressed a recurring sense of puzzlement during this phase. After the first question posed by the robotic

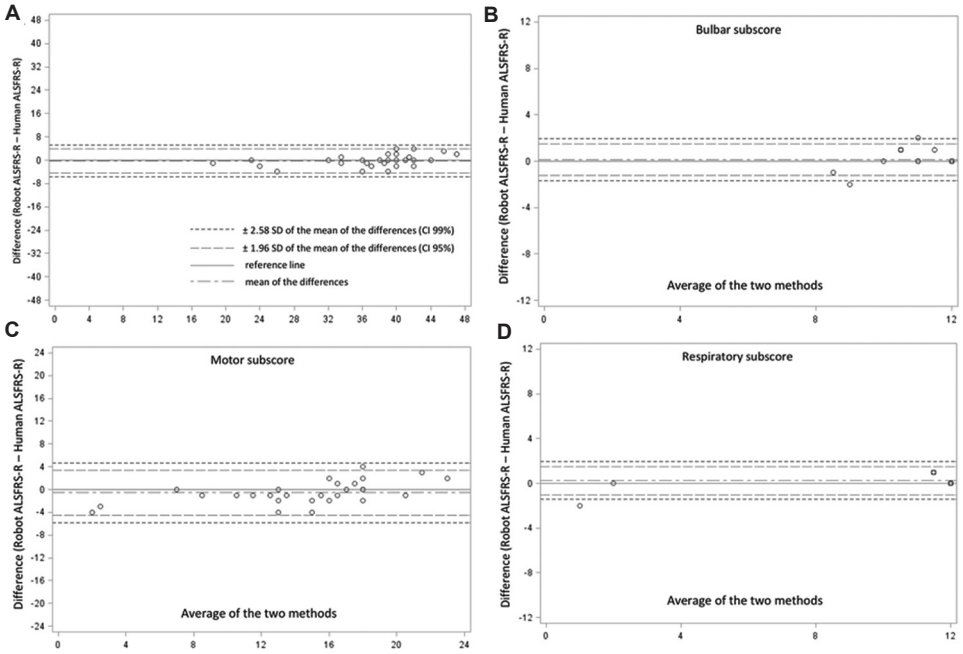


Figure 3. Evaluation of agreement between robotic and human-administered ALSFRS-R total scores and subscores. (A) Bland–Altman plot illustrates a bias of −0.18 points with 95% LOA ranging from −4.35 to 3.99 points for the total score. Intraclass correlation coefficient of 0.95 indicates good to excellent agreement. Separate Bland–Altman plots for (B) bulbar, (C) motor, and (D) respiratory subscores demonstrate biases and 95% LOA for each domain.
Abbreviations: ALSFRS-R: Amyotrophic lateral sclerosis functional rating scale-revised; AI: Artificial intelligence; LOA: Limits of agreement.

Table 3. Comparison between longitudinal rates of change during follow-up period

	Human operator			Robotic operator		
	Slope* (95% CI)	*P*-value**	CoV	Slope (95% CI)	*P*-value**	CoV
ALSFRS-R total score	−0.50 (−0.77 – −0.23)	0.0002	1.24	−0.55 (−0.86 – −0.24)	0.0006	1.27
ALSFRS-R bulbar subscore	−0.06 (−0.12 – −0.01)	0.0273	2.00	−0.08 (−0.14 – −0.02)	0.0137	1.75
ALSFRS-R motor subscore	−0.36 (−0.50 – −0.22)	<0.0001	0.89	−0.38 (−0.59 – −0.17)	0.0020	1.26
ALSFRS-R AASS	−0.15 (−0.23 – −0.07)	0.0034	1.27	−0.16 (−0.28 – −0.03)	0.0166	1.75
ALSFRS-R AAII	−0.21 (−0.32 – −0.10)	0.0008	0.95	−0.22 (−0.35 – −0.09)	0.0028	1.36
ALSFRS-R respiratory subscore	−0.07 (−0.20 – 0.05)	0.6250	4.14	−0.09 (−0.24 – 0.06)	0.5000	3.89

Notes: * Slope is the mean monthly rate of change during follow-up; ** *P* value of slope.
Abbreviations: ALSFRS-R: Amyotrophic lateral sclerosis Functional Rating Scale-Revised; CI: Confidence interval; CoV: Coefficient of variation.

operator (T2), curiosity and interest remained prevalent (67.85%), with some patients also experiencing amusement (10.71%) and neutrality (7.14%). By the midpoint of the questionnaire (T3), a decrease in curiosity and interest (25%) and puzzlement (17.85%) was observed compared to earlier stages. Confidence in using the instrument was reported by 46.42% of patients at this stage. At the conclusion of the test (T4), a significant portion of patients (64.28%) exhibited a sense of puzzlement, while confidence levels decreased (14.28%). Some patients also displayed amusement (10.71%) and curiosity/interest (10.71%). Further analysis revealed that emotional reactions peaked during questions 4 (handwriting) and 7 (turning in bed and adjusting bed clothes), with patients often seeking clarification from the human operator due to the questions' ambiguous nature (Figure 4).

During follow-up observations, shifts in emotional states were noted. In the initial training phase (T1), patients exhibited a more neutral attitude (28.57%) compared to the baseline period. However, interest and attention persisted (17.85%).

Similarly, after the second question (T2), patients remained attentive and interested (25%), though some displayed a neutral attitude (14.28%) possibly due to increased familiarity with the process. At T3, patients

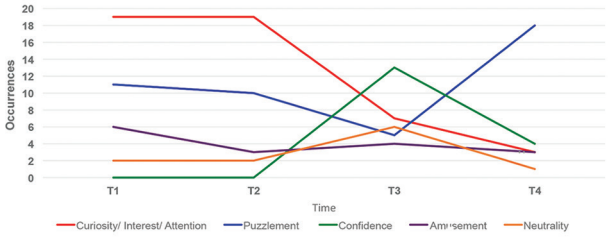


Figure 4. Observational grid analysis of patient interactions with the robotic device throughout the ALSFRS-R questionnaire administration process. Emotional reactions and engagement levels are depicted across various stages (T1–T4) of the interaction. Peaks in emotional responses coincide with specific questionnaire items, highlighting areas of confusion and the need for clarification. Notable shifts in patient engagement and confidence levels are observed, shedding light on the dynamics of human-robot interaction during medical assessments.
Abbreviation: ALSFRS-R: Amyotrophic lateral sclerosis functional rating scale-revised.

maintained their focus on providing answers (14.28%), while some displayed neutrality (14.28%), puzzlement (10.71%), and confidence (10.71%). By the end of the questionnaire (T4), patients demonstrated confidence in using the robot (17.85%), accompanied by amusement (14.28%) and interest in the experience (10.71%), with fewer instances of puzzlement compared to baseline evaluation.

6. Discussion

The urgent need for precise symptom monitoring in ALS due to its progressive and ultimately fatal nature underscores the quest for more accurate assessment tools. While the ALSFRS-R is widely regarded as reliable, concerns regarding subjectivity in score attribution persist. Moreover, recent advancements in the realm of neuromuscular diseases have seen the emergence of robotics integrated with AI technology.

To address the challenges posed by subjective scoring and explore the potential of robotics in neuromuscular care, our study delved into the accuracy and agreement of ALSFRS-R scores obtained from a robotic operator enhanced with AI algorithms, employing the binary tree method, compared to those from a human operator. In addition, we examined patients' emotional states and perceptions during interactions with the robotic system.

By investigating the efficacy of AI-powered robotics in symptom assessment, we aim to mitigate scoring arbitrariness and enhance the precision of ALS monitoring. Moreover, understanding patients' experiences and emotions during interactions with robotic technology provides crucial insights into the feasibility and acceptance of such innovations in clinical practice.

The agreement analysis between human and robotic ALSFRS-R assessment demonstrated promising results. Indeed, the Bland–Altman plot and ICC revealed good to excellent agreement between the robotic and human-administered questionnaires for both total scores and individual domains (bulbar, motor, and respiratory). This indicates that the robotic device equipped with AI technology can effectively administer the ALSFRS-R questionnaire, providing comparable results to those obtained by human operators. Furthermore, the Vivaldi AI system based on dichotomous answers ("yes" or "no") reduces answers ambiguity.

In addition, longitudinal analysis showed significant declining trends in functional status over time, with no significant differences observed between human- and robot-based administration methods. This suggests that the robotic device can accurately track disease progression, making it a valuable tool for monitoring ALS patients over time.

The study also evaluated the impact of robotic device usage on patients' anxiety and openness to experience. The findings revealed a statistically significant decrease in state anxiety after robotic administration of the ALSFRS-R questionnaire, although anxiety levels remained below the clinical threshold even before robotic administration. This suggests that the use of robotic technology in healthcare settings may contribute to reducing patient anxiety levels, potentially improving the overall patient experience.

There were no significant correlations between patients' openness to experience and the difference in ALSFRS-R scores measured by human operators versus robotic devices, indicating that patients' personality traits did not influence the accuracy of robotic assessments.

Finally, the observational grid analysis provided valuable insights into patients' emotional reactions during interactions with the robotic device. Patients initially exhibited curiosity and interest, although some also experienced puzzlement. However, as patients became more familiar with the robotic device, confidence in using the instrument increased, accompanied by reduced levels of puzzlement and increased interest and amusement. This suggests that with proper training and familiarization, patients can become more comfortable with robotic technology, enhancing the acceptability and usability of such devices in clinical settings.

7. Conclusion

The study highlights the potential of robotic technology integrated with AI to enhance functional assessment and care for ALS patients. The findings support the feasibility

and efficacy of using robotic devices for administering the ALSFRS-R questionnaire and tracking disease progression over time. In addition, it can be assumed that high levels (not clinically significant) of anxiety were initially due to the novelty and curiosity of the first experience that triggered a momentary state of activation and tension but decreased subsequently because they became familiar with the experience. Specifically, before the robot-administered questionnaire, 60.71% of ALS patients exhibited no clinical state anxiety, while 39.29% reported mildly elevated anxiety levels. In post-robotic administration, 17.86% experienced a reduction in anxiety levels below the clinical threshold. Furthermore, in the longitudinal study involving 16 patients who underwent psychological evaluations at both baseline and follow-up, no significant differences in state anxiety were observed before and after robotic administration at either time point. However, there was a more pronounced trend of anxiety reduction during the initial evaluation (effect size = 0.42) compared to the final assessment (effect size = 0.36).

However, further research is needed to validate these findings and explore the long-term impact of robotic technology on patient outcomes in ALS management; using a larger sample size would provide greater statistical power and enhance the generalizability of the findings. Despite this, ongoing advancements in robotics and AI technology present exciting opportunities for further innovation in ALS care. Research efforts should focus on refining robotic systems to address specific needs and challenges faced by ALS patients, such as respiratory support, mobility assistance, and communication aids. Collaborative efforts between clinicians, engineers, and patients are essential to develop tailored solutions that maximize the benefits of robotic technology while addressing the unique requirements of individuals living with ALS. In addition, longitudinal studies tracking the real-world implementation of robotic AI-enabled systems in ALS and neuromuscular diseases in general will be crucial to evaluate their long-term efficacy, cost-effectiveness, and impact on patient outcomes.

Finally, the assistive and collaborative robots have the potential to take on a larger role in various tasks in the future. We believe that our research can contribute to this by exploring how humans and robots can effectively collaborate in a clinical setting. The goal is not necessarily a complete replacement, but rather a complementary approach where robots handle repetitive or hazardous tasks, freeing humans for more strategic endeavors.

Acknowledgments

None.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: J.L. Casiraghi, A. Lizio, F. Cerri
Investigation: S. Bolognini, A. Lizio, J.L. Casiraghi, D. Tessaro, M. Xia, G. Sommavilla, M. Cestari, E. Carraro, F. Gerardi, V.A. Sansone, F. Cerri
Methodology: J.L. Casiraghi, A. Lizio
Formal analysis: S. Bolognini, A. Lizio, F. Cerri, R. Pugliese
Writing–original draft: J. Casiraghi, F. Cerri, S. Bolognini, A. Lizio, V.A. Sansone, R. Pugliese
Writing–review & editing: R. Pugliese, S. Regondi, F. Cerri

Ethics approval and consent to participate

Before participation, all patients provided informed consent, which was approved by the Local Ethics Committee (Protocol Number: 404-092019).

Consent for publication

All patients provided written informed consent, which was approved by the Local Ethics Committee (Protocol Number: 404-092019).

Availability of data

Data are available from the corresponding author on reasonable request.

References

1. Hardiman O, Al-Chalabi A, Chio A, *et al*. Amyotrophic lateral sclerosis. *Nat Rev Dis Primers*. 2017;3:17071.
doi: 10.1038/nrdp.2017.71
2. Kiernan MC, Vucic S, Cheah BC, *et al*. Amyotrophic lateral sclerosis. *Lancet*. 2011;377(9769):942-955.
doi: 10.1016/S0140-6736(10)61156-7
3. Brown RH, Al-Chalabi A. Amyotrophic lateral sclerosis. *N Engl J Med*. 2017;377(2):162-172.
doi: 10.1056/NEJMra1603471
4. Chio A, Calvo A, Moglia C, Mazzini L, Mora G, PARALS Study Group. Phenotypic heterogeneity of amyotrophic lateral sclerosis: A population based study. *J Neurol Neurosurg Psychiatry*. 2011;82(7):740-746.
doi: 10.1136/jnnp.2010.235952
5. Franchignoni F, Mora G, Giordano A, Volanti P, Chio A.

Evidence of multidimensionality in the ALSFRS-R Scale: A critical appraisal on its measurement properties using Rasch analysis. *J Neurol Neurosurg Psychiatry*. 2013;84(12):1340-1345.

doi: 10.1136/jnnp-2012-304701

6. European College of Neuropsychopharmacology. Clinical investigation of medicinal products for treatment of Amyotrophic Lateral Sclerosis (ALS). *Eur Neuropsychopharmacol*. 2001;11(2):187-189.

doi: 10.1016/s0924-977x(01)00067-0

7. Gordon PH, Miller RG, Moore DH. Alsfrs-R. *Amyotroph Lateral Scler Other Motor Neuron Disord*. 2004;5(Suppl 1):90-93.

doi: 10.1080/17434470410019906

8. Kaufmann P, Levy G, Thompson JLP, *et al*. The ALSFRSr predicts survival time in an ALS clinic population. *Neurology*. 2005;64(1):38-43.

doi: 10.1212/01.WNL.0000148648.38313.64

9. Miller RG, Moore DH 2nd, Gelinas DF, *et al*. Phase III randomized trial of gabapentin in patients with amyotrophic lateral sclerosis. *Neurology*. 2001;56(7):843-848.

doi: 10.1016/s0022-510x(01)00632-3

10. Pugliese R, Sala R, Regondi S, Beltrami B, Lunetta C. Emerging technologies for management of patients with amyotrophic lateral sclerosis: From telehealth to assistive robotics and neural interfaces. *J Neurol*. 2022;269(6):2910-2921.

doi: 10.1007/s00415-022-10971-w

11. Maier A, Eicher C, Kiselev J, *et al*. Acceptance of enhanced robotic assistance systems in people with amyotrophic lateral sclerosis-associated motor impairment: Observational online study. *JMIR Rehabil Assist Technol*. 2021;8(4):e18972.

doi: 10.2196/18972

12. Coser O, Tamantini C, Soda P, Zollo L. AI-based methodologies for exoskeleton-assisted rehabilitation of the lower limb: A review. *Front Robot AI*. 2024;11:1341580.

doi: 10.3389/frobt.2024.1341580

13. Botta M, Camilleri D, Cena F, *et al*. Cloud-based user modeling for social robots: A first attempt. *arXiv*. 2022.

14. Beraldo G, Menegatti E, de Tommasi V, Mancin R, Benini F. A Preliminary Investigation of Using Humanoid Social Robots as Non-pharmacological Techniques With Children. In: *15th IEEE International Conference on Advanced Robotics and its Social Impacts*; 2019.

doi: 10.1109/ARSO46408.2019.8948760

15. Kimmig R, Verheijen RHM, Rudnicki M, for SERGS Council. Robot assisted surgery during the COVID-19 pandemic, especially for gynecological cancer: A statement of the Society of European Robotic Gynaecological Surgery (SERGS). *J Gynecol Oncol*. 2020;31(3):e59.

doi: 10.3802/jgo.2020.31.e59

16. Bauer J, Dengler S, Faubel L, *et al*. Pandemic robot. *Curr Dir Biomed Eng*. 2021;7(2):601-604.

doi: 10.1515/cdbme-2021-2153

17. Di Napoli C, Ercolano G, Rossi S. Personalized home-care support for the elderly: A field experience with a social robot at home. *User Model User Adap Inter*. 2022;33:405-440.

doi: 10.1007/s11257-022-09333-y

18. Gena C, Botta M, Cena F, Mattutino C. User modeling for social robots. *arXiv*. 2021.

doi: 10.5281/zenodo.4781442

19. Gao A, Murphy RR, Chen W, *et al*. Progress in robotics for combating infectious diseases. *Sci Robot*. 2021;6(52):eabf1462.

doi: 10.1126/scirobotics.abf1462

20. Amabili G, Maranesi E, Margaritini A, *et al*. Usability and feasibility assessment of a social assistive robot for the older people: Results from the GUARDIAN project. *Bioengineering* (*Basel*). 2023;11(1):20.

doi: 10.3390/bioengineering11010020

21. Luperto M, Monroy J, Renoux J, *et al*. Integrating social assistive robots, IoT, virtual communities and smart objects to assist at-home independently living elders: The movecare project. *Int J Soc Robot*. 2023;15(3):517-545.

doi: 10.1007/s12369-021-00843-0

22. Ghafurian M, Hoey J, Dautenhahn K. Social robots for the care of persons with dementia: A systematic review. *ACM Trans Hum Robot Interact*. 2021;10(4):1-31.

doi: 10.1145/346965

23. Saunders J, Syrdal DS, Koay KL, Burke N. "Teach me-show me"-End-user personalization of a smart home and companion robot. *IEEE Trans Hum Mach Syst*. 2016;46(1):27-40.

24. Schroeter C, Mueller S, Volkhardt M, *et al*. Realization and User Evaluation of a Companion Robot for People with Mild Cognitive Impairments. In: *2013 IEEE International Conference on Robotics and Automation*. Vol. 1. IEEE; 2013.

doi: 10.1109/ICRA.2013.6630717

25. Fischinger D, Einramhof P, Papoutsakis K, *et al*. Hobbit, a care robot supporting independent living at home: First prototype and lessons learned. *Robot Auton Syst*. 2016;75(Part A):60-78.

doi: 10.1016/j.robot.2014.09.029

26. Casey D, Felzmann H, Pegman G, *et al*. What people with dementia want: Designing MARIO an acceptable robot companion. In: *Computers Helping People with Special*

Needs. Vol. 1. Cham: Springer; 2016.

doi: 10.1007/978-3-319-41264-1_44

27. Neerincx MA, van Vught W, Blanson Henkemans O, *et al*. Socio-cognitive engineering of a robotic partner for child's diabetes self-management. *Front Robot AI*. 2019;6:118.

doi: 10.3389/frobt.2019.00118

28. Brooks BR, Miller RG, Swash M, Munsat TL, World Federation of Neurology Research Group on Motor Neuron Diseases. El Escorial revisited: Revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Other Motor Neuron Disord*. 2000;1(5):293-299.

doi: 10.1080/146608200300079536

29. Shahid A, Wilkinson K, Marcu S, Shapiro CM. *State-Trait Anxiety Inventory* (*STAI*). United Kingdom: Psychology Press; 2011.

doi: 10.1007/978-1-4419-9893-4_90

30. John OP, Naumann LP. Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In: *Handbook of Personality: Theory and Research*. New York: The Guilford Press; 2008.

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

32. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135-160.

doi: 10.1177/096228029900800204

33. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: Measures of agreement. *Perspect Clin Res*. 2017;8(4):187-191.

doi: 10.4103/picr.PICR_123_17

34. Koo TK, Li MY. A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163.

doi: 10.1016/j.jcm.2016.02.012

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Leveraging summary of radiology reports with transformers

Raul Salles de Padua[1]* and Imran Qureshi[2]*

[1]Quod Analytics, Niterói, Rio de Janeiro, Brazil

[2]Department of Computer Science, University of Texas Austin, Austin, Texas, United States of America

Abstract

Two fundamental problems in health-care stem from patient handoff and triage. Doctors are often required to perform complex findings summarization to facilitate efficient communication with specialists and decision-making on the urgency of each case. To address these challenges, we present a state-of-the-art radiology report summarization model utilizing adjusted bidirectional encoder representation from transformers BERT-to-BERT encoder–decoder architecture. Our approach includes a novel method for augmenting medical data and a comprehensive performance analysis. Our best-performing model achieved a recall-oriented understudy for gisting evaluation-L F1 score of 58.75/100, outperforming specialized checkpoints with more sophisticated attention mechanisms. We also provide a data processing pipeline for future models developed on the MIMIC-chest X-ray dataset. The model introduced in this paper demonstrates significantly improved capacity in radiology report summarization, highlighting the potential for ensuring better clinical workflows and enhanced patient care.

Keywords: Text summarization; Natural language processing; Deep learning; Artificial intelligence; Health care; Bidirectional encoder representations from transformers; MIMIC-chest X-ray

***Corresponding authors:**
Raul Salles de Padua
(raul.padua@iese.net)
Imran Qureshi
(imranq@utexas.edu)

Citation: de Padua RS, Qureshi I. Leveraging summary of radiology reports with transformers. *Artif Intell Health*. 2024;1(4):85-96. doi: 10.36922/aih.3846

Received: June 4, 2024

Accepted: August 5, 2024

Published Online: September 26, 2024

1. Introduction

Text summarization helps people devote attention to the most important parts of books, large bodies of text, and documents. In the process of radiology reporting, doctors need to painstakingly summarize complex findings to facilitate communication with specialists, and the resulting technical reports, which are typically long, are largely obscure for most patients. The task of report summarization is an extremely crucial part of radiology reporting, but the complexities and challenges involved in reporting are further compounded by a shortage of practicing radiologists in the United States.[1]

Despite the advances attained, natural language processing (NLP) has rarely been applied to medical summarization tasks, particularly in radiology; therefore, very little is known about language models that are specifically trained for radiology summarization. Radiology reports are typically lengthy and filled with technical jargons, making them difficult for non-specialists to interpret. Patients, in particular, may find it challenging to understand their medical conditions and treatment plans after reading these detailed

reports. Therefore, summarizing these reports not only aids in the clinical workflow but also enhances patients' understanding of their medical conditions and improves their engagement with health-care personnels.

