Am J Cardiovasc Dis 2026;16(4):289-302
www.AJCD.us /ISSN:2160-200X/AJCD0167605

Original Article

Comparative performance of machine learning and deep learning models for heart disease prediction in a small clinical dataset

Erfan Barootchi[1*], Pardis Zamani[1*], Roshan Birgani[2*], Aida Bakhshi[3], Pariya Zibaei[4], Armin Tafazolimoghadam[5], Yasamin Moeinipour[6], MohammadMahdi Doosti[7], Niloofar Deravi[8], Yaser Khakpour[9], Zahra Farrokhi[8], Mobina Fathi[8], Danyal Yarahmadi[10,11]

[1]School of Medicine, Tehran University of Medical Sciences, Tehran, Iran; [2]Abadan University of Medical Sciences, Ābādān, Khuzestan, Iran; [3]Clinical Research Development Unit, Ghaem Hospital, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran; [4]School of Medicine, Dezful University of Medical Science, Dezful, Iran; [5]Tehran University of Medical Science (TUMS), School of Medicine, Tehran, Iran; [6]Cardiothoracic Surgery Department, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran; [7]Faculty of Pharmacy, Cyprus International University, Nicosia 99258, Cyprus; [8]Student Research Committee, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran; [9]Faculty of Medicine, Guilan University of Medical Sciences, Rasht, Iran; [10]School of Medicine, Iran University of Medical Sciences, Tehran, Iran; [11]Razi Herbal Medicines Research Center, School of Medicine, Lorestan University of Medical Sciences, Khorramabad, Iran. [*]Equal contributors and co-first authors.

Received July 21, 2025; Accepted July 8, 2026; Epub August 15, 2026; Published August 30, 2026

Abstract: Background: Machine learning (ML) and deep learning (DL) models are increasingly applied to cardiovascular risk prediction, yet their comparative performance in small tabular clinical datasets remains uncertain. This study compared conventional machine learning models and a deep neural network for predicting angiographically defined heart disease using the original UCI Heart Disease dataset. Methods: The dataset was loaded directly from the UCI Machine Learning Repository and included 303 patient records. The original multiclass angiographic disease variable was converted into a binary outcome indicating absence or presence of heart disease. Continuous predictors included age, resting blood pressure, serum cholesterol, maximum heart rate, and exercise-induced ST depression. Categorical predictors were defined according to UCI attribute documentation and one-hot encoded. The dataset was divided using a stratified 70/15/15 train-validation-test split. Logistic regression, decision tree, random forest, XGBoost, and deep neural network models were trained. Machine learning hyperparameters were selected using 5-fold stratified cross-validation with AUC optimization. Classification thresholds were selected on the validation set by maximizing F1-score and then applied unchanged to the independent test set. Performance was assessed using accuracy, precision, recall, F1-score, and AUC with bootstrap 95% confidence intervals. Pairwise comparisons used McNemar's test and paired bootstrap AUC testing with Holm correction. Results: The deep neural network achieved the highest test-set accuracy, precision, and F1-score, each reaching 91.30%, 90.48%, and 90.48%, respectively, with an AUC of 0.958. Logistic regression showed the highest recall, 95.24%, and AUC, 0.968. XGBoost achieved 71.74% accuracy and 0.825 AUC. Conclusion: Deep learning and regularized logistic regression showed the strongest performance, while XGBoost did not demonstrate superiority in this small clinical dataset.

Keywords: Heart disease prediction, machine learning, deep learning, XGBoost, random forest, logistic regression, small dataset

Introduction

Cardiovascular diseases (CVDs) remain among the most prevalent causes of death in the United States and Europe, accounting for more than 17 million deaths worldwide each year according to the World Health Organization (WHO) [1-3]. Beyond their effect on survival and quality of life (QOL), CVDs impose a substantial economic burden on patients, families, and healthcare systems [4]. Early diagnosis and risk assessment of heart diseases will assist

https://doi.org/10.62347/ICWV2651

healthcare professionals in providing more proper treatment approaches to the patients, thus diminishing the mortality rate and the healthcare system's burden [5].

Heart disease is influenced by several clinical and behavioral risk factors, including hypercholesterolemia, hypertension, smoking, family history, diabetes, obesity, alcohol use, and physical inactivity. These factors are commonly categorized as modifiable and non-modifiable risk factors [6, 7]. Modifiable risk factors such as smoking; high blood pressure; high cholesterol levels; overweight or obese; and lack of exercise are factors that can be changed with the help of your physician and/or changes in lifestyle [8]. Non-modifiable risk factors are those characteristics of the patient that cannot be changed. Recognizing these two categories of risk factors is essential to create appropriate interventions aimed at decreasing the incidence of CVD through effective preventive strategies or optimal treatment of those patients who already have CVD established [9]. In contrast, non-modifiable risk factors function primarily as predictors for the baseline risk categorization of a patient and projections for the long-term outcome of his/her CVD risk [10]. Because these factors often interact in complex and non-linear ways, accurate heart disease prediction requires multivariable approaches that can integrate clinical and diagnostic information efficiently [5].

Artificial intelligence (AI) is being utilized extensively nowadays in every domain of expertise. Machine Learning (ML) and Deep Learning (DL) are two AI subsets that enabled us to present a strong tool in medical domains, particularly in the prediction of various diseases like CVD [11-14]. These methods can identify patterns within clinical data and may support decision-making when used as adjuncts to conventional clinical evaluation. Therefore, carefully developed intelligent prediction systems may contribute to earlier recognition of high-risk patients and more efficient diagnostic assessment [12].

Over the last few years, owing to the significant role of timely heart disease prediction in public health, a number of ML and DL models have been proposed and utilized in assisting physicians in the early prediction of CVD by combining large medical data with high-level algorithms. ML models including Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), Logistic Regression (LR), Adaptive Boosting (AB), Naive Bayes (NB), and Extra Tree (ET), DL models including Deep Neural Network (DNN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), autoencoder, Feed Forward Neural Network (FNN), ResNet, InceptionNet, DenseNet, and hybrid model, and ultimately ensemble models which are proposed to provide a more robust and enhanced prediction than any one model by combining weak and strong models [7, 11, 15]. Extreme Gradient Boosting, or XGBoost, is also a scalable ensemble ML model inspired by boosted decision trees [16, 17].

Nonetheless, much of the current research has been conducted on large-scale or augmented datasets, where complex architectures may benefit from high-dimensional information. In many clinical settings, particularly in routine cardiology and small institutional datasets, only a limited amount of structured clinical data is available. Under these data-constrained conditions, risk stratification using modifiable and non-modifiable factors must be performed carefully [18, 19]. Consequently, it remains uncertain whether classical ML models, ensemble methods, or DL models provide the most reliable prediction of heart disease in small tabular clinical datasets.

Therefore, this study aimed to compare the predictive performance of conventional machine learning models and a deep neural network for angiographically defined heart disease using the original UCI Heart Disease dataset. Specifically, we evaluated logistic regression, decision tree, random forest, XGBoost, and deep neural network models under a unified validation framework. By incorporating stratified train-validation-test splitting, validation-based threshold selection, bootstrap confidence intervals, ROC/AUC analysis, and formal paired model comparisons, we sought to determine which approaches remain most reliable when clinical data are structured, tabular, and limited in size.

Methodology

Data source and study population

Our data was collected from the publicly available UCI Heart Disease dataset that has 303 patients' records of individuals who were referred for coronary artery disease assessment and diagnosis. The data includes demographic (age and sex) and cardiovascular risk factors (blood pressure, fasting blood glucose, serum cholesterol), plus results of non-invasive exercise test variables (resting electrocardiogram, maximum heart rate achieved during exercise, exercise-induced angina, and ST-segment depression), representing a standard clinical group of patients undergoing a diagnostic assessment to evaluate if they have heart disease [20].

Inclusion and exclusion criteria

All patient records available in the original UCI Heart Disease dataset were eligible for analysis because each contained the angiographic outcome variable required to define heart disease status. No record was excluded because of missing outcome data, and duplicate observations were not identified. Records with missing predictor values were retained to preserve the original dataset structure and avoid unnecessary loss of information; these values were handled during model preprocessing through prespecified imputation. The final analytical cohort therefore included 303 patients. Eligible predictors were limited to the 13 UCI clinical and diagnostic attributes used in the standard heart disease prediction task, and the outcome was binarized as absence versus presence of angiographic disease.

Outcome definition and grouping

The primary outcome was the presence of heart disease, defined according to the angiographic findings reported in the UCI dataset. We treated the outcome as a binary variable, where patients either had heart disease (1) or did not (0), and all models were trained to distinguish between these two groups.