In addition, although there is an evident need for effective summarization tools in radiology, advancements in NLP have been slow to penetrate this domain. The application of state-of-the-art NLP models to radiology report summarization remains relatively unexplored. Addressing this gap presents a significant opportunity for innovation and improvement in medical communication and patient care.

The present study addresses this gap by developing and evaluating a novel summarization model tailored specifically for radiology reporting. By utilizing the MIMIC – chest X-ray (CXR) dataset,[3] we fine-tuned a bidirectional encoder representation from transformers BERT-based model[2] to create Biomedical-BERT2BERT, achieving state-of-the-art performance in generating concise and accurate summaries of radiology findings. Our approach not only leverages advanced NLP techniques but also introduces a novel data augmentation strategy to enhance model performance. This model takes multiple free-text radiology report fields as input and uses a sequence-to-sequence architecture to output abstract summaries. The main contributions of our work are as follows:

(1) We developed Biomedical-BERT2BERT, an adjusted BERT-based model with state-of-the-art performance in radiology text summarization, assisting radiologists in generating concise impressions from reports
(2) We introduced a novel data augmentation strategy to improve performance on related tasks with MIMIC-CXR reports
(3) We conducted an in-depth analysis of Biomedical-BERT2BERT's performance, knowledge gain, and limitations with respect to disease distribution and other architectures.

Our data processing pipeline and model training code are also provided.[4]

2. Related works

Initial work in this domain was conducted by Chen *et al.*,[5] who reported promising results in predicting radiologist impressions from raw findings using fine-tuned BERT-based encoder–decoder models. We extended this work by experimenting with different architectures, understanding the limitations of applying language models in this domain, and investigating the effectiveness of modern linear attention mechanisms on MIMIC-CXR.

Chen *et al.*[5] addressed the challenge of abstractive summarization in radiology reporting using pre-trained BERT-based models. They developed ClinicalBioBERTSum, which incorporates domain-specific BERT models into the BERTSum architecture. This model was applied to the MIMIC-CXR dataset, focusing on predicting the "Impression" section of radiology reports based on the "Indication" and "Findings" sections. Their model achieved a recall-oriented understudy for gisting evaluation (ROUGE)-L F1 score of 57.37 and introduced ClinicalBioBERTScore to better evaluate the semantic quality of the summaries. Their work emphasizes the importance of domain-specific pre-training and fine-tuning in improving summarization performance for clinical texts. By leveraging ClinicalBioBERT, a model fine-tuned on MIMIC-III clinical texts, they improved the representation of medical semantics in radiology reports. The study also explored the use of custom tokenizers and a two-stage fine-tuning process, which showed that combining extractive and abstractive summarization objectives could enhance model performance.

Devlin *et al.*[2] introduced BERT, a novel transformer-based model pre-trained on large text corpora for various NLP tasks. BERT's bidirectional training allows it to capture context from both left and right surroundings, significantly enhancing its performance on tasks such as question answering, language inference, and more. The model sets new benchmarks on several NLP tasks, demonstrating the effectiveness of pre-training on large datasets followed by fine-tuning on specific tasks. The study's impact on the field is profound, as BERT has become a foundational model for many subsequent NLP advancements. Its architecture, which utilizes multiple layers of transformers and a masked language model objective, enables it to learn deep contextual representations. This work paved the way for numerous domain-specific adaptations, such as BioBERT and ClinicalBERT, which tailor the model to specific fields by further pre-training on domain-specific texts.

Alsentzer *et al.*[6] explored the application of contextual word embedding models, such as ELMo and BERT, to clinical text, addressing the gap in publicly available pre-trained models for clinical NLP tasks. They developed and released BERT models specifically trained on clinical text, demonstrating significant improvements over traditional BERT and BioBERT models on several clinical NLP tasks. These tasks included named entity recognition (NER) and medical natural language inference, where their domain-specific models outperformed general domain models. The study highlights the challenges of applying general pre-trained models to domain-specific texts due to linguistic differences. By training BERT models on the MIMIC-III dataset, which comprises approximately two million clinical notes, they tailored the embeddings to better fit the clinical context. This work is notable for providing publicly

accessible resources that can be utilized by the wider research community to advance clinical NLP applications. The models showed robust performance across various tasks, although they noted limitations in de-identification tasks due to differences in data characteristics between training and task datasets.

The T5 (Text-To-Text Transfer Transformer) model, created by Raffel *et al.*,[8] frames all NLP tasks as text-to-text problems. This approach allows for a unified framework where both inputs and outputs are treated as text strings, simplifying the architecture and training process. T5 is pre-trained on a large dataset (C4) and fine-tuned on various downstream tasks, achieving state-of-the-art results across a wide range of benchmarks. The study highlights the versatility and efficiency of the text-to-text framework, demonstrating its applicability to tasks such as translation, summarization, and question–answering. Using a consistent model structure for different tasks, T5 reduces the complexity of developing task-specific models. The success of T5 underscores the potential of transfer learning and model unification in advancing the capabilities of NLP systems.

Li *et al.*[9] investigated the adaptation of long-sequence transformer models, such as Longformer and BigBird, to clinical NLP tasks. These models address the limitations of traditional transformers such as BERT, which are constrained by a maximum input sequence length of 512 tokens. By employing sparse attention mechanisms, Clinical-Longformer and Clinical-BigBird can handle sequences up to 4096 tokens, making them suitable for the lengthy documents common in clinical contexts. Their study involved pre-training these models on large-scale clinical corpora and evaluating them on a variety of NLP tasks, including NER, question answering, and document classification. The results demonstrated that both Clinical-Longformer and Clinical-BigBird significantly outperformed ClinicalBERT and other short-sequence transformers across all tasks. This work underscores the potential of long-sequence models to improve the processing and analysis of extensive clinical texts, paving the way for more effective NLP tools in health care.

The application of transformer-based models, such as BERT, GPT-3, and T5, in medical text summarization has been explored by Yalunin *et al.*[10] They found that fine-tuning these models on medical datasets significantly improves their performance. Compared to their findings, our Biomedical-BERT2BERT model demonstrates superior performance due to our novel data augmentation techniques. Kraljevic *et al.*[11] proposed a multimodal approach combining text and image data for summarizing medical documents. While their method shows promise, our current work focuses on text-only data. Future work could explore incorporating multimodal data to enhance our model further.

Separately, a comprehensive review by Zhang *et al.*[12] recent advancements in NLP for medical text processing highlights the latest trends and future directions, contextualizing our work within the broader landscape of NLP advancements in medical text processing. Our contributions align with and extend these current trends, offering novel solutions for radiology report summarization.

2.1. Tokenizers

BERT-based models have been trained on word-splits tokenizers on several corpora, mainly wiki-data and literature datasets in the process usually called tokenization. Tokenization is breaking the raw text into small chunks. Tokenization breaks the raw text into words and sentences called tokens. These tokens help in understanding the context or developing the model for the NLP. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of the words.

2.2. Pre-trained language models (PLMs)

PLMs are large neural networks that are used in a wide variety of NLP tasks. They operate under a pre-train-finetune paradigm: Models are first pre-trained over a large text corpus and then fine-tuned on a downstream task using additional datasets. Most common architectures, such as BERT[6] and T5,[8] have not been pre-trained on specialized medical corpora. We have fine-tuned our model in the MIMIC-CXR dataset, which is a large publicly available dataset of chest radiographs, free-text radiology reports, and structured labels.

2.3. Evaluation metrics

We evaluated summarization generation performance with a recall-oriented understudy for gisting evaluation, or ROUGE[7] on F1 metrics. Historically, ROUGE has shown a good correlation with human-evaluated summaries and is a canonical metric for summarization evaluation. We focused on a variant of ROUGE, called ROUGE-L which measures the longest common subsequence (LCS) overlap between the predicted and reference summaries to evaluate the informativeness of the summary.

3. Approaches

3.1. Text summarization

Our task, text summarization for biomedical documents, can be approached by either extractive or abstractive methods. Extractive summaries are snippets taken directly

from the source text that best preserve the salient aspects of the document. The abstractive method, on the other hand, may generate text which may not be included in the source text.

Biomedical summaries require an abstractive approach. Radiology settings in particular require the interpretation of many data points to identify the most important aspects of a patient's condition and implications for future care. For instance, a radiologist's report may include only physical details of lung nodules, but the summary may conclude that "pneumonia is or is not present."

3.2. Baselines

We began by leveraging several pre-trained models from HuggingFace and built our own data processing pipeline[4] to extract sections and identify relevant training data from the MIMIC-CXR dataset. We baselined our project by fine-tuning a T5 encoder–decoder[8] and Meta's BART[13] in the MIMIC-CXR dataset, implementing Teacher-Forcing[14] on our training batches with HuggingFace's DataCollatorForSeq2Seq.[15] This baseline already performed near SOTA at 47.55 ROUGE-L.

3.3. Main approach

To improve on these results, we experimented with approaches in model architecture including:

- Larger variants of common architectures
- Custom encoder–decoder models
- Specialized checkpoints in medical data
- Models using linear attention mechanisms.

We then moved on to a data-centric approach by shuffling all input fields for each of our highest-performing model architectures: T5, BERT2BERT, and a BigBird-PubMed-Base model,[16] with the latter model chosen because it relies on block sparse attention instead of normal attention and can handle longer sequences. This data-centric approach proved to be key to reaching a new state-of-art performance, improving the previous SOTA work[5] ROUGE-L performance by 1.38 points.

We leveraged several pre-trained models from HuggingFace and built our own data processing pipeline to extract sections and identify relevant training data from the MIMIC-CXR dataset. We have baselined our project by fine-tuning a T5 encoder–decoder[8] and Facebook' BART[13] in the MIMIC-CXR dataset.

To perform the summarization task, the input fields were fed with sentence-level embeddings created by the encoder to a transformer decoder initialized randomly.

Encoders and decoders were fine-tuned end-to-end; Figure 1 for the whole process.

We used cross-entropy loss[17] with a weighted average of distances to a reference word in the vector space, outlined in Equation I. Y_i is the i-th word in the predicted summary and V_k is the k-th word in the reference vocabulary. $E(w)$ is a vector of the word w, V_k, and y_i, in the equation, and ed is the Euclidean distance computation.

$$Loss = \sum_{i=0}^{I} \sum_{k=0}^{K} p\left(y_{<i}, X\right) ed(E\left(V_k\right), E(y_i) \quad \text{(I)}$$

Table 1 provides an overview of the model performance.

3.3.1. Data pipeline

We built a comprehensive data processing pipeline to handle the MIMIC-CXR dataset. The pipeline includes the following steps: (1) Data extraction when free-text radiology reports are extracted with associated labels from the MIMIC-CXR dataset. (2) Preprocessing that tokenizes the text and removes any unnecessary characters or noise. (3) Section extraction from reports, in which relevant sections such as FINDINGS, IMPRESSION, INDICATION, and TECHNIQUE are extracted from the radiology reports. (4) Data augmentation, where data augmentation techniques are applied to create new training examples, improving model robustness and generalization (detailed in the following subsection). (5) Model training, which uses the processed and augmented data to train the Biomedical-BERT2BERT model.

3.3.2. Data augmentation techniques

Significant improvements in our model are attributed to the data augmentation techniques utilized. Data augmentation in NLP involves creating new training examples by altering the existing data to improve model robustness and generalization.

We implemented input's fields shuffling. For each model, we trained on different epochs with shuffled input fields as new examples (*i.e.*, FINDINGS, IMPRESSION, INDICATION, and TECHNIQUE; Figure 2). This technique ensures that the model learns to understand the context and meaning of the text regardless of the order in which the sentences are presented, thereby improving its ability to generalize across different sentence structures.

We acknowledge that class imbalance in the dataset has affected our model's performance, particularly in the "No Findings" category. Initially, we did not address this imbalance to observe the model's natural learning patterns. However, to improve the model, we implemented the fields shuffling augmentation technique to mitigate underrepresented categories.

3.3.3. Rationale for model configurations

The rationale behind choosing specific model configurations is based on balancing model complexity

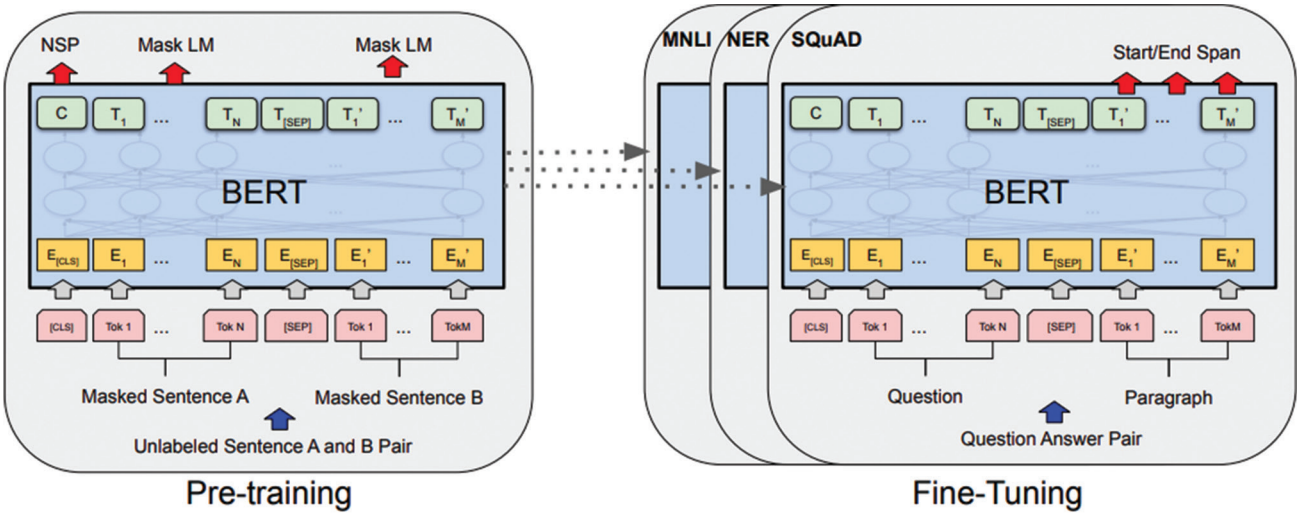


Figure 1. Pre-training and fine-tuning procedures of bidirectional encoder representations from transformers.[6] The same pre-trained model parameters are used to initialize models for different downstream tasks

Input Data

COMPARISON:Chest radiograph ___
FINDINGS: Scoliosis of the thoracic spine and consequent asymmetry in the rib spaces. The compression fracture in the thoracic vertebral body is stable. Normal size of the cardiac silhouette. Normal hilar and mediastinal structures, no pulmonary edema. No pleural effusions. No pneumonia. IMPRESSION: Chronic scoliosis and stable compression fracture of a thoracic vertebra. Otherwise normal chest radiograph. No evidence pneumonia.
INDICATION: ___ year old man with two weeks of productive cough, diffuse expiratory low pitched lung sounds on exam. // r/o pneumonia,
TECHNIQUE: Chest PA and lateral

Labels

IMPRESSION: No evidence of pneumonia.

Figure 2. An example of MIMIC-CXR[3] radiologist report with all input fields. Blank fields represent censored patient demographic information

Table 1. Training time of the models

#	Model name	Training time per epoch
1	Base-T5-Small (3 epochs)	0 h
2	Fine-tuned T5-Small (3 epochs)	1.5 h
3	DistillBART (3 epochs)	2.5 h
4	Base-T5-long (12 epochs)	0.63 h
5	Fine-tuned BERT2BERT (6 epochs)	0.65 h
6	Fine-tuned PubMed BigBird (7 epochs)	1.55 h
7	ClinicalLongFormer2ClinicalLongFormer (4 epochs)	1.6 h
8	ClinicalBioBert2Transformer (previous SOTA)	-

Abbreviation: BERT: Bidirectional encoder representations from transformers.

and performance. Larger models with more parameters typically perform better due to their capacity to learn complex patterns. However, they also require more computational resources and are prone to overfitting. Custom encoder–decoder models allow for flexibility in a design tailored specifically for our summarization task. Utilizing specialized checkpoints like ClinicalBERT leverages domain-specific pre-training, potentially improving performance on medical texts. Finally, models with linear attention mechanisms such as BigBird are chosen to handle longer input sequences efficiently, which is crucial for summarizing lengthy radiology reports.

4. Experiments

4.1. Data

The dataset used in this work is the MIMIC-CXR dataset, a collection of 377,110 CXR images and 227,827 associated free-text radiology reports and structured labels.[3] The dataset is intended to support a wide body of research in medicine including image understanding, NLP, and decision support.

We focused on the free-text reports, each comprising sections such as the radiologists' image observations, history, comparisons between images, and final impressions. For the purposes of our baselines, we sampled approximately 97,000 reports that included both FINDINGS and IMPRESSION sections and pre-processed them to extract sections and prepare models for generating an IMPRESSION from FINDINGS.

To improve on these results, we expanded our input fields to include FINDINGS, INDICATION, TECHNIQUE, and COMPARISON. An example layout is illustrated in Figure 2. In the final phase of the project, we designed a text augmentation technique by training different epochs with input fields shuffled ordering.

4.2. Evaluation method

In text summarization tasks, various evaluation metrics were used to assess the quality of the generated summaries. Common metrics include BLEU, METEOR, and ROUGE.[18] BLEU measures the correspondence between machine-generated text and human-written reference text using n-gram overlaps, making it a popular choice for evaluating machine translation. However, BLEU has limitations in text summarization as it primarily focuses on precision and does not adequately capture recall, which is crucial for summarization tasks. METEOR addresses some of BLEU's shortcomings by considering synonyms and stemming, thereby providing a more nuanced evaluation. However, METEOR is still more suited for translation tasks than for summarization.

For our study, we chose ROUGE, specifically ROUGE-L, as the primary evaluation metric. ROUGE-L focuses on the LCS between the predicted and reference summaries, which is particularly effective in measuring the informativeness and fluency of the summaries. Unlike n-gram-based metrics, ROUGE-L captures the overall structure and coherence of the text by considering the longest matching sequence of words, making it well-suited for evaluating abstractive summarization tasks where the generated text may not have exact n-gram matches with the reference. This characteristic of ROUGE-L is crucial for radiology report summarization, where the goal is to produce coherent and informative summaries that may not directly match the source text verbatim.

By selecting ROUGE-L, we ensured that our evaluation metric aligned with the specific requirements of medical text summarization. The ability of ROUGE-L to balance precision and recall makes it an ideal choice for capturing the quality of summaries in terms of both completeness and relevance. Our study leveraged ROUGE-L to provide a comprehensive assessment of our model's performance, ensuring that the generated summaries effectively convey the critical information contained in radiology reports.

Equations II and III compute ROUGE precision and recall, respectively, where MaxLCS is the maximum length of LCS between the reference summary (R) and the candidate summary (C). r and c are the lengths of the reference and candidate summaries, respectively.

$$ROUGE_{precision} = \frac{Max_{LCS}(R,C)}{r} \quad \text{(II)}$$

$$ROUGE_{recall} = \frac{Max_{LCS}(R,C)}{c} \quad \text{(III)}$$

$$ROUGE_{F1} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad \text{(IV)}$$

Precision and recall scores are combined into an F1 score, as seen in Equation IV. We implemented ROUGE-L leveraging the base rouge_score calculator available from HuggingFace (huggingface.co/metrics/rouge). Our baseline models use cross-entropy loss on Teacher-Forcing masked spans while training. We implemented Teacher-Forcing on our training batches with HuggingFace DataCollatorForSeq2Seq. We also used BERTScore[16] in our model, to learn contextual embeddings for the reference and predicted summarization for the radiology reports and, thus, mitigate drawbacks of pure ROUGE score evaluation.

4.3. Experimental details

Our experiments established summarization from the fine-tuned BERT2BERT model and reproduced medical summarization work by Chen *et al.*[5]

We applied a set of hyperparameters, as illustrated in Table 2, to the entire dataset with a ratio of train/test split of 5:1 for developing the model. We also built a custom section extractor to automate the processing of whole reports into various components such as FINDINGS and IMPRESSION. These reports were filtered for those that included ALL INPUT FIELDS and IMPRESSION sections (Figure 2). The reports were then sampled, tokenized, and batched for training. Finally, we shuffled inputs for the last 2 – 3 epochs of every model trained.

4.4. Results

Our final set of results on ROUGE are shown in Table 3. We found that a fine-tuned T5 model performs best among our baseline models. We achieved state-of-the-art ROUGE-L F1 performance with a BERT2BERT model after 6 epochs, with each epoch using a different ordering of input fields.

5. Analysis and discussion

Our experiments yielded several non-intuitive results across model architecture, pre-training, and attention context. For instance, larger specialized models like ClinicalLongFormer[9] significantly underperformed baselines.

We investigated these results across disease types and analyzed why the vanilla BERT-to-BERT model trained directly on this task outperformed models that had specialized checkpoints on clinical data and used architectures with more sophisticated attention mechanisms.

5.1. Summarization across disease types

Alongside patient radiology reports, the MIMIC-CXR dataset provides extracted disease metadata that is either

Table 2. Hyperparameters used for top 3 best performing models trained

Hyperparameters	Bidirectional encoder representations from transformer's values	BigBird values	T5 values
Epochs	6	7	12
Batch size	8	8	2
Learning rate	1.0e-5	1.0e-5	1.0e-5
Gradient accumulation steps	2	2	4
Optimizer	AdamW	AdamW	AdamW
Training time per epoch	0.65 h	1.55 h	0.63 h

Table 3. ROUGE scores among different models we experimented with

#	ROUGE-1	ROUGE-2	ROUGE-L	Baseline/rank
1	0.67	0.15	0.63	Yes
2	48.71	37.98	47.42	Yes
3	30.81	19.51	26.88	Yes
4	55.68	**45.52**	**54.74**	**3rd**
5	**59.61**	**48.22**	**58.75**	1st
6	57.83	47.12	56.66	2nd
7	43.8	30.98	41.7	-
8	58.97	47.06	57.37	-

Note: Bold values highlight the best performance in the selected metrics. We found that a fine-tuned bidirectional encoder representation from transformers (BERT) 2BERT model performs the best.
Abbreviation: ROUGE: Recall-Oriented Understudy for Gisting Evaluation.

indicated or negated by the radiologist. For instance, the radiologist might note that the patient's chest "had no indications for pneumonia," which would be provided as "pneumonia: -1."