Dataset processing and covariates

The dataset was loaded directly from the UCI Machine Learning Repository, and abbreviated UCI variable names were converted into clinically readable labels before analysis. The original angiographic disease severity variable was preserved for documentation and then transformed into a binary outcome: no heart disease for severity level 0 and heart disease presence for severity levels 1 through 4. Predictors were classified according to the UCI attribute definitions. Age, resting blood pressure, serum cholesterol, maximum heart rate achieved, and exercise-induced ST depression were treated as continuous variables. Sex, chest pain type, fasting blood sugar status, resting electrocardiographic findings, exercise-induced angina, ST-segment slope, number of major vessels colored by fluoroscopy, and thalassemia status were treated as categorical variables. Continuous predictors were converted to numeric format, imputed using the median, and standardized with Z-score scaling. Categorical predictors were preserved with human-readable labels, imputed using the most frequent category, and one-hot encoded for model development.

Model development

Five predictive models were developed under a unified analytical framework: logistic regression, decision tree, random forest, XGBoost, and a deep neural network. The dataset was divided into training, validation, and test sets using a stratified 70/15/15 split to maintain the proportion of patients with and without heart disease across all partitions. Machine learning models were implemented as preprocessing-model pipelines, ensuring that imputation, scaling, and one-hot encoding were learned only from the training data and then applied to validation and test data. Hyperparameters for logistic regression, decision tree, random forest, and XGBoost were selected using grid search with 5-fold stratified cross-validation in the training set, with AUC used as the optimization criterion. The deep neural network included dense layers with ReLU activation, dropout, L2 regularization, class weighting, and early stopping based on validation loss. For all models, classification thresholds were selected on the validation set by maximizing F1-score and were then applied unchanged to the independent test set.

Statistical analysis and performance evaluation

Statistical analysis and model evaluation were performed using Python version 3.13.7 using

pandas, numpy, scipy, scikit-learn, statmodels, matplotlib, ucimlrepo, tensorflow, and *xgboost* packages. Continuous variables were summarized using mean, standard deviation, median, interquartile range, and observed range, whereas categorical variables were summarized as frequencies and percentages. The dataset was loaded directly from the UCI Machine Learning Repository using the ucimlrepo Python package. Based on the UCI Heart Disease attribute definitions, age, resting blood pressure, serum cholesterol, maximum heart rate, and exercise-induced ST depression were treated as continuous variables. Sex, chest pain type, fasting blood sugar status, resting electrocardiographic results, exercise-induced angina, ST-segment slope, number of major vessels colored by fluoroscopy, and thalassemia status were treated as categorical variables and one-hot encoded.

The dataset was divided into training, validation, and test sets using a stratified 70/15/15 split to preserve the distribution of heart disease status. Numerical predictors were imputed using the median and standardized using Z-score scaling. Categorical predictors were imputed using the most frequent category and one-hot encoded. Logistic regression, decision tree, random forest, XGBoost, and deep neural network models were trained and evaluated. Hyperparameters for machine learning models were selected using grid search with 5-fold stratified cross-validation within the training set, using AUC as the optimization metric. Classification thresholds were selected using the validation set only according to the F1 criterion and then applied unchanged to the independent test set.

Final model performance was assessed on the independent test set using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve. Ninety-five percent confidence intervals for performance metrics were calculated using bootstrap resampling of the test set. Receiver operating characteristic curves were generated using predicted probabilities. Formal paired comparisons between models were performed on the same test-set observations. McNemar's test was used to compare paired classification correctness between models. Paired bootstrap testing was used to compare AUC values between models. Holm adjustment was applied for multiple pairwise comparisons. A two-sided *P*-value less than 0.05 was considered statistically significant.

Results

Study population and outcome distribution

The analysis included 303 patient records from the original UCI Heart Disease dataset. No duplicate records were identified, and all patients had an available angiographic disease outcome. The study population was clinically heterogeneous, reflecting a typical diagnostic cohort evaluated for suspected coronary artery disease. The mean age was 54.44 years (SD, 9.04), and 206 participants (67.99%) were male. Mean resting blood pressure was 131.69 mmHg (SD, 17.60), mean serum cholesterol was 246.69 mg/dL (SD, 51.78), mean maximum heart rate was 149.61 beats per minute (SD, 22.88), and mean exercise-induced ST depression was 1.04 (SD, 1.16). The most frequent chest pain category was asymptomatic chest pain, observed in 144 patients (47.52%). In the binary outcome definition, 164 patients (54.13%) had no heart disease and 139 patients (45.87%) had heart disease. Baseline characteristics and categorical predictor distributions are presented in **Table 1**.

Data partitioning and model optimization

The dataset was divided using a stratified 70/15/15 train-validation-test split. This produced 212 patients in the training set, 45 in the validation set, and 46 in the independent test set. The proportion of patients with heart disease was well preserved across the three subsets: 45.75% in the training set, 46.67% in the validation set, and 45.65% in the test set, confirming appropriate stratification (Table S1).

During model development, hyperparameters for the machine learning models were selected using 5-fold stratified cross-validation within the training set. Logistic regression achieved the highest cross-validated AUC among the machine learning models (0.8949), followed closely by random forest (0.8923) and XGBoost (0.8891). Decision tree showed a lower cross-validated AUC of 0.8579. Validation-selected thresholds were then identified by maximizing F1-score and applied unchanged to the independent test set. The selected thresholds were

Table 1. Baseline characteristics

Variable Sample Size =303	Level	Mean, N*	SD, Percentage*
Continuous Variables			
Age		54.44	9.04
BP		131.69	17.60
Cholesterol		246.69	51.78
Max HR		149.61	22.87
ST depression		1.04	1.16
Categorical Variables			
Heart Disease Status	Absence	164	54.13
	Presence	139	45.87
Sex	Female	97	32.01
	Male	206	67.99
Chest pain type	Asymptomatic	144	47.52
	Atypical angina	50	16.50
	Non-anginal pain	86	28.38
	Typical angina	23	7.59
FBS	Fasting blood sugar ≤ 120 mg/dl	258	85.15
	Fasting blood sugar > 120 mg/dl	45	14.85
EKG results	Left ventricular hypertrophy	148	48.84
	Normal	151	49.83
	ST-T wave abnormality	4	1.32
Exercise angina	Exercise-induced angina	99	32.67
	No exercise-induced angina	204	67.33
Slope of ST	Downsloping	21	6.93
	Flat	140	46.20
	Upsloping	142	46.86
Number of vessels	0 major vessels	176	58.09
	1 major vessel	65	21.45
	2 major vessels	38	12.54
	3 major vessels	20	6.60
	None	4	1.32
Thallium	Fixed defect	18	5.94
	Normal	166	54.79
	Reversible defect	117	38.61
	None	2	0.66

*Continuous variables are presented with mean and SD and categorical Variables are presented with number (N) and Percentage.

0.3338 for logistic regression, 0.3488 for decision tree, 0.5300 for random forest, 0.3612 for XGBoost, and 0.5975 for the deep neural network. Hyperparameter tuning results and validation-selected thresholds are summarized in **Table 2**.

Independent test-set performance

Model performance was evaluated on the independent test set of 46 patients, including 21 patients with heart disease and 25 without heart disease. The deep neural network achieved the highest overall classification performance, with an accuracy of 91.30% (95% CI, 82.61-97.83), precision of 90.48% (95% CI, 75.00-100.00), recall of 90.48% (95% CI, 76.16-100.00), F1-score of 90.48% (95% CI, 79.16-98.04), and AUC of 0.958 (95% CI, 0.892-1.000). Logistic regression showed the highest discriminative performance by AUC, reaching 0.968 (95% CI, 0.911-1.000), and

Table 2. Hyperparameter tuning and validation selected thresholds

Model	Best cross-validated AUC	Validation-selected threshold	Threshold selection rule
Logistic Regression	0.8949	0.3338	Validation-selected threshold maximizing F1-score; validation F1=0.8444
Decision Tree	0.8579	0.3488	Validation-selected threshold maximizing F1-score; validation F1=0.7556
Random Forest	0.8923	0.5300	Validation-selected threshold maximizing F1-score; validation F1=0.8085
XGBoost	0.8891	0.3612	Validation-selected threshold maximizing F1-score; validation F1=0.7692
Deep Neural Network	Validation-based training only	0.5975	Validation-selected threshold maximizing F1-score; validation F1=0.8372