By comparing performance across these disease profiles (Figure 3), we observed the Biomedical-BERT2BERT model performance has a slight positive correlation with the number of examples per disease type. This indicates that while the model gains knowledge with more examples, there is potentially a model saturation point. One interpretation is that BERT-to-BERT has reached an architectural limit to improve the summarization of these complex disease types.

In contrast, the model performs almost twice as well on reports with "No Findings" (77.6 ROUGE- LSUM). "No Findings" reports are those where the radiologist still describes X-ray features, but interprets them to be normal and without disease. This summarization improvement is likely due to the following reasons:

- "No Findings" reports account for the majority of the examples
- Each report with no findings has a much smaller space of possible impressions compared to other disease-type impressions. Other diseases have a variety of nuances that are harder for the model to capture.

In fact, solely summarizing reports with the phrase "There is no cardiopulmonary process" could achieve a ROUGE-L of 53.2 on these reports. It is possible that the model has learned an efficient classification strategy to detect "No Findings" reports and respond with a few canonical phrases in those cases.

5.2. Specialized and general checkpoints

Interestingly, fine-tuned checkpoints on PubMed and other clinical data underperformed the base BERT2BERT model on this task. One example is the Clinical LongFormer model, which is pre-trained on large-scale clinical corpora[9] and achieves SOTA performance on many biomedical tasks. Similar performance was observed with BioClinical BERT.

However, radiology text summarization is a highly specialized task. One interpretation of this result is that other clinical checkpoints may only contain a fraction of the information required to summarize radiology notes effectively. As a result, these specialized checkpoints can easily fall into local minima with respect to the loss function, whereas a more general language checkpoint can optimize more for global minima.

For future studies, this evidence points to the importance of using a variety of pre-trained checkpoints, and not solely relying on fine-tuned variants for specialized tasks.

5.3. Limitations of linear attention mechanisms

Attention is the key mechanism underlying transformers. However, the time and memory complexity to calculate attention is scaled with $O(n^2)$, which restricts models such as BERT to a limited context size (*i.e.*, 512 tokens).

Many models such as Linformer, Reformer, and Perceiver[19] have been formulated to use linear attention methods by indirectly calculating "full attention" by approximation. Google's BigBird is the latest of such models,[16]

which uses random attention, windowed attention, and global attention to generate a sparse attention representation (Figure 4). The value of this approach is the ability to process 4096 tokens with sparse attention at approximately the same time complexity as with 512 tokens with full attention. Theoretically, this provides better information capture for longer documents. This is relevant for our task, as radiology reports can exceed the 512-token limit.

For BigBird, however, complete parity with full attention with n-tokens is only realized with n hidden attention layers.[16] This means at m < n layers, BigBird performance relies on the larger context size to have much more relevant information for the task than the 512 token limit. At m = n layers, we lose the performance advantage of linear attention as $O(n \nleftarrow m) = O(n^2)$.

By evaluating the information distribution in radiology text data, we found that the majority of IMPRESSION information can be derived from only two to three sections (*i.e.*, FINDINGS, COMPARISON, and INDICATION), whose size totaled 200 – 300 tokens, well within the BERT full attention limit. As a result, while BigBird might eventually achieve the Biomedical-BERT2BERT performance given more compute and scaling laws,[20] the larger context size effectively acted as statistical noise, rather than providing an information advantage. In contrast, since we provided key sections to BERT directly, the Biomedical-BERT2BERT model learned summarization more efficiently with full attention.

For future studies, the limited effectiveness of linear attention points to the importance of evaluating the information distribution within a dataset. Likely, the more concentrated relevant information is in a dataset, the less likely a larger context transformer will outperform.

5.4. Learning radiology from summarization

While transformers tend to find uninterpretable statistical patterns in the training data, we found that our model has learned a few radiology facts. A few notable observations that hint at some of the operating mechanisms for Biomedical-BERT2BERT are as follows:

- Pneumonia corresponds to pleural surfaces
- Negation for disease is entailed by phrasing normal physiology (*e.g.*, No pneumonia = Normal heart and lungs)
- "Chest" pertains to both heart and lung anatomical features.

Figure A1 provides more information in this regard. Visualizations were created by extracting cross-attention matrices between our BERT2BERT Encoder Decoder components and plotted with BERTViz.[21] We also sampled model outputs with a medical resident who found that the generated summaries encapsulate the source text well for a medical setting (Figure A2). This points to an exciting future direction to extract knowledge from radiology

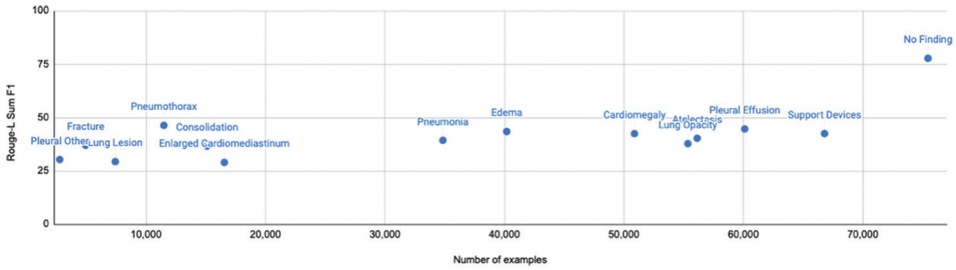


Figure 3. Performance distribution of ROUGE-L SUM scores versus the number of examples in the dataset. Image created with Google Sheets

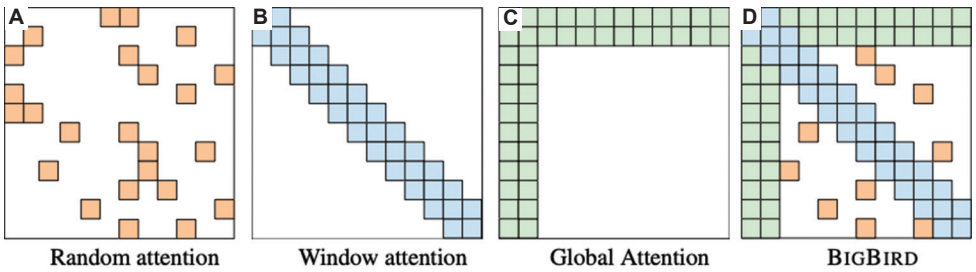


Figure 4. (A-D) Multiple attention mechanisms in the BigBird[16] linear attention calculation, which did not show improved performance for our summarization task

language models and provide interpretable information for users.

5.5. Practical implications and integration into clinical workflows

The practical implications of our findings are significant for clinical workflows. Our Biomedical-BERT2BERT model can be integrated into radiology departments to automate the summarization of radiology reports, thereby reducing the workload on radiologists and allowing them to focus on more critical tasks. This semi-automation can enhance the efficiency of patient handoffs and triage processes by providing clear and concise summaries of radiology findings.

Integrating our model into clinical workflows involves several steps. First, the model can be integrated into the existing radiology information systems platforms to automatically generate summaries as radiologists input their findings. Second, there is a need to develop a user interface that allows radiologists to review and edit the generated summaries before finalizing them. This ensures accuracy and allows radiologists to add and/or edit any additional context or information. Finally, it is crucial to provide training and support for radiologists and clinical staff on using the new system, along with ongoing support to address any issues or questions that arise during implementation.

By improving the efficiency and accuracy of radiology report summarization, our model has the potential to significantly impact patient care. Concise and clear summaries can help other healthcare providers quickly understand the radiologist's findings and make informed decisions about patient management. In addition, enhanced communication can lead to better patient outcomes by reducing the likelihood of misinterpretation and ensuring timely interventions.

6. Conclusion

In this work, we built a biomedical BERT2BERT text summarization model by performing fine-tuning with an end-to-end deep learning approach in a data-centric fashion. The input fields for the model are COMPARISON, FINDINGS, IMPRESSION, INDICATION, and TECHNIQUE fields, producing IMPRESSION predictions as outputs through abstractive summarization method. We believe that it will help to reduce human labor resources in medical settings. We used ROUGE scores as an evaluation metric to capture exact word-matching in a patient findings interpretation task that entails a high level of sensitivity. We believe that slight changes in this exact word matching with reference summaries may mislead patients and medical professionals. Our model generates state-of-the-art abstractive summarization by achieving a ROUGE-L score of 58.75/100.

After thorough experimentation, we found that a data-centric approach significantly improves the output quality of the radiology report summarization task. Our fine-tuned model may serve as a good checkpoints for other NLP endeavors in the medical space dealing with examination reports, such as summary predictions or even auto labeling.

Future works for this line of research include leveraging our data-centric approach combined with upsampling of the minority classes and downsampling of the majority class to create a more balanced training set and harnessing more data-centric approaches for improving the model's performance by addressing class imbalance with a higher proportion of "No Finding" impressions. In addition, due to computation constraints, we did not consider experimenting with large language models such as GPT-4 and Llama 3 in the current work. Although such models are decoder-only architectures, they learned during pre-training to understand the context deeply, a process that makes it perform tasks such as summarization and translation more effectively. Finally, incorporating human evaluation, more specifically from radiologists, may provide deeper insights into the quality of the summaries generated by the developed model. Another direction for future work is to conduct simple baseline measurements to gain a sense of how the model learns and then conduct a deep investigation to thoroughly understand the knowledge generated by the model.

Acknowledgments

The authors extend their appreciation to the Stanford University Computer Science Department for offering all the related conceptual guidance and structure in the cloud with computing resources as well as all the staff of CS224N – NLP with deep learning course.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: All authors
Investigation: All authors
Methodology: 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

MIMIC-CXR can be accessed at https://physionet.org/content/mimic-cxr/2.0.0/.

Further disclosure

Part of the findings have been presented in a final project event at the Stanford University campus in Palo Alto, California, in May 2022. The paper has been uploaded to a preprint server (doi: 10.48550/arXiv.2405.06802).

References

1. Takacs N, Makary MS. *Are we Prepared for a Looming Radiologist Shortage? Radiology Today*. Available from: https://www.radiologytoday.net/archive/rt0619p10.shtml [Last accessed on 2024 Jun 03].
2. Devlin J, Chang MW, Lee K, Toutanova K. BERT: *Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint* arXiv:1810.04805; 2019. Available from: https://aclanthology.org/N19-1423 [Last accessed on 2024 Jun 03].
3. Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *Nat Sci Data*. 2019;24(1):1-18.
4. De Padua RS, Qureshi I. *Colab Notebook with Fine-Tuned T5 Model for Radiology Summarization*. Available from: https://colab.research.google.com/drive/14A3j4bsTiC3hh3GdbLxwWGtwZoFiwciv [Last accessed on 2024 Jun 03].
5. Chen Z, Gong Z, Zhuk A. *Predicting Doctor's Impression for Radiology Reports with Abstractive Text Summarization. CS224N: Natural Language Processing with Deep Learning*. Stanford University; 2021. Available from: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/reports/final_reports/report005.pdf [Last accessed on 2024 Jun 03].
6. Alsentzer E, Murphy JR, Boag W, *et al*. *Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop (ClinicalNLP)*; 2019. p. 72-78. Available from: https://aclanthology.org/W19-1909 [Last accessed on 2024 Jun 03].
7. Lin CY. *ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out (2004): 74-81. Barcelona, Spain: Association for Computational Linguistics. Proceedings of the ACL Workshop: Text Summarization Braches Out*; 2004. p. 10.
8. Raffel C, Shazeer N, Roberts A, *et al*. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res*. 2020;21(140):1-67.
9. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *J Am Med Inform Assoc*. 2022;29(2):273-281.

   doi: 10.48550/arXiv.2201.11838
10. Yalunin A, Umerenkov D, Kokh V. Abstractive Summarization of Hospitalisation Histories with Transformer Networks.

    doi: 10.48550/arXiv.2204.02208
11. Kraljevic Z, Newham M, Fox D, *et al*. Multimodal representation learning for medical text summarization. *J Biomed Inform*. 2021;116:103713.
12. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. *BERTScore: Evaluating Text Generation with BERT. In: International Conference on Learning Representations (ICLR)*; 2020.

    doi: 10.48550/arXiv.1904.09675
13. Lewis M, Liu Y, Goyal N, *et al*. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; 2020. p. 7871-7880.

    doi: 10.48550/arXiv.1910.13461
14. Lamb AM, Goyal A, Zhang Y, Zhang S, Courville A, Bengio Y. Professor forcing: A new algorithm for training recurrent networks. *Adv Neural Inform Process Syst*. 2016;29:4601-4609.

    doi: 10.48550/arXiv.1610.09038
15. Wolf T, Debut L, Sanh V, *et al*. *Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*; 2020. p. 38-45.

    doi: 10.48550/arXiv.1910.03771
16. Zaheer M, Guruganesh G, Dubey A, *et al*. Big bird: Transformers for longer sequences. *Adv Neural Inform Process Syst*. 2020;33:17283-17297.

    doi: 10.48550/arXiv.2007.14062
17. Dahal P. *Classification and Loss Evaluation - Softmax and Cross Entropy Loss*. Available from: https://deepnotes.io/softmax-crossentropy [Last accessed on 2024 Jun 03].
18. Wolk K, Marasek K. Enhanced Bilingual evaluation understudy. arXiv preprint arXiv: 1509.09088; 2015.

    doi: 10.48550/arXiv.1509.09088
19. Tay Y, Dehghani M, Bahri D, Metzler D. Efficient

Transformers: A Survey. arXiv preprint arXiv:2009.06732; 2020.

doi: 10.48550/arXiv.2009.06732

20. Kaplan J, McCandlish S, Henighan T, *et al*. Scaling Laws for Neural Language Models. arXiv preprint arXiv:2001.08361; 2020.

doi: 10.48550/arXiv.2001.08361

21. Vig J. *A Multiscale Visualization of Attention in the Transformer Model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*; 2019. p. 37-42.

doi: 10.48550/arXiv.1906.05714

Appendix

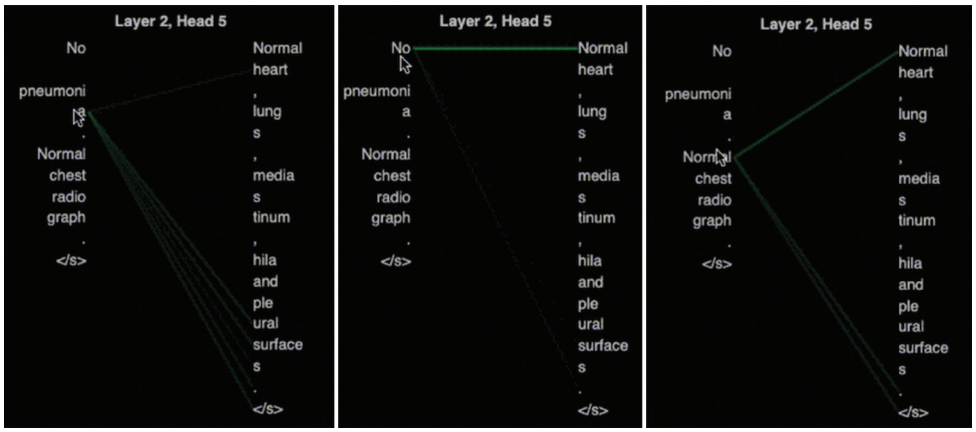


Figure A1. Visualization of BERT2BERT cross-attention weights using BERTViz[21]

Input Data

FINDINGS: There is mild cardiomegaly. Pulmonary markings are likely accentuated by lower lung volumes. There is no consolidation or pleural effusion. No pneumothorax. There are bilateral healed rib fractures and left clavicular healed rib fracture.

Ground Truth

IMPRESSION: No evidence of pneumonia.

Model Generated

IMPRESSION: Mild cardiomegaly. No evidence of pneumonia

Figure A2. An example of a radiologist's report with model outputs

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

An exploratory study on the potential of ChatGPT as an AI-assisted diagnostic tool for visceral leishmaniasis

Paulo Adriano Schwingel[1,2,3,4†*], Dino Schwingel[1,2†], Samuel Ricarte de Aquino[1,5†], Aline Rafaela Soares da Silva[1,2,3], Pedro Paulo Ramos da Silva[1,2], Renato Augusto da Cruz Pereira[1,2,6], Daniela Conceição Gomes Gonçalves e Silva[1,2,4], Amanda Alves Marcelino da Silva[1,2,3,4], Flavia Emília Cavalcante Valença Fernandes[1,2], Maria Jacqueline Silva Ribeiro[1,2,6], Paulo Ditarso Maciel Júnior[1,7], Paulo Gustavo Serafim de Carvalho[1,8], Ricardo Kenji Shiosaki[1,2], Rogério Fabiano Gonçalves[1,2], Bruno Bavaresco Gambassi[1,2,6], and Paula Andreatta Maduro[1,2,4]

[1]AI-assisted Diagnostics Research Group, Universidade de Pernambuco, Petrolina, Pernambuco, Brazil

[2]Human Performance Research Laboratory, Universidade de Pernambuco, Petrolina, Pernambuco, Brazil

Abstract

Visceral leishmaniasis (VL) is a severe parasitic disease that poses significant diagnostic challenges due to its complex presentation and the necessity for comprehensive diagnostic methods. This exploratory study investigates the potential of Chat Generative Pre-trained Transformer (ChatGPT)/GPT-4, an artificial intelligence (AI) chatbot, in assisting the diagnostic process for VL. We evaluated the diagnostic accuracy of ChatGPT/GPT-4 in generating differential diagnosis lists for eight clinical vignette cases of VL, authored by a Brazilian infectious disease doctor. Our findings reveal that ChatGPT/GPT-4 included VL in the top five differential diagnoses in 75% of the cases (95% confidence interval [CI]: 40.1 – 93.7%) and identified VL as the top diagnosis in 50% of the cases (95% CI: 30.3 – 86.5%). These results underscore the high potential of ChatGPT/GPT-4 as an AI-assisted diagnostic tool, which is capable of providing accurate differential diagnoses and assisting healthcare professionals in resource-limited settings. The study highlights the broader applicability of AI chatbots in medical diagnostics, not only for common conditions but also for specialized and less prevalent diseases like VL. By integrating AI tools into the diagnostic workflow, healthcare providers can enhance their diagnostic accuracy and efficiency, ultimately improving patient outcomes. This research contributes to the growing body of evidence supporting the utility of AI in healthcare and underscores the need for further studies to validate these findings across larger and more diverse clinical scenarios.

Keywords: Tropical neglected diseases; Artificial neural network; Differential diagnosis; Artificial intelligence-assisted diagnosis; Healthcare technology

†*These authors contributed equally to this work.*

***Corresponding author:**
Paulo Adriano Schwingel
(paulo.schwingel@upe.br)

Citation: 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

Received: June 13, 2024

Accepted: September 20, 2024

Published Online: October 16, 2024

[3]Postgraduate Program in Rehabilitation and Functional Performance, Universidade de Pernambuco, Petrolina, Pernambuco, Brazil
[4]Postgraduate Program in Health Sciences, Universidade de Pernambuco, Recife, Pernambuco, Brazil
[5]Dr. Washington Antônio de Barros Teaching Hospital, Brazilian Hospital Services Company, Petrolina, Pernambuco, Brazil
[6]Postgraduate Program on Management and Health Programs and Services, CEUMA University, São Luís, Maranhão, Brazil
[7]Postgraduate Program in Information Technology, Federal Institute of Paraíba, João Pessoa, Paraíba, Brazil
[8]College of Agricultural and Environmental Sciences, Federal University of Vale do São Francisco, Juazeiro, Bahia, Brazil

1. Introduction

Visceral leishmaniasis (VL) is a serious parasitic disease caused by protozoa of the genus *Leishmania*. The primary causative agents are *Leishmania donovani* in Africa and Asia; *Leishmania infantum* in Asia, Europe, and Africa; and *Leishmania chagasi* in the Americas.[1-3] These protozoans exhibit pleomorphic characteristics, with promastigote and paramastigote forms developing in the digestive tracts of insect vectors, while amastigote forms reside and multiply in the phagocytic cells of vertebrate hosts.[4] The life cycle of *Leishmania* involves binary fission in both hosts, highlighting the complex nature of the parasite transmission and infection process. Molecular studies have shown that *L. chagasi* and *L. infantum* are considered the same species, commonly known as *L. infantum* (syn. *chagasi*).[5]

The vectors responsible for the transmission of *Leishmania* show significant regional variation.[4-6] In Europe, Asia, and Africa, species of the genus *Phlebotomus* are the primary vectors, whereas species of the genus *Lutzomyia* predominate in the Americas.[1-3,6] This geographic differentiation in vector species underscores the importance of developing region-specific vector control strategies, as the efficacy of such interventions may vary significantly depending on the local vector ecology.

VL is a severe zoonotic disease that primarily affects impoverished regions and is endemic in approximately one hundred countries.[1-3] It has an alarming mortality rate in untreated cases, often due to complications such as multi-organ failure and secondary infections.[1-5] The disease is transmitted to humans primarily through biting by infected female sandflies, specifically *Lutzomyia longipalpis* in Brazil.[6] These vectors belong to the order *Diptera*, family *Psychodidae*, and subfamily *Phlebotominae*, and are characterized by their small size and pale-yellow coloration.[6-8]

The global burden of VL is significant, with millions of people at risk, particularly in tropical and subtropical regions.[9-11] The disease predominantly affects poor and marginalized communities, exacerbating existing health disparities and imposing a significant socioeconomic burden.[1-6] VL is also associated with alarmingly high rates of morbidity and mortality, particularly in endemic regions.[1-11] According to the World Health Organization Leishmaniasis Control Team,[9] the disease remains a leading cause of death, with mortality rates as high as 90% in untreated cases. This stark reality highlights the urgent need for timely diagnosis and effective treatment interventions to reduce the devastating impact of the disease on vulnerable populations.