Table 3. Test performance and ROC/AUC values

Model	Threshold	Accuracy, % (95% CI)	Precision, % (95% CI)	Recall, % (95% CI)	F1-score, % (95% CI)	AUC (95% CI)
Logistic Regression	0.3338	80.43 (67.39-91.30)	71.43 (53.85-87.50)	95.24 (84.62-100.00)	81.63 (68.08-91.80)	0.968 (0.911-1.000)
Decision Tree	0.3488	67.39 (54.35-80.43)	59.38 (41.94-76.67)	90.48 (76.19-100.00)	71.70 (55.81-84.38)	0.812 (0.672-0.933)
Random Forest	0.5300	84.78 (73.91-95.65)	79.17 (61.90-95.24)	90.48 (76.16-100.00)	84.44 (70.97-94.74)	0.927 (0.838-0.989)
XGBoost	0.3612	71.74 (58.70-84.78)	62.50 (45.15-78.79)	95.24 (84.62-100.00)	75.47 (60.47-87.10)	0.825 (0.687-0.931)
Deep Neural Network	0.5975	91.30 (82.61-97.83)	90.48 (75.00-100.00)	90.48 (76.16-100.00)	90.48 (79.16-98.04)	0.958 (0.892-1.000)

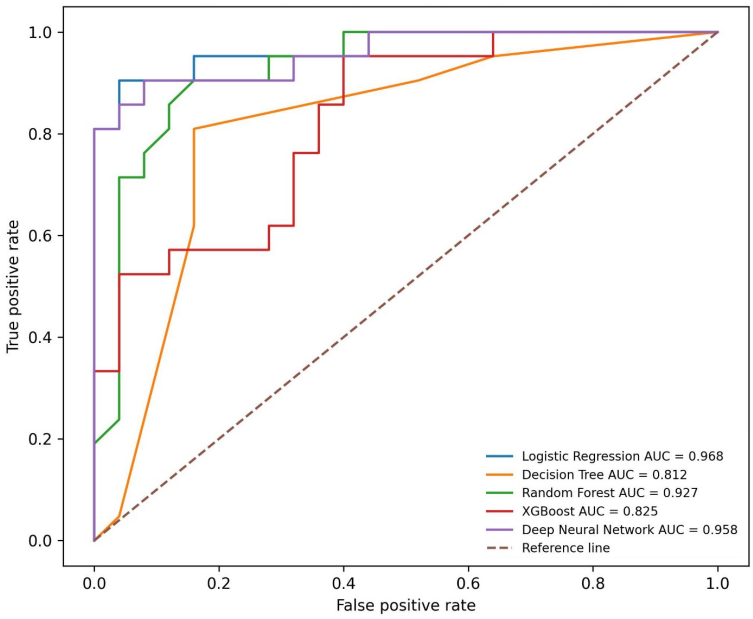

Figure 1. ROC curves for heart disease prediction models.

also achieved the highest recall, 95.24% (95% CI, 84.62-100.00). Its overall accuracy was 80.43% (95% CI, 67.39-91.30).

Random forest also performed strongly, with an accuracy of 84.78% (95% CI, 73.91-95.65), recall of 90.48% (95% CI, 76.16-100.00), F1-score of 84.44% (95% CI, 70.97-94.74), and AUC of 0.927 (95% CI, 0.838-0.989). XGBoost did not show the expected superior performance. Although it achieved high recall, 95.24% (95% CI, 84.62-100.00), its accuracy was 71.74% (95% CI, 58.70-84.78), precision was 62.50% (95% CI, 45.15-78.79), F1-score was 75.47% (95% CI, 60.47-87.10), and AUC was 0.825 (95% CI, 0.687-0.931). Decision tree had the lowest overall accuracy, 67.39% (95% CI, 54.35-80.43), and an AUC of 0.812 (95% CI, 0.672-0.933). Full test-set performance metrics are reported in **Table 3**, and ROC curves are shown in **Figure 1**.

Class-specific performance and confusion matrix findings

Class-specific evaluation provided additional insight into model behavior. The deep neural network showed the most balanced classification profile, with precision, recall, and F1-score of 0.92 for the absence class and 0.9048 for the presence class. Random forest also showed balanced performance, although its presence-class precision was lower than that of the deep neural network. Logistic regression and XGBoost were highly sensitive for detecting heart disease, each identifying 20 of 21 disease cases, but both produced more false positives than the deep neural network. Logistic regression generated 8 false positives and 1 false negative, while XGBoost generated 12 false positives and 1 false negative. In contrast, the deep neural network produced only 2 false positives and 2 false negatives, indicating a better balance between sensitivity and specificity. Class-specific metrics and confusion matrix components are presented in **Tables 4**, **5**, respectively.

Formal model comparisons

Formal paired comparisons demonstrated that differences between models varied according to the evaluation metric. For classification accuracy, the deep neural network significantly outperformed the decision tree after Holm correction (adjusted McNemar P=0.0342). The comparison between the deep neural network and XGBoost favored the deep neural network numerically, but the Holm-adjusted McNemar *P*-value did not remain below 0.05. In AUC comparisons, logistic regression significantly outperformed decision tree and XGBoost after adjustment. The deep neural network also showed significantly higher AUC than decision tree and XGBoost, while random forest showed significantly higher AUC than XGBoost. These findings indicate that XGBoost was not statistically superior to the competing models and, in several AUC-based comparisons, performed significantly worse than logistic regression, random forest, and the deep neural network (**Table 6**).

Feature importance

Model-based feature importance was examined for XGBoost and random forest. In XGBoost, thallium status was the dominant predictor, with “normal” and “reversible defect” categories showing the highest importance

Table 4. Class-specific classification report for each model

Model	Class	Precision	Recall	F1-score	Support
Logistic Regression	Absence	0.9444	0.6800	0.7907	25
	Presence	0.7143	0.9524	0.8163	21
Decision Tree	Absence	0.8571	0.4800	0.6154	25
	Presence	0.5938	0.9048	0.7170	21
Random Forest	Absence	0.9091	0.8000	0.8511	25
	Presence	0.7917	0.9048	0.8444	21
XGBoost	Absence	0.9286	0.5200	0.6667	25
	Presence	0.6250	0.9524	0.7547	21
Deep Neural Network	Absence	0.9200	0.9200	0.9200	25
	Presence	0.9048	0.9048	0.9048	21

This table presents class-specific precision, recall, F1-score, and support for absence and presence of heart disease.

Table 5. Confusion metrics

Model	True negative	False positive	False negative	True positive
Logistic Regression	17	8	1	20
Decision Tree	12	13	2	19
Random Forest	20	5	2	19
XGBoost	13	12	1	20
Deep Neural Network	23	2	2	19

values. In random forest, maximum heart rate, ST depression, age, thallium status, blood pressure, and cholesterol contributed most strongly to classification. These patterns suggest that both exercise-related variables and perfusion-related diagnostic indicators were central to prediction. The complete feature-importance rankings are provided in Table S2, and the visual comparison is presented in **Figure 2**.

Overall, the results showed that the deep neural network achieved the strongest balanced test-set performance, while logistic regression provided the highest sensitivity and AUC. Random forest also performed well. In contrast, XGBoost did not demonstrate superiority and was outperformed in several AUC-based comparisons.

Discussion

Summary of findings

In this comparative analysis of the original UCI Heart Disease dataset, model performance differed meaningfully across algorithms, and the findings did not support the initial assumption of XGBoost superiority. The deep neural network achieved the strongest balanced test-set performance, with 91.30% accuracy, 90.48% precision, 90.48% recall, and an F1-score of 90.48%. Logistic regression showed the highest discrimination and sensitivity, with an AUC of 0.968 and recall of 95.24%, suggesting strong clinical utility when minimizing missed heart disease cases is prioritized. Random forest also performed well, reaching 84.78% accuracy and an AUC of 0.927. In contrast, XGBoost achieved high recall but lower overall performance, with 71.74% accuracy and an AUC of 0.825. Formal paired comparisons reinforced this pattern: XGBoost was not statistically superior to the competing models and showed lower AUC than logistic regression, random forest, and the deep neural network. Overall, the results favor a more nuanced interpretation of model performance in small tabular clinical datasets.

Comparison with previous research

Previous studies have reported strong performance for both machine learning and deep learning models in cardiovascular prediction, although the leading algorithm has varied substantially across datasets, feature spaces, and optimization strategies. Several recent works have emphasized that complex preprocessing and optimization pipelines can markedly improve model performance. Al-Alshaikh et al. proposed a heart disease prediction framework combining genetic algorithm-based feature selection, recursive feature elimination, class-imbalance handling, and an optimized deep convolutional neural network, reporting high accuracy and sensitivity [21]. Similarly, Xia et al. developed an ant colony optimization-enhanced deep learning model with Bayesian hyperparameter tuning, highlighting the value of automated feature selection and parameter optimization in cardiovascular diagnosis [22]. Cao et al. focused specifically on XGBoost and

Table 6. Model comparisons (Exact McNemar Test)