VL can be diagnosed by immunologic and parasitological methods.[9] Immunologic diagnosis involves the detection of anti-Leishmania antibodies using techniques such as indirect immunofluorescence assay and rapid immunochromatographic tests, both of which are available through the Brazilian Unified Health System (SUS).[11-13] Parasitological diagnosis, which provides definitive evidence, involves the identification of amastigote forms of the parasite in biological samples, typically obtained from bone marrow due to its relative safety.[7,13,14] This process includes direct examination, culture isolation (*in vitro*), and isolation from susceptible animals (*in vivo*), as well as new diagnostic methods.[8,10,12]

Effective management of VL relies on both vector control strategies and targeted chemotherapeutic interventions. Vector control measures, such as insecticide spraying, the use of insecticide-treated bed nets, and environmental management, are critical in reducing sandfly populations and interrupting transmission.[6] On the chemotherapeutic front, the primary treatments include the use of antimonial compounds, Amphotericin B, and miltefosine.[5,7,8] Treatment regimens are tailored to individual patients, taking into account drug resistance patterns and the patient's overall health status to ensure optimal outcomes.[8]

Given the complexity of VL diagnosis, a comprehensive approach that includes anamnesis, palpation, biological specimens, serological testing, and biomarkers is essential.[12] Healthcare professionals must consider both endogenous and exogenous factors, including environmental conditions, in the prevention and control of VL.[10,12,14] Effective prevention relies on vector control, supported by community involvement in maintaining environmental hygiene.[9,12,14]

In the context of the evolving artificial intelligence (AI) landscape, tools such as large language models (LLMs) have been applied to a range of healthcare activities, including answering medical questions, examination, and diagnosis in hospitals.[15,16] In this sense, LLMs such as Chat Generative Pre-trained Transformer (ChatGPT) offer significant potential to aid in medical diagnosis.[15,17,18] ChatGPT, an AI chatbot running on a transformer architecture,[19,20] has shown promise in interpreting and integrating medical data to assist healthcare professionals.[15] While not a substitute for clinical judgment, AI tools can enhance the diagnostic process by providing differential diagnoses and refining clinical suspicions.[21,22]

In the dynamic field of AI, ChatGPT represents a significant advancement in improving human-machine communication.[18,23] By employing deep learning principles, ChatGPT leverages a comprehensive neural network model that is capable of understanding and generating text with nuanced context, tone, and intent.[19,20] The application of ChatGPT in medicine is growing, with research highlighting its utility in patient education, interaction, and health information dissemination.[15,18,22,24-31] In particular, ChatGPT has shown varying degrees of accuracy in diagnosing medical conditions, underscoring its potential role in the complex disease diagnostic processes.[17,21,22,26,32,33] However, despite its potential, research on the application of ChatGPT for the diagnostics of neglected tropical diseases remains limited, especially for the diagnosis of VL.

In this sense, this exploratory study evaluates the diagnostic accuracy of differential diagnosis lists generated by ChatGPT for clinical vignette cases of VL. Given the complexity of VL diagnosis, this research explores the integration of AI tools to assist healthcare professionals in making accurate and timely diagnoses to improve patient outcomes. In addition, this study aims to fill this gap by exploring how ChatGPT/GPT-4 can effectively integrate various medical data into the assisted diagnostic process of VL.

2. Data and methods

2.1. Study design

In this exploratory study, we evaluated the diagnostic accuracy of differential diagnosis lists generated by ChatGPT (GPT 4.0, https://chatgpt.com/, OpenAI OpCo, LLC, San Francisco, CA, United States of America [USA]) for clinical vignette cases of VL formulated in Brazilian.

Portuguese on March 31, 2024, the study was conducted at the Dr. Washington Antônio de Barros Teaching Hospital (HU-UNIVASF) of the Brazilian Hospital Services Company in Petrolina, Pernambuco, Brazil. Although the study involved the use of clinical vignette cases and thus did not require individual informed consent, approval was obtained from the Ethics Committee of the HU-UNIVASF (approval number 6.967.834) to ensure that the study was conducted according to the highest standards of research integrity.

2.2. Case materials

A Brazilian infectious disease physician, an expert in the diagnosis of VL and other neglected tropical diseases, formulated eight Brazilian clinical case studies. These studies were based on the common sociodemographic and clinical characteristics of outpatients diagnosed with VL at HU-UNIVASF.

The infectious disease specialist, a third investigator in this study (S.R.deA.), wrote the eight clinical vignette cases in Brazilian Portuguese. Each vignette included information on the patient's age, biological sex, place of birth, and other sociodemographic indicators identifying information, social history, history of present illness, past medical history, and physical examination.

It is important to emphasize that the infectious disease physician did not change the geographical scope of the study (São Francisco Valley, Brazil). Table 1 describes the eight clinical vignette cases used in the present study. The original texts, written by the physician in Brazilian Portuguese, were translated into English for presentation in this study.

2.3. Diagnosis lists generated by ChatGPT-4

On March 31, 2024, we utilized the ChatGPT (GPT 4.0, https://chatgpt.com/, OpenAIOpCo, LLC, San Francisco, CA, USA) for our research. ChatGPT-4 is an advanced natural language processing chatbot developed by OpenAI OpCo, LLC that builds on the success of previous models such as GPT-3. ChatGPT/GPT-4 is an LLM that has been trained on large amounts of text data, enabling it to generate human-like responses across multiple domains.[15,34,35] The findings from the Global Burden of Disease Study have been integrated into the ChatGPT/GPT-4 platform to enhance personalized healthcare planning through the use of AI-assisted disease burden assessment and planning tools.[32]

The differential diagnoses generated by ChatGPT/GPT-4 are derived from its extensive training on a wide array of medical literature, including significant studies such as the Global Burden of Disease study.[36] It must also be noted that in this study, ChatGPT/GPT-4 did not access external sources in real time;[34] instead, it produced responses solely based on the knowledge acquired during its training phase.[15]

Table 1. Clinical description of the eight viral leishmaniasis vignettes

Case	Description
01	ID: Female, 52 years old, Caucasian, single, lives in Petrolina, Pernambuco (urban area), homemaker. HPI: The patient presents with a 60-day history of daily fever, associated with asthenia, weight loss, diarrhea, abdominal discomfort, anorexia, and chills. PMH: HIV-positive since 2007, diagnosed with systemic lupus erythematosus in 2019, treated for visceral leishmaniasis 6 times; current medications: dolutegravir, darunavir, and ritonavir. SH: Denies alcohol use and smoking. PE: Weight: 56 kg; height: 1.56 mm; BP: 110/70 mmHg; abdomen: hepatomegaly 4 cm below the costal margin, splenomegaly 5 cm below the costal margin, no ascites, abdominal tenderness on palpation; cardiovascular: regular cardiac rhythm, heart rate 72 bpm; respiratory: clear breath sounds, no wheezes or crackles; skin/mucous membranes: no jaundice, pale, no mucosal lesions.
02	ID: Male, 67 years old, African American, married, lives in Petrolina, Pernambuco (rural area), farmer. HPI: The patient presents with a 30-day history of daily fever, associated with asthenia, weight loss, diarrhea, and nausea. PMH: Hypertension, hypercholesterolemia, glaucoma, nephrectomy 9 years ago; current medications: losartan, simvastatin. SH: Denies alcohol use and smoking. PE: Weight: 60 kg; height: 1.53 m; BP: 130/80 mmHg; abdomen: normal liver span, splenomegaly 5 cm below the costal margin, no ascites, no abdominal tenderness on palpation; cardiovascular: regular cardiac rhythm, heart rate 78 bpm; respiratory: clear breath sounds, no wheezes or crackles; skin/mucous membranes: no jaundice, well-perfused, no mucosal lesions.
03	ID: Male, 36 years old, Caucasian, single, lives in Petrolina, Pernambuco (rural area), painter. HPI: The patient presents with a 60-day history of daily fever, associated with asthenia, weight loss, and diarrhea. PMH: HIV-positive; current medications: dolutegravir, tenofovir, lamivudine. SH: Moderate alcohol use, smokes cigarettes, and marijuana user. PE: Weight: 51 kg; height: 1.65 m; BP: 110/70 mmHg; abdomen: hepatomegaly 10 cm below the costal margin, splenomegaly 15 cm below the costal margin, no ascites, no abdominal tenderness on palpation; cardiovascular: regular cardiac rhythm, heart rate 98 bpm; respiratory: clear breath sounds, no wheezes or crackles; skin/mucous membranes: no jaundice, pale, and no mucosal lesions.
04	ID: Female, 51 years old, African American, single, lives in Juazeiro, Bahia (urban area), homemaker. HPI: The patient presents with a 60-day history of daily fever, associated with asthenia, weight loss, diarrhea, abdominal discomfort, anorexia, chills, and bleeding. PMH: Liver cirrhosis, hypertension, diabetes, chronic kidney disease; current medications: unable to specify. SH: Denies alcohol use and smoking. PE: Weight: 81 kg; height: 1.57 m; BP: 100/80 mmHg; abdomen: normal liver span, splenomegaly 5 cm below the costal margin, ascites, abdominal tenderness on palpation; cardiovascular: regular cardiac rhythm, heart rate 78 bpm; respiratory: clear breath sounds, no wheezes or crackles; skin/mucous membranes: mild jaundice, pale, and no mucosal lesions.
05	ID: Female, 39 years old, mixed race, single, lives in Sento Sé, Bahia (urban area), homemaker. HPI: The patient presents with a 65-day history of daily fever, associated with asthenia, weight loss, chills, and bleeding. PMH: Denies any previous diseases; current medications: none. SH: denies alcohol use and smoking. PE: Weight: 63 kg; height: 1.65 m; BP: 120/70 mmHg; abdomen: Normal liver span, splenomegaly 4 cm below the costal margin, no ascites, no abdominal tenderness on palpation; cardiovascular: regular cardiac rhythm, heart rate 80 bpm; respiratory: clear breath sounds, no wheezes or crackles; skin/mucous membranes: no jaundice, pale, no mucosal lesions.
06	ID: Male, 50 years old, Caucasian, married, lives in Cabrobó, Pernambuco (urban area), farmer. HPI: The patient presents with a 90-day history of daily fever, associated with asthenia, weight loss, abdominal discomfort, and bleeding. PMH: Denies any previous diseases; current medications: None. SH: Light alcohol use, smoker for 15 years, 10 cigarettes per day. PE: Weight: 60 kg; height: 1.66 m; BP: 130/80 mmHg; abdomen: hepatomegaly 12 cm below the costal margin, splenomegaly 8 cm below the costal margin, no ascites, no abdominal tenderness on palpation; cardiovascular: regular cardiac rhythm, heart rate 82 bpm; respiratory: clear breath sounds, no wheezes or crackles; skin/mucous membranes: no jaundice, pale, no mucosal lesions.
07	ID: Male, 49 years old, Caucasian, married, lives in Petrolina, Pernambuco (rural area), farmer. HPI: The patient presents with a 35-day history of daily fever, associated with weight loss, cough, and sweating. PMH: Denies any previous diseases; current medications: None. SH: Moderate alcohol use, smoker. PE: Weight: 67 kg; height: 1.75 m; BP: 100/70 mmHg; abdomen: hepatomegaly 6 cm below the costal margin, splenomegaly 7 cm below the costal margin, no ascites, abdominal tenderness on palpation; cardiovascular: regular cardiac rhythm, heart rate 100 bpm; respiratory: clear breath sounds, no wheezes, presence of crackles; skin/mucous membranes: no jaundice, pale, and no mucosal lesions.
08	ID: Male, 42 years old, mixed race, single, lives in Petrolina, Pernambuco (rural area); occupation not reported. HPI: The patient presents with asthenia, weight loss, and abdominal discomfort. PMH: Previous visceral leishmaniasis; current medications: None. SH: Light alcohol use, denies smoking.

(Cont'd...)

Table 1. ***(Continued)***

Case	Description
	PE: Weight: 69 kg; height: 1.56 m; BP: 100/60 mmHg; abdomen: hepatomegaly 3 cm below the costal margin, splenomegaly 5 cm below the costal margin, no ascites, abdominal tenderness on palpation; cardiovascular: regular cardiac rhythm, heart rate 76 bpm; respiratory: clear breath sounds, no wheezes or crackles; skin/mucous membranes: no jaundice, pale, no mucosal lesions.

Abbreviations: BP: Blood pressure; HPI: History of present illness; ID: Identifying information; PE: Physical examination; PMH: Past medical history; SH: Social history.

The second investigator in this study (D.S.) employed a similar methodology to that performed by Hirosawa *et al.*[21] by typing the following text into the ChatGPT (GPT 4.0, OpenAI OpCo, LLC) prompt in Brazilian Portuguese: "Please provide me with the five most likely diagnoses for the following symptoms: (copy and paste each clinical vignette)." The order of the clinical vignettes presented to ChatGPT/GPT-4 was randomized using a computer-generated order table (Case 02, 08, 04, 01, 06, 05, 03, and 07). To ensure the integrity of the data and to avoid any influence of previous interactions, each clinical vignette was presented to ChatGPT/GPT-4 only once in a new chat session. This approach was employed to prevent any potential influence of previous interactions on the AI's responses.[21]

2.4. Measurements and definitions

The accuracy of the VL diagnosis was evaluated based on the inclusion of the correct diagnosis within the top five differential diagnoses generated by ChatGPT (GPT 4.0, OpenAI OpCo, LLC). This approach employed a binary scoring system, whereby the presence of a diagnosis in the list was scored as one, and its absence was scored as zero. Furthermore, the position of the VL diagnosis within the lists, classified between first and fifth, was analyzed sequentially.

2.5. Statistical analysis

The responses were entered into regular Excel spreadsheets (Microsoft Corporation, Redmond, WA, USA, Release 12.0.6662, 2012) and exported to the Statistical Package for the Social Science for Windows (SPSS Inc., Chicago, Illinois, USA, Release 16.0.2, 2008) for statistical analysis. Descriptive statistical analysis was performed on categorical variables, which were presented as absolute and relative frequencies. The accuracy of ChatGPT/GPT-4 as an AI-assisted diagnostic tool for VL was calculated using the prevalence ratio, and its inaccuracy was estimated using a 95% confidence interval (95% CI). Statistical analyses were conducted in a two-tailed manner, and statistical significance was set at $P < 0.05$.

3. Results

The correct diagnosis of VL among the five differential diagnoses generated by ChatGPT (GPT 4.0, OpenAI OpCo, LLC) was presented 6 times, representing 75% of the total number of cases (95% CI: 40.1 – 93.7%). Table 2 shows the five differential diagnoses presented by ChatGPT/GPT-4 for each clinical vignette.

While ChatGPT/GPT-4 did not provide an accurate representation of VL as a diagnostic possibility for the clinical vignettes containing cases 03 and 04, it did report VL as the top diagnosis for four cases (50.0%; 95% CI9: 30.3 – 86.5%). Figure 1 shows the accuracy of ChatGPT/GPT-4 in presenting VL as a differential diagnosis (Figure 1A) and as the principal diagnosis (Figure 1B).

4. Discussion

The ability of ChatGPT to provide diagnostic support, especially in resource-limited settings where access to specialized medical expertise is limited, is one of its most promising contributions to healthcare. By providing reliable differential diagnoses, ChatGPT has the potential to bridge gaps in medical expertise, enabling more timely and accurate clinical decision-making in underserved areas.

This exploratory study evaluated the diagnostic accuracy of ChatGPT/GPT-4 in generating differential diagnosis lists for clinical vignettes of VL. The results showed that ChatGPT/GPT-4 correctly included VL in the top five differential diagnoses in 75% of cases. Notably, ChatGPT/GPT-4 identified VL as the top diagnosis in 50% of these cases. These results indicate that ChatGPT (GPT 4.0, OpenAI OpCo, LLC) has a high potential to aid in the diagnosis of VL, as evidenced by its significant accuracy in generating relevant differential diagnoses.

The findings of our study are consistent with a growing body of research demonstrating the diagnostic capabilities of AI chatbots. For example, Hirosawa *et al.*[21] evaluated the diagnostic accuracy of differential diagnosis lists generated by ChatGPT/GPT-3.5 on January 5, 2023, for clinical vignettes with common chief complaints. Their results showed that the correct diagnosis was included within the top ten differential diagnoses in 93.3% of cases. Similarly, a study by Mizuta *et al.*[22] showed that ChatGPT/GPT-4 had an elevated level of agreement (95.9%) with physicians in determining whether the correct diagnosis was included in the top ten differential diagnosis lists.

Table 2. List of the five differential diagnoses generated by ChatGPT-4 on March 31, 2024 for the visceral leishmaniasis clinical vignettes

Case	Diagnoses				
	1st	2nd	3rd	4th	5th
01	Visceral leishmaniasis	Opportunistic infections	Systemic lupus erythematosus	Tuberculosis	Gastrointestinal infection
02	HIV infection	Tuberculosis	Chagas disease	Visceral leishmaniasis	Leukemia or lymphoma
03	Opportunistic infections	Lymphoma	Tuberculosis	Viral hepatitis	Chagas disease
04	Endocarditis	Chronic liver disease	Tuberculosis	Lymphoma	HIV infection
05	Visceral leishmaniasis	Tuberculosis	Lymphoma	Chagas disease	Typhoid fever
06	Visceral leishmaniasis	Chagas disease	Schistosomiasis	Tuberculosis	Lymphoma
07	Tuberculosis	Visceral leishmaniasis	Viral hepatitis	Chagas disease	Lymphoma
08	Visceral leishmaniasis	Viral hepatitis	Chagas disease	Cirrhosis	Lymphoma

Abbreviation: HIV: Human immunodeficiency virus.

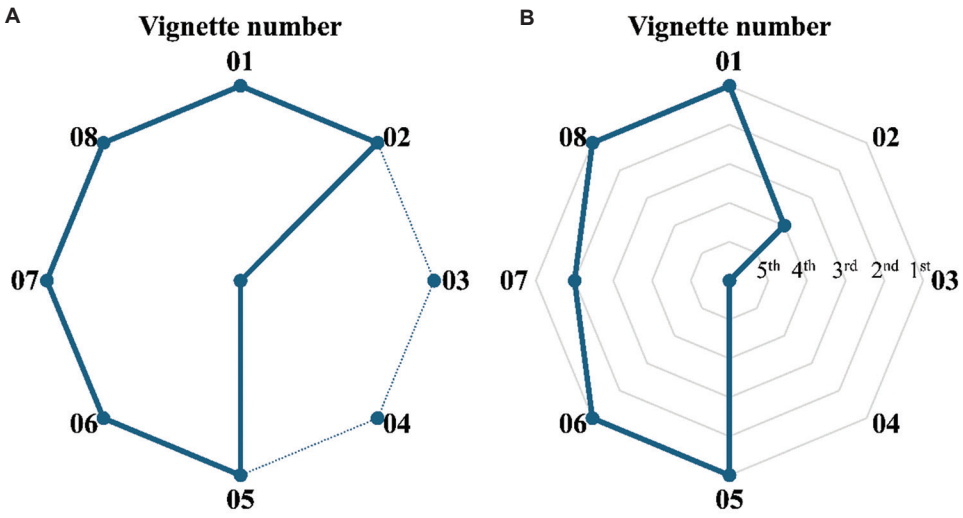


Figure 1. Rates of correct diagnoses of viral leishmaniasis (VL) generated by ChatGPT/GPT-4 for each clinical vignette. (A) Shows the generated results as Yes or No for the diagnosis of VL in the list of five differential diagnoses. (B) Shows the position of the generated diagnosis of VL between the first and fifth

In a different context, Biswas[30] explored the potential of ChatGPT in radiology and found that the AI could help generate accurate and consistent radiology reports. This corroborates our findings that AI chatbots can provide valuable diagnostic assistance in various medical fields. In addition, Cheng *et al.*[29] explored the potential of chat-GPT/GPT-4 in sports medicine, highlighting its role in diagnostic imaging and exercise prescription. This further underscores the broad applicability of AI in medical diagnostics.

The present study focuses on the diagnosis of VL, a tropical neglected disease that presents unique diagnostic challenges. The high accuracy rates observed indicate that AI chatbots, such as ChatGPT/GPT-4, are not only capable of handling common medical conditions but also excel in the diagnosis of more specialized and less common diseases. Other research conducted on AI-based chatbots also demonstrates their significant potential in specialized medical fields, including neurology,[28] infectious diseases,[27] and clinical scenarios.[26] Our findings and the literature suggest that AI chatbots have the potential to expand beyond the diagnosis of common medical conditions and into more niche areas of medicine. This suggests that they could be useful in a wide range of healthcare settings, thereby expanding the potential utility of AI chatbots beyond general practice.[15,24,25] They could also assist healthcare professionals in effectively diagnosing a wide range of diseases.[15,24-26]

The results of this study demonstrate the potential of ChatGPT/GPT-4 as a valuable AI-assisted diagnostic tool for the diagnosis of VL. By providing accurate differential diagnoses, ChatGPT/GPT-4 can assist physicians, especially in resource-limited settings where expert opinion and laboratory testing may not be readily available.

A notable strength of this study is the use of authentic clinical scenarios created by an infectious disease specialist with extensive experience in the diagnosis of VL. This approach ensures that the cases presented to ChatGPT/GPT-4 closely resemble real-world clinical scenarios. In addition, randomizing the order of case presentations and using a new chat session before entering each case helped to minimize potential biases in the AI's responses.

The potential for AI-assisted medical diagnosis to transform healthcare delivery is significant.[37,38] LLMs are capable of processing vast amounts of medical data with remarkable speed and precision, offering several advantages over traditional diagnostic methods.[15,34,35] One of the most significant advantages of AI in medical diagnosis is that it can provide diagnostic support in resource-limited settings where access to specialist medical knowledge is scarce.[21,22] AI can serve as a bridge to provide expert diagnostic suggestions, thereby improving patient outcomes and healthcare efficiency.[16,25,28,37,39] Furthermore, AI-based diagnostic tools can facilitate clinicians' decision-making processes. By generating comprehensive differential diagnosis lists, AI helps clinicians consider a wider range of potential conditions, thereby reducing the likelihood of misdiagnosis.[18,21,22,38] This is particularly important in cases where multiple conditions may present with similar symptoms.

Despite the encouraging results, it is important to note that our study is subject to certain limitations. First, the study employed a vignette-based methodology,[21,40] rather than involving real patient interactions, which may limit the generalizability of the findings. Second, the sample size was relatively limited, consisting of only eight clinical cases. The limited sample size of eight clinical cases limits the generalizability of the study's findings. To validate these findings, further studies with larger and more diverse samples are required to ensure the robustness of the conclusions. In addition, the selection of clinical vignettes reflecting common symptoms of VL may have contributed to an overestimation of the diagnostic capabilities of ChatGPT. Future studies should include a broader range of case presentations to evaluate the AI's performance in more varied clinical scenarios. Moreover, the binary scoring system used in this study, while simple, may not fully capture the nuances of differential diagnosis accuracy.

While AI in medical diagnosis offers significant benefits, several ethical issues must be addressed to ensure its responsible use. A primary concern is the potential for AI algorithms to reflect existing biases in medical practice and societal inequalities, which could lead to unequal treatment.[41,42] To prevent this, it is essential that the datasets used to train AI are representative and that algorithms are regularly audited to identify any potential bias. It is also critical to ensure the transparency and accountability of AI-driven diagnoses.[43] Clinicians and patients must be able to understand how the AI arrived at its conclusions to trust and effectively use these tools. It is critical to develop AI systems that provide clear and interpretable reasoning, as this is essential for informed decision-making.

Another crucial issue in the field of AI-assisted medical diagnosis is privacy and security, as it involves the processing of sensitive patient information.[41,44,45] Robust measures must be implemented to protect data from breaches and misuse.[46] There is also a clear need to establish transparent policies and regulations for the use and sharing of data in AI applications.[46] Moreover, the potential for AI to replace human clinicians raises ethical questions about the future of the medical profession.[47] It must be highlighted that AI should be used to enhance and reinforce clinical decision-making, instead of replacing the critical thinking, empathy, and nuanced understanding that human clinicians can provide. The role of AI in healthcare should be to assist and enhance the skills of healthcare professionals, ensuring that patient care remains human-centered.[47]

Finally, future research should focus on expanding the sample size and diversity of clinical cases to better understand the generalizability of ChatGPT's diagnostic capabilities. In addition, it would be beneficial to explore the integration of AI-generated diagnoses into clinical workflows and assess the impact on clinical decision-making and patient outcomes. Moreover, further studies should also consider the potential biases and ethical implications of using AI in healthcare. It is of the utmost importance that these tools are used responsibly and equitably.

5. Conclusion

This exploratory study demonstrates that ChatGPT/GPT-4 can generate an accurate differential diagnosis for VL, correctly identifying the disease in a considerable proportion of cases. Further research is necessary to confirm these findings. This study also substantiates that ChatGPT/GPT-4 is a promising AI-assisted diagnostic tool with the potential to improve clinical decision-making and healthcare delivery.

Acknowledgments

None.

Funding

This study received financial support from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under grant number 408003/2023-5 and from the

Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE) under grant numbers APQ-1413-4.08/21 and APQ-0238-4.01/24. Additionally, Paulo Adriano Schwingel was awarded a Research Productivity Grant (BPP) from the FACEPE under number BPP-0003-4.01/24 and Daniela Conceição Gomes Gonçalves e Silva was awarded a Technical Cooperation Grant (BCT) from the FACEPE under number BCT-0355-4.08/23.

Conflict of interest

Paulo Adriano Schwingel is an editorial board member of this journal but was not in any way involved in the editorial and peer-review process conducted for this paper, directly or indirectly. Separately, other authors declared that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Author contributions

Conceptualization: Paulo Adriano Schwingel, Dino Schwingel, Samuel Ricarte de Aquino, Bruno Bavaresco Gambassi, Paula Andreatta Maduro

Formal analysis: Paulo Adriano Schwingel, Dino Schwingel, Samuel Ricarte de Aquino, Maria Jacqueline Silva Ribeiro, Paula Maduro

Investigation: Paulo Adriano Schwingel, Dino Schwingel, Samuel Ricarte de Aquino, Aline Rafaela Soares da Silva, Pedro Paulo Ramos da Silva, Daniela Conceição Gomes Gonçalves e Silva, Paulo Ditarso Maciel Júnior, Paulo Serafim de Carvalho, Paula Andreatta Maduro

Methodology: Paulo Adriano Schwingel, Dino Schwingel, Samuel Ricarte de Aquino, Amanda Alves Marcelino da Silva, Flavia Emília Cavalcante Valença Fernandes, Ricardo Kenji Shiosaki, Rogério Fabiano Gonçalves, Bruno Bavaresco Gambassi, Paula Andreatta Maduro

Writing–original draft: Paulo Adriano Schwingel, Aline Rafaela Soares da Silva, Pedro Paulo Ramos da Silva, Renato Augusto da Cruz Pereira, Daniela Conceição Gomes Gonçalves e Silva, Bruno Bavaresco Gambassi, Paula Andreatta Maduro

Writing–review & editing: Dino Schwingel, Samuel Ricarte de Aquino, Amanda Alves Marcelino da Silva, Flavia Emília Cavalcante Valença Fernandes, Maria Jacqueline Silva Ribeiro, Paulo Ditarso Maciel Júnior, Paulo Gustavo Serafim de Carvalho, Ricardo Kenji Shiosaki, Rogério Fabiano Gonçalves

Ethics approval and consent to participate

This study involved the use of clinical vignette cases and thus did not require individual informed consent. Research approval was obtained from the Ethics Committee of the HU-UNIVASF (approval number 6.967.834).

Consent for publication

Not applicable.

Availability of data

The data used to support the findings of this study are included within the article.

References

1. Dawkins M, Lin Z, Cohen C, Mikkilineni S, Shakil F, Tewari V. A rare case of Visceral Leishmaniasis diagnosed by endoscopy in an anemic patient with HIV/AIDS. *ACG Case Rep J*. 2023;10(7):e01108.

   doi: 10.14309/crj.0000000000001108

2. Rodrigues Monteiro M, Serra JT, Gomes F, Tinoco J. Visceral Leishmaniasis in an immunocompetent patient: A case report. *Acta Med Port*. 2023;36:835-840.

   doi: 10.20344/amp.19010

3. Pandey K. Emerging association between serum vitamin D and degree of anemia in visceral leishmaniasis. *Biomed J Sci Tech Res*. 2023;49(2):40519-40521.

   doi: 10.26717/bjstr.2023.49.007781

4. Gonçalves C, Diniz B, Guerreiro B, *et al*. A case of visceral leishmaniasis in a child on platelet recovery after treatment with filgrastim. *Resid Pediatr*. 2023;13(1):1-4.

   doi: 10.25060/residpediatr-2023.v13n1-485

5. Silveira FT, Sousa Junior EC, Silvestre RV, *et al*. Comparative genomic analyses of new and old world viscerotropic leishmanine parasites: Further insights into the origins of visceral leishmaniasis agents. *Microorganisms*. 2022;11(1):25.

   doi: 10.3390/microorganisms11010025

6. Lainson R, Rangel BF. *Lutzomyia longipalpis* and the eco-epidemiology of American visceral leishmaniasis, with particular reference to Brazil: A review. *Mem Inst Oswaldo Cruz*. 2005;100(8):811-827.

   doi: 10.1590/s0074-02762005000800001

7. Ali N, Nakhasi HL, Valenzuela JG, Reis AB. Targeted immunology for prevention and cure of VL. *Front Immunol*. 2014;5:660.

   doi: 10.3389/fimmu.2014.00660

8. Singh OP, Sundar S. Immunotherapy and targeted therapies in treatment of visceral leishmaniasis: Current status and future prospects. *Front Immunol*. 2014;5:296.

   doi: 10.3389/fimmu.2014.00296

9. Alvar J, Vélez ID, Bern C, *et al*. Leishmaniasis worldwide and global estimates of its incidence. *PLoS One*. 2012;7(5):e35671.

   doi: 10.1371/journal.pone.0035671

10. Paul A, Singh S. Visceral leishmaniasis in the COVID-19

pandemic era. *Trans R Soc Trop Med Hyg*. 2023;117(2):67-71.

doi: 10.1093/trstmh/trac100

11. Rahim S, Sharif MM, Amin MR, Rahman MT, Karim MM. Real time PCR-based diagnosis of human visceral leishmaniasis using urine samples. *PLOS Glob Public Health*. 2022;2(12):e0000834.

doi: 10.1371/journal.pgph.0000834

12. Chappuis F, Sundar S, Hailu A, *et al*. Visceral leishmaniasis: What are the needs for diagnosis, treatment and control? *Nat Rev Microbiol*. 2007;5(11):873-882.

doi: 10.1038/nrmicro1748

13. Cavalcanti MP, Barros De Lorena VM, Gomes YD. Biotechnological advances for the diagnosis of infectious and parasitic diseases. *J Trop Pathol.* 2008;37(1):1-14.

doi: 10.5216/rpt.v37i1.4026

14. Makau-Barasa LK, Ochol D, Yotebieng KA, Adera CB, De Souza DK. Moving from control to elimination of Visceral Leishmaniasis in East Africa. *Front Trop Dis*. 2022;3:965609.

doi: 10.3389/fitd.2022.965609

15. Mumtaz U, Ahmed A, Mumtaz S. LLMs-healthcare: Current applications and challenges of large language models in various medical specialties. *Artif Intell Health*. 2024;1(2):16-28.

doi: 10.36922/aih.2558

16. Sukeda I, Suzuki M, Sakaji H, Kodera S. Development and analysis of medical instruction-tuning for Japanese large language models. *Artif Intell Health*. 2024;1(2):107-116.

doi: 10.36922/aih.2695

17. Meral G, Ateş S, Günay S, Öztürk A, Kuşdoğan M. Comparative analysis of ChatGPT, Gemini and emergency medicine specialist in ESI triage assessment. *Am J Emerg Med*. 2024;81:146-150.

doi: 10.1016/j.ajem.2024.05.001

18. Chavez MR, Butler TS, Rekawek P, Heo H, Kinzler WL. Chat generative pre-trained transformer: Why we should embrace this technology. *Am J Obstet Gynecol*. 2023;228(6):706-711.

doi: 10.1016/j.ajog.2023.03.010

19. Vaswani A, Shazeer N, Parmar N, *et al*. *Attention is all You Need*. ArXiv; 2017.

doi: 10.48550/arXiv.1706.03762

20. Li N, Liu S, Liu Y, Zhao S, Liu M. Neural Speech Synthesis with Transformer Network. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*; 2018. p. 6706-6713.

doi: 10.1609/aaai.v33i01.33016706

21. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by Generative Pretrained Transformer 3 chatbot for clinical vignettes with common chief complaints: A pilot study. *Int J Environ Res Public Health.* 2023;20(4):3378.

doi: 10.3390/ijerph20043378

22. Mizuta K, Hirosawa T, Harada Y, Shimizu T. Can ChatGPT-4 evaluate whether a differential diagnosis list contains the correct diagnosis as accurately as a physician? *Diagnosis* (*Berl*). 2024;11(3):321-324.

doi: 10.1515/dx-2024-0027

23. Ouyang L, Wu J, Jiang X, *et al*. *Training Language Models to Follow Instructions with Human Feedback*. ArXiv; 2022.

doi: 10.48550/arXiv.2203.02155

24. Johnson D, Goodman R, Patrinely J, *et al*. Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the Chat-GPT model. *Res Sq*. 2023:rs.3.rs-2566942.

doi: 10.21203/rs.3.rs-2566942/v1

25. Shahsavar Y, Choudhury A. User Intentions to use ChatGPT for self-diagnosis and health-related purposes: Cross-sectional survey study. *JMIR Hum Factors*. 2023;10:e47564.

doi: 10.2196/47564

26. Ueda D, Walston SL, Matsumoto T, Deguchi R, Tatekawa H, Miki Y. Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz. *BMC Digital Health*. 2024;2(1):4.

doi: 10.1186/S44247-023-00058-5

27. Cheng K, Li Z, He Y, *et al*. Potential Use of artificial intelligence in infectious disease: Take chatGPT as an example. *Ann Biomed Eng*. 2023;51(6):1130-1135.

doi: 10.1007/s10439-023- 03203-3

28. Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK neurology specialty certificate examination. *BMJ Neurol Open*. 2023;5(1):e000451.

doi: 10.1136/bmjno-2023-000451

29. Cheng K, Guo Q, He Y, *et al*. Artificial intelligence in sports medicine: Could GPT-4 make human doctors obsolete? *Ann Biomed Eng*. 2023;51(8):1658-1662.

doi: 10.1007/s10439- 023-03213-1

30. Biswas S. ChatGPT and the future of medical writing. *Radiology*. 2023;307(2):e223312.

doi: 10.1148/radiol.223312

31. Heng JJ, Teo DB, Tan LF. The impact of chat generative pre-trained transformer (ChatGPT) on medical education. *Postgrad Med J*. 2023;99(1176):1125-1127.

doi: 10.1093/postmj/qgad058

32. Kaliyadan F, Seetharam KA. ChatGPT-Quo vadis? *Indian Dermatol Online J*. 2023;14(4):457-458.

    doi: 10.4103/idoj.idoj_344_23

33. El Haj M, Boutoleau-Bretonnière C, Gallouj K, *et al*. ChatGPT as a diagnostic aid in Alzheimer's disease: An exploratory study. *J Alzheimers Dis Rep*. 2024;8(1):495-500.

    doi: 10.3233/adr-230191

34. Egli A. ChatGPT, GPT-4, and other large language models: The next revolution for clinical microbiology? *Clin Infect Dis*. 2023;77(9):1322-1328.

    doi: 10.1093/cid/ciad407

35. Bahrini A, Khamoshifar M, Abbasimehr H, *et al*. ChatGPT: Applications, Opportunities, and Threats. In: *2023 Systems and Information Engineering Design Symposium (SIEDS)*. United States: IEEE; 2023. p. 274-279.

    doi: 10.1109/sieds58326.2023.10137850

36. Temsah MH, Jamal A, Aljamaan F, Al-Tawfiq JA, Al-Eyadhy A. ChatGPT-4 and the global burden of disease study: Advancing personalized healthcare through artificial intelligence in clinical and translational medicine. *Cureus*. 2023;15(5):e39384.

    doi: 10.7759/cureus.39384

37. Santhoshkumar SP, Beaulah HL, Susithra K. A study on scope of artificial intelligence in diagnostic medicine. *Recent Res Rev J*. 2023;2(1):39-53.

    doi: 10.36548/rrrj.2023.1.04

38. Rathore FA, Rathore MA. The emerging role of artificial intelligence in healthcare. *J Pak Med Assoc*. 2023;73(7): 1368-1369.

    doi: 10.47391/JPMA.23-48

39. Jain P, Zameer F, Khan K, *et al*. Artificial intelligence in diagnosis and monitoring of atopic dermatitis: From pixels to predictions. *Artif Intell Health*. 2024;1(2):48-65.

    doi: 10.36922/aih.2775

40. Van Sassen C, Mamede S, Bos M, Van den Broek W, Bindels P, Zwaan L. Do malpractice claim clinical case vignettes enhance diagnostic accuracy and acceptance in clinical reasoning education during GP training? *BMC Med Educ*. 2023;23(1):474.

    doi: 10.1186/S12909-023-04448-1

41. Elasan S, Ateş Y. Artificial intelligence (AI) and ethics in medicine at a global level: Benefits and risks. In: *Current Researches in Health Sciences-II*. Türkiye: Özgür Yayınları; 2023.

    doi: 10.58830/ozgur.pub128.c508

42. Aquino YS. Making decisions: Bias in artificial intelligence and data-driven diagnostic tools. *Aust J Gen Pract*. 2023;52(7):439-442.

    doi: 10.31128/ajgp-12-22-6630

43. Chakraborty S, Chopra H, Akash S, Chakraborty C, Dhama K. Advances in artificial intelligence (AI)-based diagnosis in clinical practice-correspondence. *Ann Med Surg* (*Lond*). 2023;85(7):3757-3758.

    doi: 10.1097/MS9.0000000000000959

44. Kanter GP, Packel EA. Health care privacy risks of AI chatbots. *JAMA*. 2023;330(4):311-312.

    doi: 10.1001/jama.2023.9618

45. Marks M, Haupt CE. AI chatbots, health privacy, and challenges to HIPAA compliance. *JAMA*. 2023;330(4): 309-310.

    doi: 10.1001/jama.2023.9458

46. Karako K, Song P, Chen Y, Tang W. New possibilities for medical support systems utilizing artificial intelligence (AI) and data platforms. *Biosci Trends*. 2023;17(3):186-189.

    doi: 10.5582/bst.2023.01138

47. Paladino MS. Artificial intelligence in medicine. Ethical reflections from the thought of Edmund Pellegrino. *Cuad Bioet*. 2023;34(110):25-35.

    doi: 10.30444/CB.140

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Discovering predictive features of multiple sclerosis from clinically isolated syndrome with machine learning

Minh Sao Khue Luu[1†*], Bair N. Tuchinov[1,2†], Anna I. Prokaeva[2,3†], Denis S. Korobko[2,3], Nadezhda A. Malkova[2,3], and Andrey A. Tulupov[1,2]

[1]Stream Data Analytics and Machine Learning Laboratory, Novosibirsk State University, Novosibirsk, Russia

[2]The Institute International Tomography Center of the Russian Academy of Sciences, Novosibirsk, Russia

[3]State Novosibirsk Regional Clinical Hospital, Novosibirsk, Russia

Abstract

Accurately predicting the progression of clinically isolated syndrome (CIS) to multiple sclerosis (MS) is crucial for early intervention and management. This study employs a range of machine learning models, including categorical boosting, extreme gradient boosting, light gradient boosting machine, random forest, support vector machine, and logistic regression, to classify CIS patients based on their likelihood of developing MS. Our best model achieves and demonstrates superior predictive accuracy of 0.9312, measured using the area under the curve metric. In addition, we apply explainability techniques to determine the most influential features driving the predictions, identifying which CISs are most indicative of MS progression. Furthermore, we explore feature interactions to detect relationships between features, providing a deeper understanding of the underlying mechanisms. The study utilizes public data from 273 CISs patients, offering significant contributions to the clinical management and early diagnosis of MS.

Keywords: Clinically isolated syndromes; Multiple sclerosis; Machine learning; Binary classification; Predictive features; Model explainability

†These authors contributed equally to this work.

***Corresponding author:**
Minh Sao Khue Luu
(khue.luu@g.nsu.ru)

Citation: 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

Received: July 16, 2024

Accepted: August 26, 2024

Published Online: September 24, 2024

1. Introduction

Multiple sclerosis (MS) is a chronic, inflammatory, demyelinating disease of the central nervous system[1,2] that affects approximately 2.9 million people worldwide.[3] MS primarily affects young adults, with 70 – 80% of patients having an age of onset between 20 and 40 years, and represents the leading cause of non-traumatic disability in young adults.[3] The consequences of MS can be profound, affecting physical abilities, cognitive functions, emotional well-being, and overall quality of life.[4-8] As such, MS exerts a large personal and societal socioeconomic burden due to the high cost of treatment and additional care related to permanent neurologic disability.[9,10] MS is often first suspected when an individual experiences a clinically isolated syndrome (CIS), which is a sudden onset of neurological symptoms lasting at least 24 h due to inflammation or demyelination in the central nervous system.[11] As these initial episodes recur and additional neurological symptoms appear over time, the condition may develop into clinically definite MS

(CDMS), a diagnosis of MS based on clinical criteria, such as the McDonald criteria.[12-14]

Predicting this progression is difficult due to the disease's heterogeneous nature and differences in lesion appearance and progression on magnetic resonance imaging (MRI) scans. MS manifests through a wide range of symptoms, including visual, sensory, and motor dysfunctions, as well as cognitive impairments, with the severity of these symptoms varying significantly between individuals.[15-17] The disease can present in different forms – such as relapsing-remitting, primary progressive, secondary progressive, and progressive relapsing – each with distinct symptoms, presentations, and progression patterns, making prediction even more complex.[18] Moreover, the progression of MS is unpredictable; some patients experience a slow decline over many years, while others deteriorate rapidly. The characteristics of MS lesions, including their number, size, and location, also differ greatly between patients, contributing to the variability in symptoms and disease progression. The progression and severity of MS vary significantly among individuals, but in the absence of timely diagnosis and treatment, patients face severe disability. Early intervention in MS is crucial, as it can significantly delay the progression of disability and improve long-term outcomes for patients. Timely treatment not only reduces the frequency of relapses but also helps in preserving neurological function, leading to a better overall quality of life for individuals with MS.[19]

There have been studies using statistical analysis to study the progression of CIS to CDMS.[20-24] Some studies specifically utilized MRI data to predict the diagnosis of CDMS evolved from CIS.[25-30] Diminished sense of vibration and proprioception, spinal cord MRI lesions were found among CIS patients that later developed MS.[31] The role of viral infections in CIS and their potential trigger mechanism in MS remains controversial; a direct relationship has been found between the history of infectious mononucleosis due to the Epstein–Barr virus and an increased risk of developing CDMS.[32] Notably, machine learning (ML) has emerged as a crucial tool in this predictive process, analyzing clinical data to identify patterns and risk factors associated with the progression to CDMS.[33-41] However, most research relies on private datasets that are not accessible to external researchers, making it difficult to reproduce results and establish benchmarks for developing ML algorithms for CDMS prediction. The existing literature still lacks a specific set of features that can accurately predict the progression of CIS to CDMS. Furthermore, there is currently no single, generally accepted method for predicting the progression of CIS to CDMS. However, it has been shown that the use of ML models or generative artificial intelligence (AI) platforms helped to speed up and facilitate the diagnosis of CDMS compared to the real-life clinical timeline.[42]

Recently, Rasouli *et al.*[43] utilized extreme gradient boosting (XGBoost) on a public dataset of 273 patients to predict CDMS, achieving an impressive area under the curve (AUC) of 0.918. They identified key predictive features using the SHapley Additive exPlanations (SHAP)[44] library. Building on this work, our study explores the same dataset with six advanced ML models: Categorical Boosting (CatBoost),[45] XGBoost,[46] light gradient boosting machine (LGBM),[47] random forest (RF),[48] support vector machine (SVM),[49] and logistic regression (LR).[50] Each model undergoes training with five-fold cross-validation.[51] To determine feature importance for predictions, we apply SHAP across the five folds of these ML models.