Model A	Model B	Accuracy A	Accuracy B	McNemar *P*-value	Discordant pairs	AUC A	AUC B	AUC difference A-B	AUC bootstrap *P*-value	McNemar Holm-adjusted *P*-value	AUC Holm-adjusted *P*-value
Logistic Regression	Decision Tree	0.8043	0.6739	0.1796	14	0.9676	0.8124	0.1552	0.003	0.8978	0.021
Logistic Regression	Random Forest	0.8043	0.8478	0.6875	6	0.9676	0.9267	0.0410	0.090	1.0000	0.360
Logistic Regression	XGBoost	0.8043	0.7174	0.2188	6	0.9676	0.8248	0.1429	< 0.001	0.8978	< 0.001
Logistic Regression	Deep Neural Network	0.8043	0.9130	0.1250	7	0.9676	0.9581	0.0095	0.394	0.7500	0.788
Decision Tree	Random Forest	0.6739	0.8478	0.0574	14	0.8124	0.9267	-0.1143	0.033	0.4590	0.165
Decision Tree	XGBoost	0.6739	0.7174	0.7539	10	0.8124	0.8248	-0.0124	0.785	1.0000	0.788
Decision Tree	Deep Neural Network	0.6739	0.9130	0.0034	13	0.8124	0.9581	-0.1457	0.004	0.0342	0.024
Random Forest	XGBoost	0.8478	0.7174	0.0703	8	0.9267	0.8248	0.1019	0.002	0.4922	0.016
Random Forest	Deep Neural Network	0.8478	0.9130	0.2500	3	0.9267	0.9581	-0.0314	0.250	0.8978	0.750
XGBoost	Deep Neural Network	0.7174	0.9130	0.0117	11	0.8248	0.9581	-0.1333	0.000	0.1055	0.000

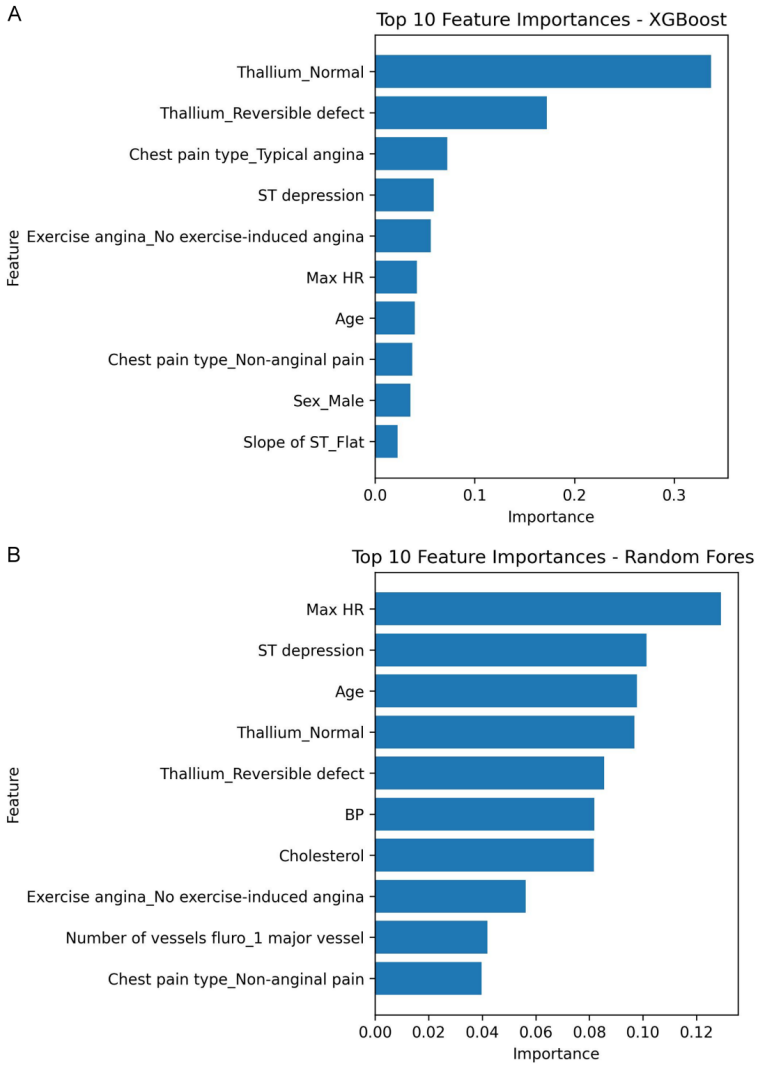

Figure 2. XGBOOST and random forest feature importances. A: XGBOOST; B: Random Forest.

showed that multiple feature selection combined with improved particle swarm optimization increased the performance of standard XGBoost, supporting the idea that boosting models are highly dependent on feature selection and tuning procedures [23].