Our study makes notable contributions to the field:

(1) Enhanced model performance: Our CatBoost model achieves an AUC of 0.9312, demonstrating superior predictive accuracy. The XGBoost model also performs well with an AUC of 0.9202, which is slightly higher than that reported in a recent study.[43]
(2) Key feature identification: We identify the most influential features contributing to CDMS prediction. We also observe that MRI-based features are universally critical, while symptom-related and schooling features vary in importance, suggesting unique model strengths and socioeconomic implications.
(3) Comprehensive feature analysis: We conduct an in-depth analysis of feature interactions and the impact of initial symptoms on the progression to CDMS, revealing new patterns and relationships that enhance our understanding of the disease.

2. Methodology

Figure 1 summarizes our workflow, which involves several critical steps to ensure robust model evaluation and insightful feature analysis. After preprocessing, we split the data into five folds for cross-validation. This ensures that each fold was used as a test set once, and the remaining folds were used for training. We then proceeded to train six ML models: CatBoost, XGBoost, LGBM, RF, SVM, and LR. We employed Optuna, a hyperparameter optimization framework, to identify the optimal hyperparameters for each model. For each of the five folds, we trained all six models on four folds and tested them on the remaining fold. We repeated this process for each fold, resulting in 30 trained models, with six models per fold. During this stage, we obtained predictions for each test fold and concatenate predictions of all folds for each model. Then, we calculated multiple evaluation metrics to compare model

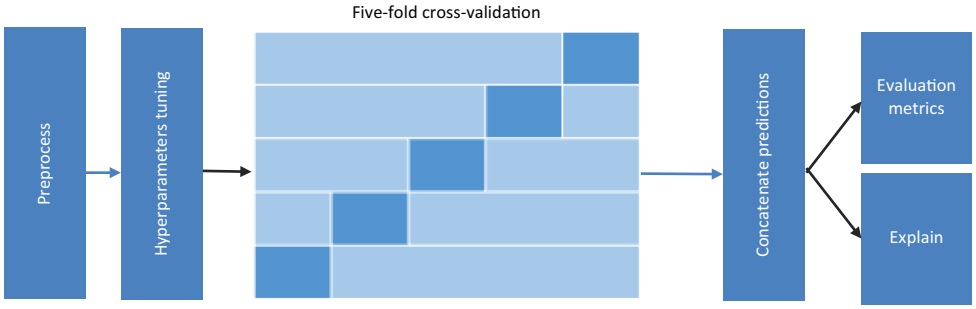


Figure 1. Diagram of the process for training and testing six classifiers for the prediction of clinically definite multiple sclerosis (CDMS) from clinically isolated syndrome (CIS) data

performance. After training and evaluating the models, we analyzed each model's feature importance to determine the most influential features in the predictions. By examining, the features that were consistently ranked highly across different models, we gained insights into the key factors that contribute to the diagnosis. We then utilized the SHAP library to further analyze the interactions between features. SHAP values help us understand not only the individual impact of each feature but also how features interact with each other, providing a deeper understanding of the data's underlying patterns. We used software tools such as Python's Pandas and Numpy libraries for data manipulation and basic statistical computations, and libraries such as Scikit-learn and Scipy for model building, evaluation, and statistical testing. In addition, we employed visualization tools such as Matplotlib and Seaborn to illustrate data analysis and model performance metrics.

2.1. Data

We used a public dataset of 273 diagnosed patients with CIS from 2006 to 2010,[52] which includes clinical and neuroimaging data from the first CIS episode and a 10-year follow-up. These patients were monitored over a period to observe whether they developed MS. The dataset includes a variety of features that are potentially relevant for predicting the progression to MS. These features encompass clinical characteristics such as the type of initial symptoms, the presence of specific neurological signs, and results from diagnostic tests like MRI scans. In addition, demographic information such as age, gender, and medical history were also included to provide a holistic view of each patient's profile. Patients were classified as CDMS or non-CDMS based on the McDonald 2010 criteria.[12] These features are described in Table 1.

2.2. Preprocessing

Our preprocessing steps included several critical transformations and imputations to ensure robust model performance and prevent data leakage. First, we removed the Initial_EDSS and Final_EDSS columns since they contain values exclusively for the CDMS class and null values for the non-CDMS class. This discrepancy can lead to overfitting, allowing the model to achieve perfect AUC without considering other features; thus, we excluded these columns from our study.

To address the missing values, we imputed the Initial_Symptom column with the mode value, acknowledging that these numbers represent categorical data. In addition, we imputed the schooling column with the median value due to its numerical nature. Specifically, there was one missing value in each of these columns.

To enhance the interpretability and granularity of our data, we split some columns into multiple binary columns. The Initial_Symptom column, which contains values from 1 to 15, indicating the presence of one or more symptoms, was divided into four binary columns: Symptom_Visual, Symptom_Sensory, Symptom_Motor, and Symptom_Others. All conversions had been verified by our expert neurologists and can be used during the clinical diagnostic process. This transformation allowed us to analyze the effect of each specific symptom on the prediction independently, thereby improving our understanding of symptom-specific impacts on the diagnosis. Similarly, we split the "Mono or Polysymptomatic" column into two binary columns: Mono_Symptomatic and Poly_Symptomatic, which differentiates between patients exhibiting a single symptom versus multiple symptoms. This split provides a clearer representation of symptom complexity in our dataset.

Moreover, we remapped certain numerical columns to a binary format to standardize the data and ensured consistency in the model input. Specifically, we remapped Oligoclonal_Bands, Gender, Breastfeeding, and Varicella columns to binary values where 0 indicates a negative response, 1 indicates a positive response, and -1 represents unknown values. This binary transformation simplified these categorical features into a consistent format that could be easily interpreted by ML algorithms. It also

Table 1. Descriptions of columns in the dataset

Column	Description
ID	Unique identifier for each patient (integer)
Age	Age of the patient (years)
Schooling	Duration of patient's education (years)
Gender	Gender of the patient (1=Male, 2=Female)
Breastfeeding	Breastfeeding history (1=Yes, 2=No, 3=Unknown)
Varicella	Varicella (chickenpox) history (1=Positive, 2=Negative, 3=Unknown)
Initial_Symptoms	Type of initial symptoms experienced
Mono_or_Polysymptomatic	Symptom presentation type (1=Monosymptomatic, 2=Polysymptomatic, 3=Unknown)
Oligoclonal_Bands	Oligoclonal bands status (0=Negative, 1=Positive, 2=Unknown)
LLSSEP	Lower limb somatosensory evoked potentials (0=Negative, 1=Positive)
ULSSEP	Upper limb somatosensory evoked potentials (0=Negative, 1=Positive)
VEP	Visual-evoked potentials (0=Negative, 1=Positive)
BAEP	Brainstem auditory-evoked potentials (0=Negative, 1=Positive)
Periventricular_MRI	MRI results for brain's periventricular area (0=Negative, 1=Positive)
Cortical_MRI	MRI results for brain's cortex (0=Negative, 1=Positive)
Infratentorial_MRI	MRI results for brain's lower regions (0=Negative, 1=Positive)
Spinal_Cord_MRI	MRI results for spinal cord (0=Negative, 1=Positive)
Initial_EDSS	Initial disability score (Expanded Disability Status scale)
Final_EDSS	Final disability score (Expanded Disability Status scale)
Group	Diagnostic group (1=CDMS, 2=Non-CDMS)

Abbreviations: CDMS: Clinically definite multiple sclerosis; MRI: Magnetic resonance imaging.

facilitated the handling of unknown values, ensuring that the presence of such values did not disrupt the model's learning process.

Finally, we converted all values to type float. This conversion is essential as many ML algorithms, including those we were using, expect input features to be in a numerical format to perform mathematical operations effectively.

2.3. Hyperparameters tuning

For each model, we ran an Optuna study with 500 iterations with the goal to maximize the average of AUC, accuracy (ACC), and F1 score. We experimented with a wide range of values for different hyperparameters of the models. In a separate file, we saved sets of the best hyperparameters for models.

2.4. Evaluation metrics

In this study, we employed several evaluation metrics to comprehensively assess the performance of our ML models in predicting CDMS from CIS. The selected metrics provide a detailed understanding of the models' capabilities in various aspects, ensuring a robust evaluation.

The AUC measures the area under the receiver operating characteristic curve, plotting the true positive rate against false-positive rate, ranging from 0 to 1.[53] Higher AUC values indicate better performance, with 1 being perfect discrimination and 0.5 indicating random chance.

The ACC measures the proportion of correct predictions out of all predictions, also ranging from 0 to 1. An accuracy of 1 denotes perfect classification, while 0.5 suggests random guessing.[54] This metric may not be reliable for imbalanced datasets.

The precision measures the accuracy of positive predictions, indicating the proportion of correct positive predictions.[54]

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

Where TP = true positives, and FP = false positives.

The recall measures the proportion of actual positives correctly identified,[54] crucial for ensuring that all positive cases are detected, particularly important in CDMS prediction where missing positive cases can have severe consequences.

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

Where TP = true positives, and FN = false negatives.

The F1 score is the harmonic mean of precision and recall,[54] balancing accuracy in positive predictions and the ability to capture all positive instances, with higher values indicating better performance.

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The specificity measures the proportion of actual negatives correctly identified,[54] crucial for minimizing false alarms.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Where TN = true negatives, and FP = false positives.

2.5. Statistical test

The statistical testing process started with a normal distribution check on all evaluation metrics for each model across all folds. We chose the Shapiro–Wilk test[55] due to its sensitivity in detecting deviations from normality in small sample sizes. On analyzing the results, we found that some metrics were not normally distributed, such as recall of LGBM and specificity of RF. Therefore, we proceeded with the non-parametric Friedman test[56] to compare models' performance. This test is ideal for comparing metrics of multiple models across different folds without relying on the assumption of normal distribution. When the Friedman test indicated significant differences among model metrics, we, further, investigated these differences using the Nemenyi *post hoc* test[57] for comparison between pairs of models.

2.6. Explainability

To explain our models, we leveraged SHAP, a highly regarded technique in Explainable AI within the medical and healthcare domain,[58] to identify important factors influencing CDMS predictions. SHAP provides both global and local insights into feature importance, helping us understand overall model behavior and individual predictions. These techniques ensure our models were not only accurate but also transparent, enhancing their trustworthiness for predicting CDMS from CIS.

For tree-based models such as CatBoost, XGBoost, LGBM, and RF, we used TreeExplainer with parameters model_output set to "raw" and feature_perturbation set to "tree_path_dependent." This setup captures the raw output of the models before applying any logistic function and uses the decision tree structures to perturb features. For SVM, we applied KernelExplainer with the number of background samples automatically selected by SHAP. This ensures that the sample size provided a satisfactory approximation without requiring excessive computation time. LR was explained with LinearExplainer to suit the linear nature of the model. The explainer was also set to consider the correlation dependence of features during the explanation.

We calculated SHAP values for each row in the test set across all folds and models. SHAP interaction values are available only for the four tree-based models, as KernelExplainer and LinearExplainer do not support interactions. For each model, we first averaged the mean absolute SHAP values for all rows and folds, followed by min-max normalization to ensure metrics were on a consistent scale. This process produced a matrix of 20 rows of features with six columns of models. This matrix allowed us to analyze the overall impact of features. Next, to rank features, we calculated the mean SHAP values of all six models for each feature, then sorted them. This step provides a clear view of which CIS is driving the prediction of CDMS across all ML models.

3. Results

Overall, we found that gradient-boosted models – CatBoost, LGBM, and XGBoost – consistently outperformed other models in predicting CDMS, though the performance differences were not statistically significant. CatBoost achieved the highest AUC and showed the best overall balance between precision and recall. Periventricular_MRI, Infratentorial_MRI, and Oligoclonal_Bands are features that have a big effect on how well different ML models predict CDMS. Among the features, Oligoclonal_Bands and Periventricular_MRI showed strong interaction.

3.1. Model performance

3.1.1. Model performance metrics

To compare model performance, we evaluated predictions across five-fold cross-validation, focusing on metrics such as AUC, ACC, F1 score, precision, recall, and specificity. Gradient-boosted tree models – CatBoost, LGBM, and XGBoost – consistently outperformed others, likely due to their iterative error-correcting nature. These models achieved higher AUCs, better precision-recall balance, and superior F1 scores, indicating stronger class separation and accuracy in predicting positive cases while reducing false positives and negatives. Table 2 presents the mean metrics for each model across the five folds.

Table 2. Evaluation metrics for six machine learning models across five folds

Model	AUC	ACC	F1 score	Precision	Recall	Specificity
CatBoost	**0.9312**	**0.8791**	**0.8675**	**0.8710**	**0.8640**	**0.8919**
XGBoost	0.9202	0.8645	0.8514	0.8548	0.8480	0.8784
LGBM	0.9150	**0.8791**	**0.8675**	**0.8710**	**0.8640**	**0.8919**
RF	0.9097	0.8388	0.8295	0.8045	0.8560	0.8243
SVM	0.8985	0.8168	0.8031	0.7907	0.8160	0.8176
LR	0.8922	0.8132	0.7935	0.8033	0.7840	0.8378

Note: The values in boldface mean highest values in the columns.
Abbreviations: CatBoost: Categorical boosting; LGBM: Light gradient boosting machine; LR: Logistic regression; RF: Random forest; SVM: Support vector machine; XGBoost: Extreme gradient boosting; AUC: Area under the curve.

CatBoost emerged as the top performer with an AUC of 0.9312, a balanced F1 score of 0.8675, precision of 0.8710, recall of 0.8640, and specificity of 0.8919, showcasing its effectiveness in both positive and negative case identification. LGBM matched CatBoost closely in these metrics, making it another strong classifier. XGBoost also performed well with an AUC of 0.9202 but showed slightly less balanced precision and recall compared to CatBoost and LGBM. RF, while having a solid AUC of 0.9097, exhibited less balance in precision and recall, resulting in more false positives.

SVM and LR demonstrated lower performance overall. SVM's AUC of 0.8985 and F1 score of 0.8031 indicated moderate effectiveness but higher false-positive rates. LR had the lowest performance, with an AUC of 0.8922 and the lowest balance between precision and recall, making it the least effective model evaluated.

The confusion matrices (Figure 2) visually confirm that CatBoost and LGBM offered the best accuracy, each with 108 true positives and 132 true negatives, while XGBoost had slightly more misclassifications. The ROC curves (Figure 3) further highlight CatBoost's superior performance, especially at lower false-positive rates, making it highly suitable for applications where minimizing false alarms is critical. Overall, CatBoost, XGBoost, and LGBM are the most reliable models for CDMS prediction, with CatBoost excelling in sensitivity, reducing false negatives, and ensuring accurate diagnoses.

3.1.2. Statistical significance of model performance

Figure 4 shows the Nemenyi *post hoc* test heatmap. It shows that CatBoost, LGBM, and XGBoost consistently ranked as the best models. However, the differences between these models were not statistically significant, as shown by higher *P*-values in most metrics. However, we observe significant differences between these top models and the lower-performing models, particularly LR and SVM. For instance, *P*-values from the Nemenyi test showed that CatBoost did much better than LR in AUC, supporting earlier findings that gradient-boosted models are better at predicting CDMS. This statistical validation underscores the reliability of the gradient-boosted models, particularly CatBoost and LGBM, which are the most effective choices for this task.

Following the statistical analysis, the critical difference plots in Figure 5, which rank the models across multiple metrics, provide a visual substantiation of the findings. In the plot, the models are ranked based on their performance and are connected by lines if the differences

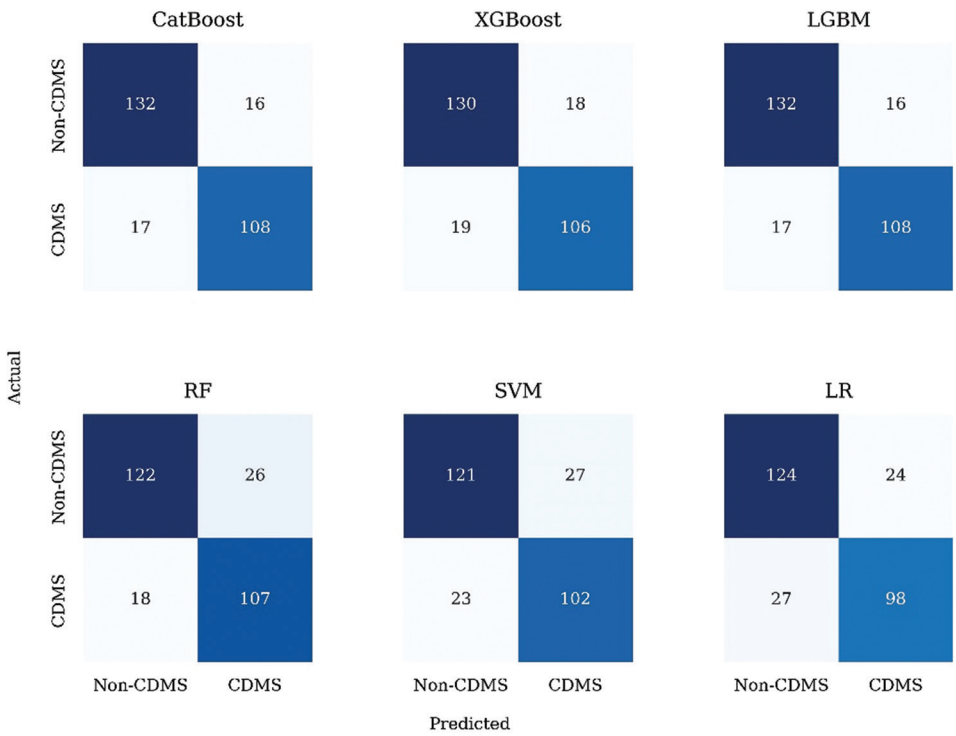


Figure 2. Confusion matrices for six machine learning models in clinically definite multiple sclerosis (CDMS) classification
Abbreviations: CatBoost: Categorical boosting; LGBM: Light gradient boosting machine; LR: Logistic regression; RF: Random forest; SVM: Support vector machine; XGBoost: Extreme gradient boosting.

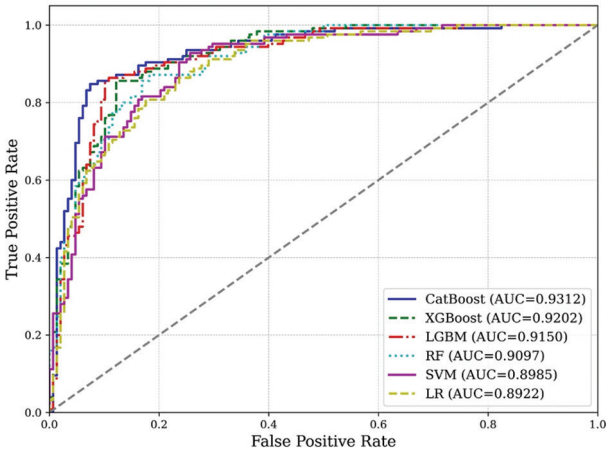


Figure 3. Comparison of receiver operating characteristic (ROC) curves for six machine learning models
Abbreviations: AUC: Area under the curve; CatBoost: Categorical boosting; LGBM: Light gradient boosting machine; LR: Logistic regression; RF: Random forest; SVM: Support vector machine; XGBoost: Extreme gradient boosting.

in their ranks are not statistically significant, meaning they perform similarly. The plot shows that CatBoost, LGBM, and XGBoost consistently rank as top performers, with minimal and statistically insignificant differences among them, indicating their similar effectiveness. In contrast, SVM and LR are consistently lower in the rankings, confirming their comparatively weaker performance.

3.2. Feature importance analysis

To identify important features for CDMS diagnosis prediction, we calculated mean absolute SHAP values of features across six ML models over five validation folds, then illustrated their rankings, as shown in Figure 6. We observed that the presence or absence of lesions in brain MRI and clinical tests is the most critical factor, while demographic features and other clinical assessments provide additional but lesser contributions.

The top three features – Periventricular_MRI, Infratentorial_MRI, and Oligoclonal_Bands – had

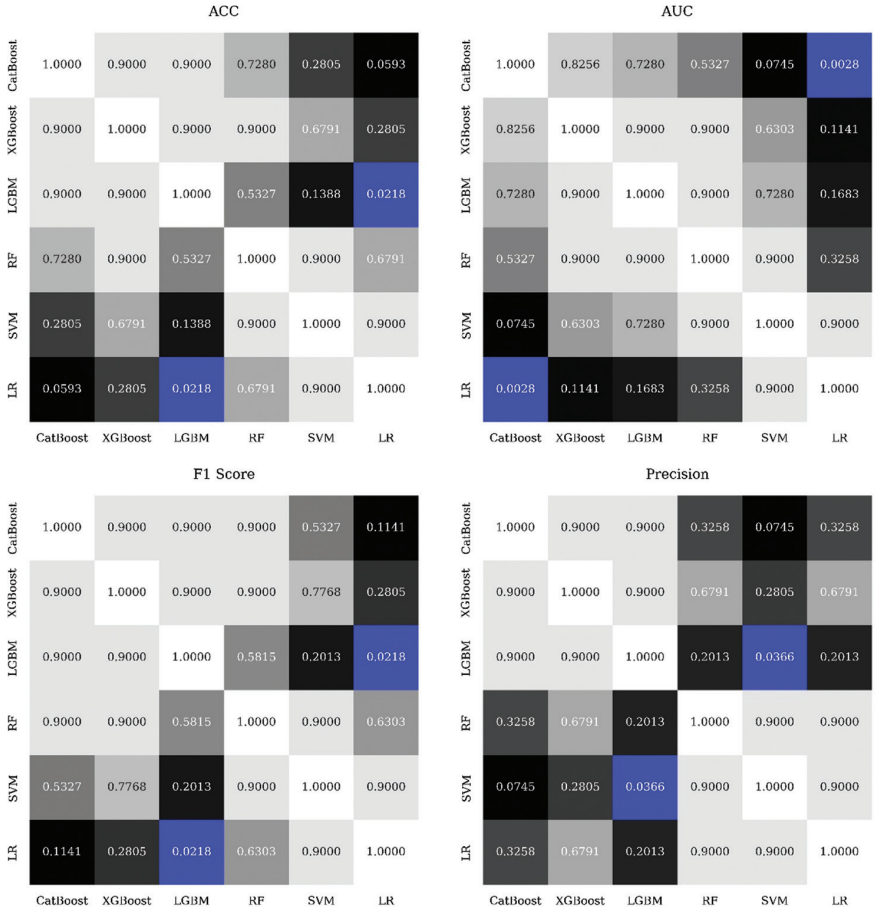


Figure 4. Nemenyi *post hoc* test heatmap for pairwise model performance comparison across multiple metrics
Abbreviations: CatBoost: Categorical boosting; LGBM: Light gradient boosting machine; LR: Logistic regression; RF: Random forest; SVM: Support vector machine; XGBoost: Extreme gradient boosting.

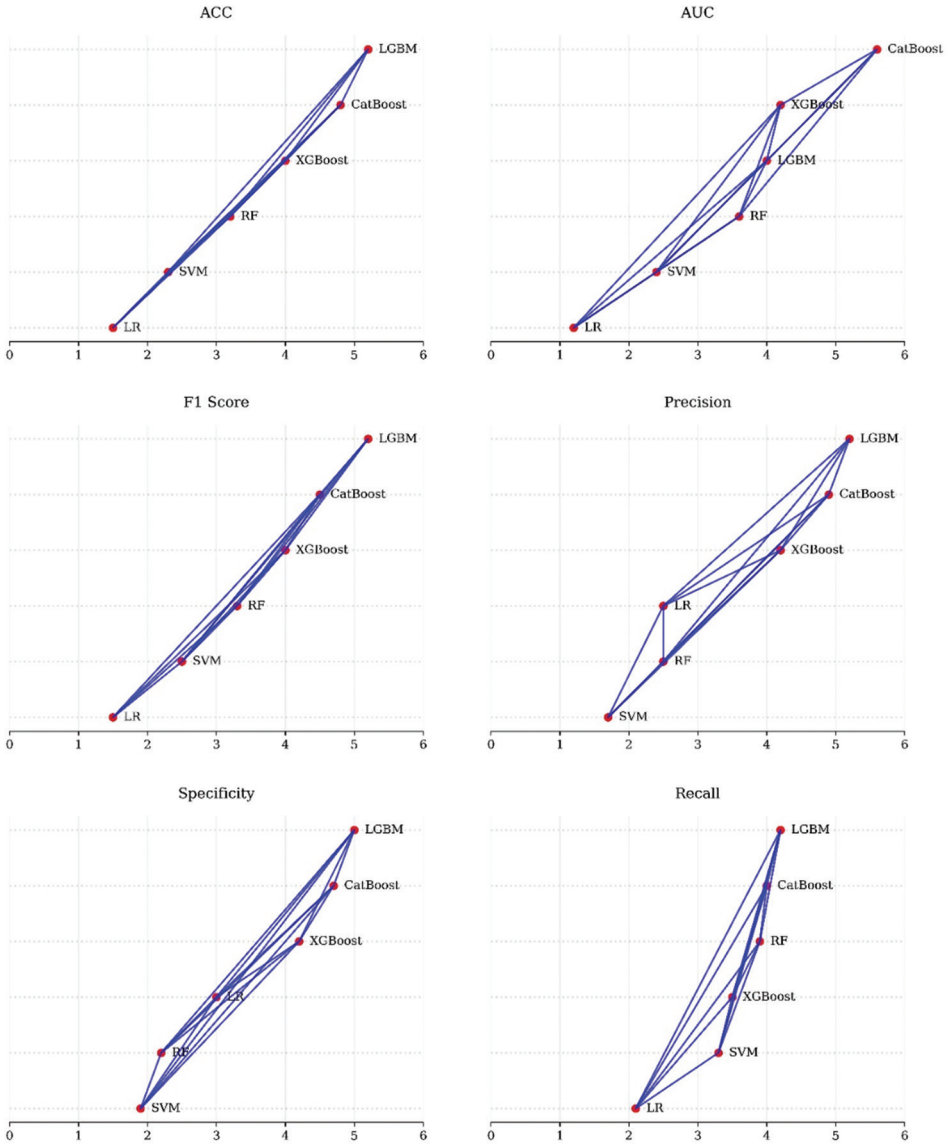


Figure 5. Critical difference plot for ranking model performance across multiple metrics
Abbreviations: CatBoost: Categorical boosting; LGBM: Light gradient boosting machine; LR: Logistic regression; RF: Random forest; SVM: Support vector machine; XGBoost: Extreme gradient boosting.

significantly higher SHAP values, indicating their strong influence on the models' predictions. Specifically, Periventricular_MRI, with a mean absolute SHAP value of 0.8501, stands out as the most influential feature, suggesting that it has the most substantial impact on the likelihood of a CDMS diagnosis. Infratentorial_MRI and Oligoclonal_Bands follow, with values of 0.5212 and 0.49, respectively, highlighting their substantial roles in the prediction process.

Schooling, with a value of 0.4388, is also notable, emphasizing the relevance of the number of years spent in school in the models' decisions. Symptom-related features, including Symptom_Motor (0.3586), Symptom_Other (0.3048), and Symptom_Sensory (0.2604), further underscore the importance of clinical presentations in diagnosing CDMS. Breastfeeding, gender, and age, with SHAP values around 0.2606, 0.2602, and 0.2293, respectively, are moderately influential, suggesting that most demographic factors play a role but are less critical than specific medical indicators.

To further analyze important features of each ML model, we generated a heatmap of SHAP values across

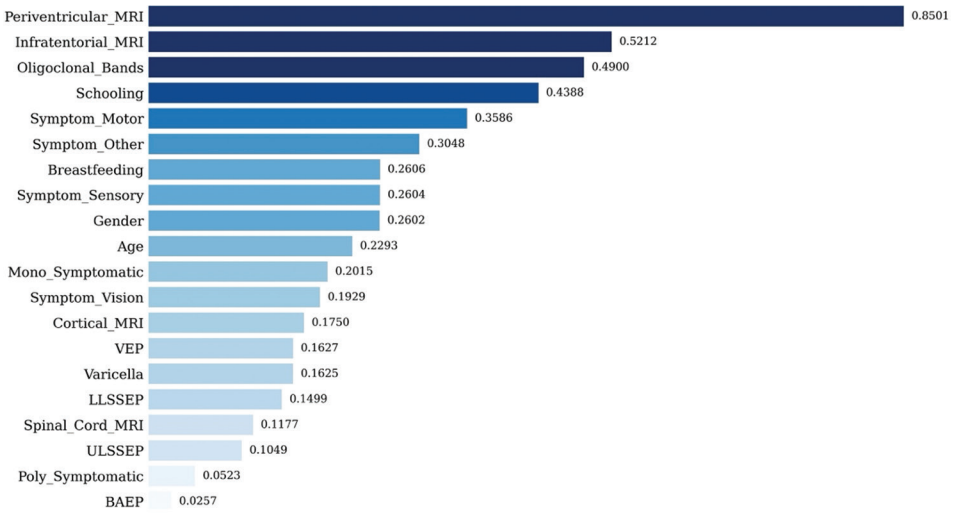


Figure 6. Feature rankings based on mean absolute SHAP values of six machine learning models.

models, as shown in Figure 7. We normalize the values to the same range to facilitate direct comparison between the features of the models. Darker colors in the cells indicate greater feature influence. If a row has many dark cells, it means that the feature is robust and significantly influences multiple models. Columns with similar color patterns suggest that those models utilize the same features for their predictions. A key observation from the heatmap is the dominant importance of MRI-based features across all models. These features consistently rank high in importance, underscoring the critical role of advanced imaging in accurately diagnosing CDMS. This consensus among different models highlights the reliability of MRI features in providing significant predictive power for CDMS. Interestingly, the group of symptom-related features also shows considerable importance across several models. This indicates that specific symptoms of CDMS are vital indicators and play a substantial role in the classification task. Moreover, the schooling feature emerges as surprisingly important in models such as LGBM and XGBoost. These points to a potential link between years spent in school and CDMS, hinting at underlying socioeconomic or lifestyle factors that could influence the disease's development or progression. Such insights emphasize the importance of considering a wide array of features, beyond just biological markers, to gain a holistic understanding of CDMS predictors.

Grouping models reveals some interesting patterns in how they prioritize features. CatBoost, XGBoost, and RF tend to group together in terms of feature importance, focusing heavily on similar features such as Periventricular_MRI, Oligoclonal_Bands, and Infratentorial_MRI. This suggests that these tree-based models may be capturing similar relationships in the data, despite having different internal mechanisms for decision-making. On the other hand, SVM and LR share some common ground, especially with features such as Infratentorial_MRI, Oligoclonal_Bands, and Schooling. Their SHAP values are closer in magnitude for these features, indicating that they might be detecting similar underlying patterns, even though SVM operates in a higher-dimensional space compared to the more straightforward linear approach of regression. Finally, LightGBM stands out, with higher importance placed on features such as Schooling and Symptom_Motor compared to other models. This could mean that LightGBM is more sensitive to different types of patterns in the data, potentially giving it an edge in certain scenarios.

3.3. Feature interaction analysis

To analyze feature interactions, we computed the mean absolute SHAP interaction values across all test data, filtered out self-interactions, and sorted the pairs by interaction value. We then extracted the top five interactions for each model, detailed in Table 3. Since KernelExplainer for SVM and LinearExplainer for LR do not support interaction values, this analysis only focused on tree-based models, which are CatBoost, XGBoost, LGBM, and RF.

3.3.1. Overall interaction

Based on the SHAP interaction values, we can evaluate the strength of feature interactions across different models. CatBoost demonstrated moderate interactions, with values ranging from about 0.05 to 0.1, indicating that Oligoclonal_Bands and Periventricular_MRI, as well as Gender and Periventricular_MRI, had notable interactions. XGBoost showed relatively stronger interactions, with several

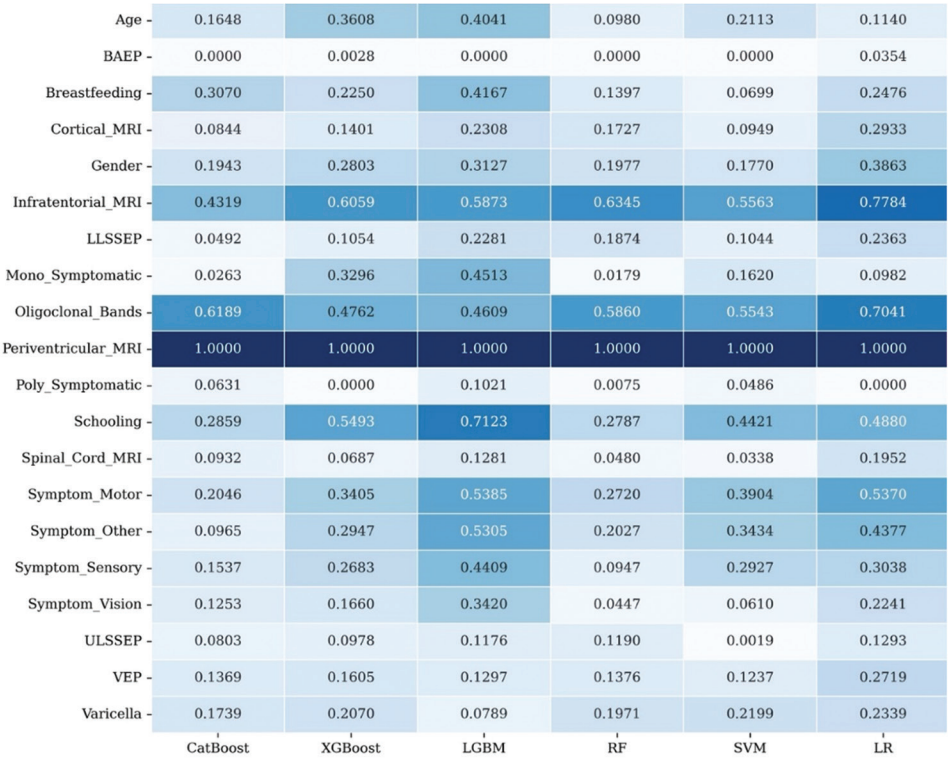


Figure 7. Heatmap of normalized SHAP values of features of six machine learning models
Abbreviations: CatBoost: Categorical boosting; LGBM: Light gradient boosting machine; LR: Logistic regression; RF: Random forest; SVM: Support vector machine; XGBoost: Extreme gradient boosting.

values exceeding 0.1, highlighting significant interactions between Oligoclonal_Bands and Periventricular_MRI, and Gender and Periventricular_MRI. LGBM presented a mix of interaction strengths, with one strong interaction above 0.1 (Symptom_Motor and ULSSEP) and others around 0.06, such as Oligoclonal_Bands and Periventricular_MRI. In contrast, the RF model exhibited much weaker interactions, with values significantly lower than those seen in CatBoost, XGBoost, and LGBM. The highest interaction value in RF, between Oligoclonal_Bands and Periventricular_MRI, was only around 0.009, indicating minimal combined influence on the model's predictions. Other interactions, like Gender with Periventricular_MRI and Infratentorial_MRI with Oligoclonal_Bands, showed similarly low values, typically below 0.008.

3.3.2. Interaction between Oligoclonal_Bands and Periventricular_MRI

Figure 8 illustrates how the interaction between Oligoclonal_Bands and Periventricular_MRI influences the prediction of CDMS across different models, with SHAP interaction values representing the mean absolute values averaged over five cross-validation folds. Each point represents a sample, with the color gradient showing the value of Periventricular_MRI – blue for lower values (closer to 0) and red for higher values (closer to 1). The Y-axis reflects the SHAP value for Oligoclonal_Bands, indicating how much this feature pushes the prediction toward CDMS (positive SHAP value) or non-CDMS (negative SHAP value). In CatBoost and LGBM, there was a noticeable spread in SHAP values when Oligoclonal_Bands equals 1, particularly with high Periventricular_MRI values (red). This suggests that the presence of oligoclonal bands strongly increases the model's confidence in classifying a patient as CDMS, especially when periventricular MRI findings are also significant. XGBoost showed a similar pattern, though with less variability, indicating a moderate interaction effect that still contributes to a higher likelihood of CDMS classification. In contrast, RF displayed much smaller changes in SHAP values, implying that the presence of oligoclonal bands and periventricular MRI findings independently contributes less to the prediction of CDMS. The weak interaction in RF suggests that this model does not heavily rely on the combined presence of these features to classify a patient as CDMS or non-CDMS.

Table 3. Top five feature interactions for tree-based models

Model	Feature 1	Feature 2	Interaction value
CatBoost	Oligoclonal_Bands	Periventricular_MRI	0.097347
	Gender	Periventricular_MRI	0.078551
	Oligoclonal_Bands	Varicella	0.070430
	Symptom_Sensory	Symptom_Vision	0.067328
	Periventricular_MRI	VEP	0.057325
XGBoost	Oligoclonal_Bands	Periventricular_MRI	0.131877
	Gender	Periventricular_MRI	0.099022
	Symptom_Motor	ULSSEP	0.077204
	Age	Cortical_MRI	0.069749
	Mono_Symptomatic	Symptom_Sensory	0.067236
LGBM	Symptom_Motor	ULSSEP	0.106938
	Oligoclonal_Bands	Periventricular_MRI	0.063874
	Mono_Symptomatic	Symptom_Sensory	0.062791
	Mono_Symptomatic	Symptom_Vision	0.062664
	Breastfeeding	Mono_Symptomatic	0.058049
RF	Oligoclonal_Bands	Periventricular_MRI	0.009170
	Gender	Periventricular_MRI	0.007391
	Infratentorial_MRI	Oligoclonal_Bands	0.006605
	Infratentorial_MRI	Periventricular_MRI	0.004550
	Oligoclonal_Bands	Varicella	0.004510

Abbreviations: CatBoost: Categorical boosting; LGBM: Light gradient boosting machine; RF: Random forest; XGBoost: Extreme gradient boosting.

3.3.3. Interaction between gender and Periventricular_MRI

Gender also has a significant interaction with Periventricular_MRI, as depicted in Figure 9. All three models with the highest prediction performance (CatBoost, XGBoost, and LGBM) effectively capture this interaction, as indicated by the distinct clusters of colors in the subplots. For females (Gender = 0), the interaction values clustered around zero, suggesting that MRI findings have a more balanced impact. In contrast, males (Gender = 1) showed more distinct interaction values, especially when periventricular lesions were present, indicating that the presence of these lesions in males has a greater impact on the prediction. This pattern suggests a potential gender-related difference in how periventricular lesions influence model predictions, which could have implications for understanding gender-specific progression or manifestation of neurological conditions associated with these features.

3.3.4. Interaction between Symptom_Motor and ULSSEP

Another significant pair of features is Symptom_Motor and ULSSEP, visualized in Figure 10. ULSSEP stands for upper limb somatosensory evoked potentials. It is a diagnostic test that measures the electrical activity in the brain in response to stimulation of the sensory nerves in

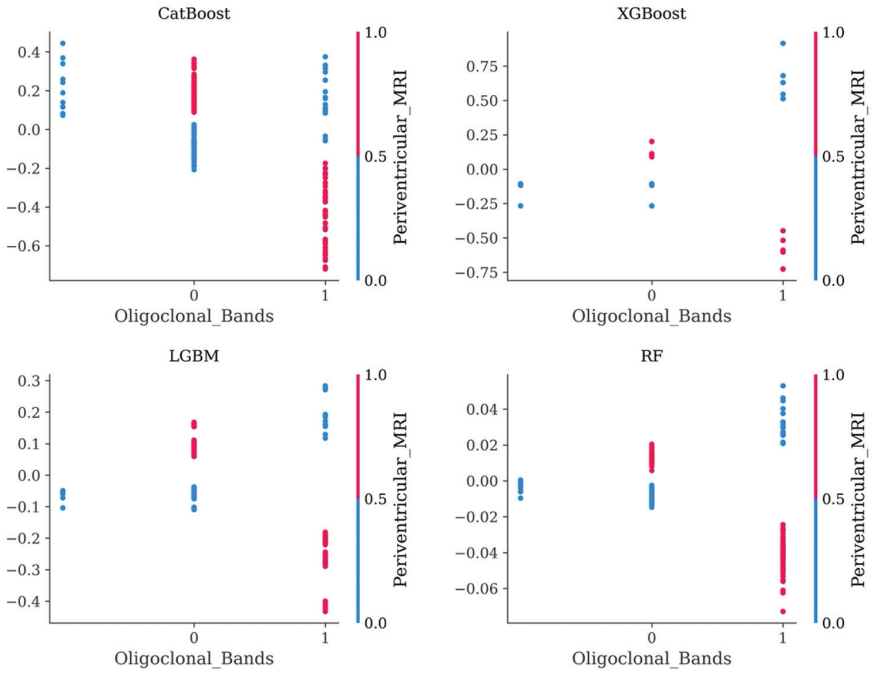


Figure 8. Interactions of Oligoclonal_Bands and Periventricular_MRI on four machine learning models
Abbreviations: CatBoost: Categorical boosting; LGBM: Light gradient boosting machine; RF: Random forest; XGBoost: Extreme gradient boosting.

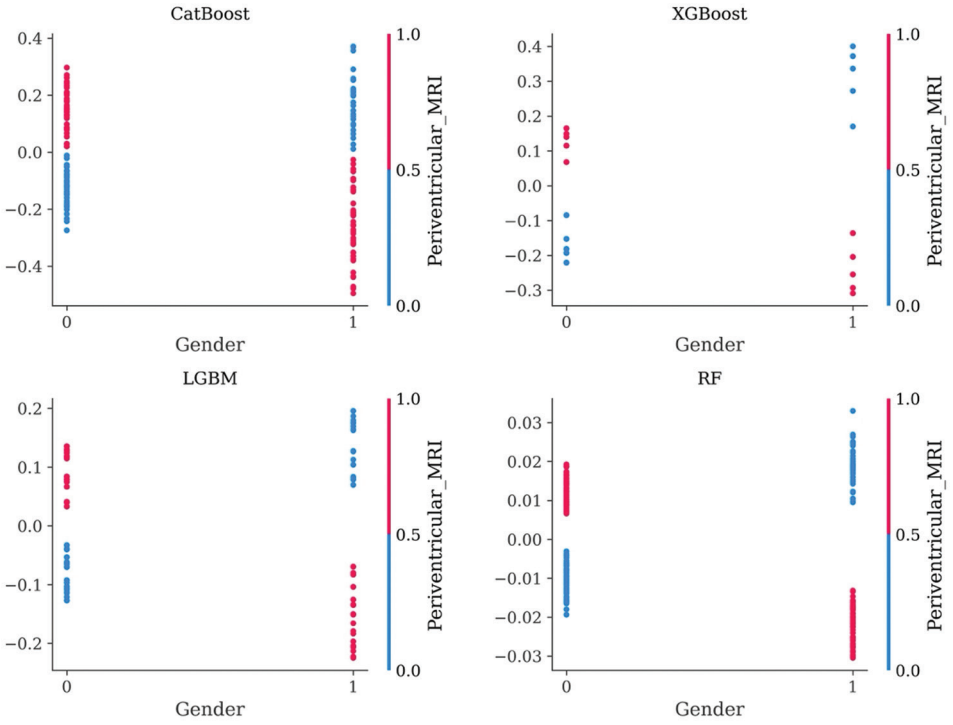


Figure 9. Interactions of Gender and Periventricular_MRI on four machine learning models
Abbreviations: CatBoost: Categorical boosting; LGBM: Light gradient boosting machine; RF: Random forest; XGBoost: Extreme gradient boosting.

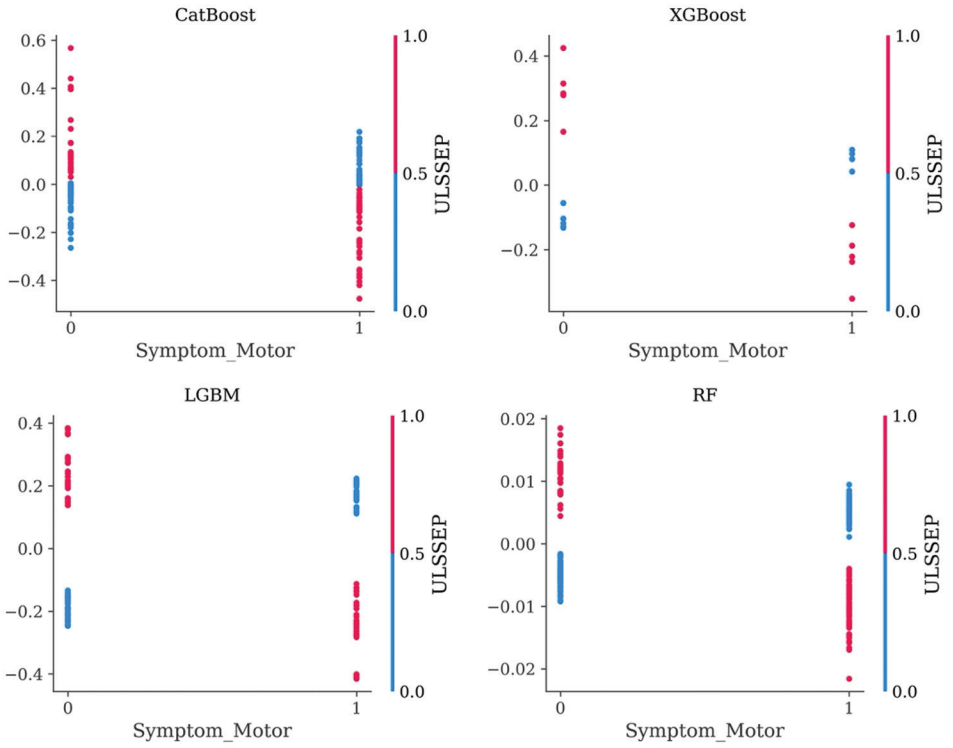


Figure 10. Interactions of Symptom_Motor and ULSSEP on four machine learning models
Abbreviations: CatBoost: Categorical boosting; LGBM: Light gradient boosting machine; RF: Random forest; XGBoost: Extreme gradient boosting.

the upper limbs (arms). A positive ULSSEP indicates an abnormality in the sensory pathways, while a negative result means normal function. The interaction between Symptom_Motor and ULSSEP reveals key neurological insights. When Symptom_Motor is 1 (motor symptoms present), ULSSEP values show a distinct separation, especially for positive responses, indicating abnormal sensory activity. Conversely, with Symptom_Motor at 0 (no motor symptoms), ULSSEP values are more clustered, showing fewer abnormalities. The presence of motor symptoms enhances the differentiation in ULSSEP results, highlighting a strong link between motor and sensory pathways.

4. Discussion

Predicting the progression of CIS to MS remains an extremely pressing issue. The use of ML models in clinical practice will help, together with clinical and radiological data, facilitate the early diagnosis of MS. Timely administration of therapy for this disease will prevent disability, maintain ability to work, and improve the quality of life of patients. In the future, these models can be integrated into diagnostic workflows to flag high-risk patients based on their clinical data and medical imaging results. They can also continuously analyze patient data to optimize treatment plans in real time, providing more responsive patient management. We believe increasing sample size and lengthening the duration of observation, coupled with the utilization of deep learning, and are key to further enhancing the predictive model. Adding more features such as MRI, serum, genetic biomarkers, and environmental factors can also provide unique insights into different aspects of CDMS progression. In addition, conducting longitudinal studies is essential to understand how CIS develops over time and distinguish between short variations and long-term trends of the disease process. This enables the development of treatment methods that tailor to different stages of CDMS.

While this study's findings are promising, there are several limitations to be acknowledged. One key limitation is the small dataset that causes the high risk of model overfitting, even when cross-validation is applied. This is particularly problematic when the dataset comes from a single location and is not representative of a diverse population. The retrospective nature of the data, which means the data is collected for purposes other than the specific research question at hand, also poses limitations. There may be inconsistencies in how data are recorded, and this can introduce noise to the models. Moreover, the analysis of features only reveals the magnitude of their importance since SHAP values are based on mean absolute values. As a result, it does not provide insights to whether the top influencing features increase or decrease the likelihood of CDMS.

5. Conclusion

The results of this study demonstrate improvements in early diagnosis accuracy and the potential of ML models in clinical integration. Specifically, our tree-based models achieve AUC scores above 0.9, with F1 scores higher than 82%, highlighting their effectiveness in predicting CDMS from CIS. We also identify key features that significantly contribute to predicting the progression of CIS to CDMS, including Periventricular_MRI, Infratentorial_MRI, Oligoclonal_Bands, Schooling, and Symptom_Motor. These features provide valuable insights into the factors most closely associated with MS progression.

Acknowledgments

None.

Funding

This work was supported by a grant from the Russian Science Foundation (RSF 23-15-00377).

Conflict of interest

The authors declare that they have no competing interests.

Authors contributions

Conceptualization: Bair N. Tuchinov
Formal analysis: Minh Sao Khue Luu
Investigation: Minh Sao Khue Luu
Methodology: Denis S. Korobko, Nadezhda A. Malkova
Project administration: Andrey A. Tulupov
Writing – original draft: Minh Sao Khue Luu, Anna I. Prokaeva
Writing – review & editing: Bair N. Tuchinov

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The data used in this study are accessible at https://data.mendeley.com/datasets/8wk5hjx7x2/1. The code used to implement the models and analyses in this study is available at https://github.com/luumsk/CIStoCDMS.git. This GitHub repository includes detailed documentation of the libraries utilized, with all codes necessary to reproduce the results, including data preparation, model

development, validation strategies, and explainability processes. All tables and figures in this article are original. Figure 1 was created using Microsoft PowerPoint, whereas other figures were produced using Matplotlib, Seaborn, and SHAP libraries.

Further disclosure

The writing of this article had been optimized with the aid of ChatGPT for language enhancement.

References

1. Compston A, Coles A. Multiple sclerosis. *Lancet*. 2008;372(9648):1502-1517.

   doi: 10.1016/S0140-6736(08)61620-7

2. Lassmann H. Multiple sclerosis pathology. *Cold Spring Harb Perspect Med*. 2018;8(3):a028936.

   doi: 10.1101/cshperspect.a028936

3. *Multiple Sclerosis International Federation. Atlas of MS*. 3rd ed.; 2020. Available from: https://www.msif.org/wp-content/uploads/2020/10/atlas-3rd-edition-epidemiology-report-en-updated-30-9-20.pdf [Last accessed on 2024 Jul 11].

4. Tozlu C, Olafson E, Jamison KW, *et al*. The sequence of regional structural disconnectivity due to multiple sclerosis lesions. *Brain Commun*. 2023;5(6):fcad332.

   doi: 10.1093/braincomms/fcad332

5. Solomon AJ, Arrambide G, Brownlee WJ, *et al*. Differential diagnosis of suspected multiple sclerosis: An updated consensus approach. *Lancet Neurol*. 2023;22(8):750-768.

   doi: 10.1016/S1474-4422(23)00148-5

6. Dickie DA, Shenkin SD, Anblagan D, *et al*. Whole brain magnetic resonance image atlases: A systematic review of existing atlases and caveats for use in population imaging. *Front Neuroinform*. 2017;11:1.

   doi: 10.3389/fninf.2017.00001

7. Wattjes MP, Ciccarelli O, Reich DS, *et al*. 2021 MAGNIMS-CMSC-NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis. *Lancet Neurol*. 2021;20(8):653-670.

   doi: 10.1016/S1474-4422(21)00095-8

8. Henriksson F, Fredrikson S, Masterman T, Jönsson B. Costs, quality of life and disease severity in multiple sclerosis: A cross-sectional study in Sweden. *Eur J Neurol*. 2001;8(1):27-35.

   doi: 10.1046/j.1468-1331.2001.00169.x

9. Paz-Zulueta M, Parás-Bravo P, Cantarero-Prieto D, Blázquez-Fernández C, Oterino-Durán A. A literature review of cost-of-illness studies on the economic burden of multiple sclerosis. *Mult Scler Relat Disord*. 2020;43:102162.

   doi: 10.1016/j.msard.2020.102162

10. Kobelt G, Thompson A, Berg J, *et al*. New insights into the burden and costs of multiple sclerosis in Europe. *Mult Scler*. 2017;23(8):1123-1136.

    doi: 10.1177/1352458517694432

11. Miller DH, Chard DT, Ciccarelli O. Clinically isolated syndromes. *Lancet Neurol*. 2012;11(2):157-169.

    doi: 10.1016/S1474-4422(11)70274-5

12. Polman CH, Reingold SC, Banwell B, *et al*. Diagnostic criteria for multiple sclerosis: 2010 revisions to the Mcdonald criteria. *Ann Neurol*. 2011;69(2):292-302.

    doi: 10.1002/ana.22366

13. Thompson AJ, Banwell BL, Barkhof F, *et al*. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol*. 2018;17(2):162-173.

    doi: 10.1016/S1474-4422(17)30470-2

14. McDonald WI, Compston A, Edan G, *et al*. Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis. *Ann Neurol*. 2001;50(1):121-127.

    doi: 10.1002/ana.1032

15. Janelle M. *Signs and Symptoms of Multiple Sclerosis* (*MS*). Available from: https://www.health.com/multiple-sclerosis-symptoms-8653986 [Last accessed on 2024 Aug 05].

16. Gómez-Melero S, Caballero-Villarraso J, Escribano BM, Galvao-Carmona A, Túnez I, Agüera-Morales E. Impact of cognitive impairment on quality of life in multiple sclerosis patients-a comprehensive review. *J Clin Med*. 2024;13(11):3321.

    doi: 10.3390/jcm13113321

17. Luchetti S, Fransen NL, van Eden CG, Ramaglia V, Mason M, Huitinga I. Progressive multiple sclerosis patients show substantial lesion activity that correlates with clinical disease severity and sex: A retrospective autopsy cohort analysis. *Acta Neuropathol*. 2018;135(4):511-528.

    doi: 10.1007/s00401-018-1818-y

18. Lublin FD, Reingold SC, Cohen JA, *et al*. Defining the clinical course of multiple sclerosis. *Neurology*. 2014;83(3):278-286.

    doi: 10.1212/WNL.0000000000000560

19. Giovannoni G, Butzkueven H, Dhib-Jalbut S, *et al*. Brain health: Time matters in multiple sclerosis. *Mult Scler Relat Disord*. 2016;9 Suppl: S5-S48.

    doi: 10.1016/j.msard.2016.07.003

20. Ro LS, Yang CC, Lyu RK, *et al*. A prospective, observational study on conversion of clinically isolated syndrome to multiple sclerosis during 4-year period (MS NEO study) in Taiwan. *PLoS One*. 2019;14(7):e0202453.

    doi: 10.1371/journal.pone.0202453

21. Piri Cinar B, Ozakbas S. Prediction of conversion from clinically isolated syndrome to multiple sclerosis according to baseline characteristics: A prospective study. *Noro Psikiyatr Ars*. 2018;55:15-21.

doi: 10.29399/npa.12667

22. Shaheen HA, Sayed SS, Daker LI, Taha MA. Early predictors of conversion in patients with clinically isolated syndrome: A preliminary Egyptian study. *Egypt J Neurol Psychiatr Neurosurg*. 2018;54(1):21.

doi: 10.1186/s41983-018-0021-3

23. Bi CF, Qian HR, Peng LJ, *et al*. The correlation factor analysis for conversion of clinically isolated syndrome to multiple sclerosis and neuromyelitis optica. *Zhonghua Nei Ke Za Zhi*. 2016;55(6):460-465.

doi: 10.3760/cma.j.issn.0578-1426.2016.06.012

24. Kuhle J, Disanto G, Dobson R, *et al*. Conversion from clinically isolated syndrome to multiple sclerosis: A large multicentre study. *Mult Scler*. 2015;21(8):1013-1024.

doi: 10.1177/1352458514568827

25. CHAMPS Study Group. MRI predictors of early conversion to clinically definite MS in the CHAMPS placebo group. *Neurology*. 2002;59(7):998-1005.

doi: 10.1212/WNL.59.7.998

26. Alroughani R, Al Hashel J, Lamdhade S, Ahmed SF. Predictors of conversion to multiple sclerosis in patients with clinical isolated syndrome using the 2010 revised McDonald criteria. *ISRN Neurol*. 2012;2012:792192.

doi: 10.5402/2012/792192

27. Kolčava J, Kočica J, Hulová M, *et al*. Conversion of clinically isolated syndrome to multiple sclerosis: A prospective study. *Mult Scler Relat Disord*. 2020;44:102262.

doi: 10.1016/j.msard.2020.102262

28. Zhang H, Alberts E, Pongratz V, *et al*. Predicting conversion from clinically isolated syndrome to multiple sclerosis-an imaging-based machine learning approach. *Neuroimage Clin*. 2019;21:101593.

doi: 10.1016/j.nicl.2018.11.003

29. Bendfeldt K, Taschler B, Gaetano L, *et al*. MRI-based prediction of conversion from clinically isolated syndrome to clinically definite multiple sclerosis using SVM and lesion geometry. *Brain Imaging Behav*. 2019;13(5):1361-1374.

doi: 10.1007/s11682-018-9942-9

30. Yoo Y, Tang LYW, Li DKB, *et al*. Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome. *Comput Methods Biomech Biomed Eng Imaging Vis*. 2019;7(3):250-259.

doi: 10.1080/21681163.2017.1356750

31. Banerjee T, Saha M, Ghosh E, *et al*. Conversion of clinically isolated syndrome to multiple sclerosis: A prospective multi-center study in Eastern India. *Mult Scler J Exp Transl Clin*. 2019;5(2):205521731984972.

doi: 10.1177/2055217319849721

32. Rommer PS, Milo R, Han MH, *et al*. Immunological aspects of approved MS therapeutics. *Front Immunol*. 2019;10:1564.

doi: 10.3389/fimmu.2019.01564

33. Pinto MF, Oliveira H, Batista S, *et al*. Prediction of disease progression and outcomes in multiple sclerosis with machine learning. *Sci Rep*. 2020;10(1):21038.

doi: 10.1038/s41598-020-78212-6

34. Zhao Y, Healy BC, Rotstein D, *et al*. Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PLoS One*. 2017;12(4):e0174866.

doi: 10.1371/journal.pone.0174866

35. Ion-Mărgineanu A, Kocevar G, Stamile C, *et al*. Machine learning approach for classifying multiple sclerosis courses by combining clinical data with lesion loads and magnetic resonance metabolic features. *Front Neurosci*. 2017;11:398.

doi: 10.3389/fnins.2017.00398

36. Wottschel V, Alexander DC, Kwok PP, *et al*. Predicting outcome in clinically isolated syndrome using machine learning. *Neuroimage Clin*. 2015;7:281-287.

doi: 10.1016/j.nicl.2014.11.021

37. Jasperse B, Barkhof F. *Machine Learning in Multiple Sclerosis*. United States: Humana Press Inc.; 2023. p. 899-919.

doi: 10.1007/978-1-0716-3195-9_28

38. Branco D, di Martino B, Esposito A, Tedeschi G, Bonavita S, Lavorgna L. Machine learning techniques for prediction of multiple sclerosis progression. *Soft Comput*. 2022;26(22):12041-12055.

doi: 10.1007/s00500-022-07503-z

39. Haouam KD, Benmalek M. Machine learning algorithms for early prediction of multiple sclerosis progression: A comparative study. *Adv Artif Intell Mach Learn*. 2024;04(01):2027-2051.

doi: 10.54364/AAIML.2024.41116

40. Vázquez-Marrufo M, Sarrias-Arrabal E, García-Torres M, Martín-Clemente R, Izquierdo G. A systematic review of the application of machine-learning algorithms in multiple sclerosis. *Neurología* (*Engl Ed*). 2023;38(8):577-590.

doi: 10.1016/j.nrleng.2020.10.013

41. Naji Y, Mahdaoui M, Klevor R, Kissani N. Artificial intelligence and multiple sclerosis: Up-to-date review. *Cureus*. 2023;15:e45412.

doi: 10.7759/cureus.45412

42. Patel MA, Villalobos F, Shan K, *et al*. Generative artificial intelligence versus clinicians: Who diagnoses multiple sclerosis faster and with greater accuracy? *Mult Scler Relat Disord*. 2024;90:105791.

doi: 10.1016/j.msard.2024.105791

43. Rasouli S, Dakkali MS, Azarbad R, *et al*. Predicting the conversion from clinically isolated syndrome to multiple sclerosis: An explainable machine learning approach. *Mult Scler Relat Disord*. 2024;86:105614.

doi: 10.1016/j.msard.2024.105614

44. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765-4774.

doi: 10.48550/arXiv.1705.07874

45. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. *CatBoost: Unbiased Boosting with Categorical Features*. arXiv [Preprint]; 2018.

doi: 10.48550/arXiv.1810.11363

46. Chen T, Guestrin C. XGBoost. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM; 2016. p. 785-794.

doi: 10.1145/2939672.2939785

47. Machado MR, Karray S, de Sousa IT. LightGBM: An Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry. In: *2019 14th International Conference on Computer Science & Education (ICCSE)*. IEEE; 2019. p. 1111-1116.

doi: 10.1109/ICCSE.2019.8845529

48. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.

doi: 10.1023/A:1010933404324

49. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273-297.

doi: 10.1007/BF00994018

50. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. United States: Wiley; 2000.

doi: 10.1002/0471722146

51. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York: Springer; 2009.

doi: 10.1007/978-0-387-84858-7

52. Chavarria V, Espinosa-Ramírez G, Sotelo J, *et al*. Conversion predictors of clinically isolated syndrome to multiple sclerosis in Mexican patients: A prospective study. *Arch Med Res*. 2023;54(5):102843.

doi: 10.1016/j.arcmed.2023.102843

53. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit*. 1997;30(7):1145-1159.

doi: 10.1016/S0031-3203(96)00142-2

54. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag*. 2009;45(4):427-437.

doi: 10.1016/j.ipm.2009.03.002

55. Shapiro SS, Wilk MB. An analysis of variance test for normality (Complete Samples). *Biometrika*. 1965;52(3/4):591.

doi: 10.2307/2333709

56. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc*. 1937;32(200):675-701.

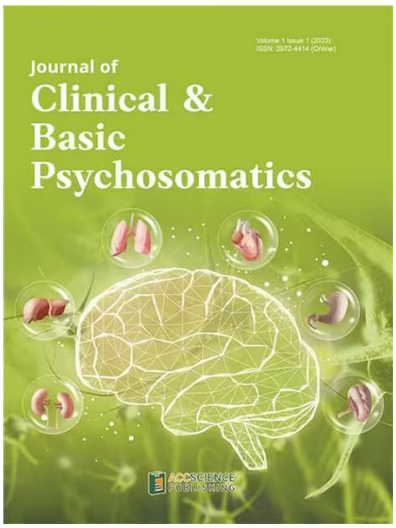
doi: 10.1080/01621459.1937.10503522

57. Nemenyi P. *Distribution-Free Multiple Comparisons*. Princeton: Princeton University; 1963.

58. Ali S, Akhlaq F, Imran AS, Kastrati Z, Daudpota SM, Moosa M. The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. *Comput Biol Med*. 2023;166:107555.

doi: 10.1016/j.compbiomed.2023.107555

Our Journals



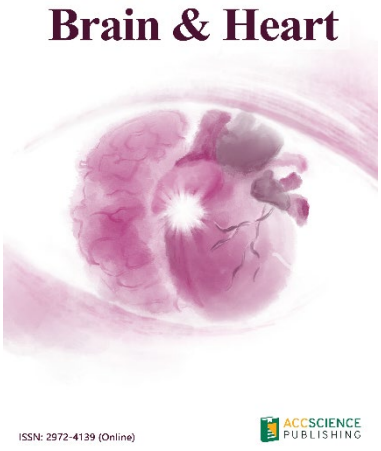
Journal of Clinical and Basic Psychosomatics (JCBP) is a quarterly journal focusing on clinical and basic research on symptoms, assessment, treatment, management, and the mechanism of psychosomatic disorders. *Journal of Clinical and Basic Psychosomatics* covers subject areas, including but not limited to the following:

- Conceptualization and classification of psychosomatic medicine
- Mechanism, biological markers, brain images, and treatment studies
- Psychosomatic reactions, syndromes, disorders, and diseases
- Psychosomatic disorders treated in general hospitals, including endocrinology, neurology, gastroenterology, dermatology, pain management, oncology, rheumatology, and other departments
- Psychological evaluation, management, rehabilitation, resilience training, and psychotherapy for general and specific populations during the pandemic
- Physiological disorders related to psychological factors (eating disorders, sleeping disorders, and sexual dysfunction)
- Somatic symptoms and related disorders and mental disorders due to somatic disease

Brain & Heart focuses on neurocardiology, a neurology and cardiology-based interdisciplinary subject that studies the circulatory mechanism of the human body, as well as the mechanisms of the interplay between the cardiovascular system and the nervous system. The journal's scope includes:

Clinical and basic research on diseases related to the circulatory and nervous systems, such as: orthostatic dizziness, orthostatic hypotension, autonomic dysfunction, and the relationship between the autonomic nervous system and the circulatory function in cerebral degeneration;

Heart-brain research on patients with syncope, autonomic dysfunction, cryptogenic stroke, and stroke with atrial fibrillation; research on the relationship between structural heart diseases and nervous system diseases, the correlation between cardiac electrophysiology and abnormal organizational structures and the pathogenesis of stroke, as well as new ways of diagnosis, treatment and prevention of unexplained stroke.



Start a new journal

Write to us via email if you are interested to start a new journal with AccScience Publishing. Please attach your CV, professional profile page and a brief pitch proposal in your email. We shall inform you of our decision whether we are interested to collaborate in starting a new journal.

Contact: info@accscience.com

https://www.accscience.com/journal/AIH

ACCSCIENCE PUBLISHING

Contact
www.accscience.com
8 Burn Road, #15-03 Trivex, Singapore 369977
Email: editorial@accscience.com
Phone: +65 8182 1586