Other studies have favored deep or hybrid architectures. Darolia et al. reported high predictive performance using a hybrid long short-term memory and quantum neural network model with optimized feature selection [24]. Bhagawati et al. found that deep learning outperformed conventional machine learning when carotid plaque and imaging-derived features were used for coronary artery disease and cardiovascular event prediction [25]. Deng et al. also showed that neural network survival models achieved similar or superior discrimination compared with pooled cohort equations for ASCVD risk prediction [26]. In line with this pattern, Meng et al. reported that neural networks provided the highest sensitivity and AUC for ASCVD detection in NHANES data, outperforming random forest and XGBoost [27].

However, the literature is not uniform. Ejiyi et al. found that XGBoost achieved higher accuracy than their proposed CardioVitalNet model on a UCI-based cardiovascular dataset [28], whereas Sadr et al. reported that an ensemble combining machine learning and deep learning models achieved the best overall performance across multiple datasets [29]. Our findings are consistent with this broader heterogeneity. In the original UCI Heart Disease dataset, the deep neural network showed the strongest balanced classification performance, logistic regression achieved the highest AUC and sensitivity, and XGBoost did not demonstrate superiority. These results suggest that model ranking in small tabular clinical datasets is context-dependent and should be judged through transparent validation rather than algorithmic expectation alone.

Strengths and limitations

This study has several strengths. First, the analysis used the original UCI Heart Disease dataset loaded directly from the UCI Machine Learning Repository, reducing uncertainty related to altered secondary copies of the dataset. Second, all models were evaluated under a unified and reproducible framework, including clinically defined covariate handling, stratified

train-validation-test splitting, validation-based threshold selection, and independent test-set evaluation. Third, the study moved beyond simple point estimates by reporting bootstrap confidence intervals, ROC/AUC values, confusion matrices, and formal paired comparisons using McNemar's test and paired bootstrap AUC testing with correction for multiple comparisons. This approach provides a more transparent assessment of model performance than isolated accuracy reporting.

Several limitations should also be acknowledged. The sample size was small, and the independent test set included only 46 patients, which widened confidence intervals and limited the stability of pairwise comparisons. The dataset is also historical and lacks contemporary clinical, laboratory, imaging, therapeutic, and follow-up information. External validation was not performed, so generalizability to broader populations remains uncertain. Finally, although feature importance was examined for tree-based models, formal explainability and calibration analyses were not fully explored.

Clinical implications

The present findings have several clinical implications. First, they support the role of predictive modeling as an adjunct to clinical risk assessment rather than a replacement for physician judgment. In vascular medicine, machine learning has already shown value for outcome prediction; for example, Heo et al. reported that a deep neural network improved long-term outcome prediction after ischemic stroke compared with an established prognostic score [30]. Similarly, Holm et al. developed an externally validated neural-network survival model for ischemic heart disease, emphasizing that richer clinical and EHR-derived data may improve individualized post-angiography prognostication [31]. These studies suggest that algorithmic models may be most useful when they add structured, patient-specific risk information to existing clinical reasoning.

Second, clinical usefulness depends on more than discrimination. Oikonomou and Khera emphasized that predictive models should be assessed through discrimination, calibration, and net clinical benefit, while also considering interpretability, bias, and regulatory safety [32]. Luo et al. illustrated this translational pathway by combining external validation, calibration, decision-curve analysis, SHAP-based interpretation, and an online calculator for ischemic stroke mortality prediction [33]. This is highly relevant to cardiovascular prediction, where a model with good AUC may still be unsuitable if it is poorly calibrated or difficult to interpret.

Third, implementation must consider accessibility and equity. Ordikhani et al. developed an interpretable CVD risk model designed to balance accuracy with clinical acceptance [34]. Wang et al. showed that routine blood and biochemical data could support cost-effective cardiovascular screening, which may be particularly valuable in resource-limited settings [35]. Non-invasive and scalable approaches are also emerging: retinal deep-learning biomarkers have shown associations with cardiovascular morbidity and mortality [36], while CNN-based ECG interpretation has achieved performance comparable to clinical standards with explainability techniques highlighting relevant ECG segments [37]. Mobile health applications may further extend risk assessment beyond traditional clinical settings, although prospective validation remains essential [38].

Finally, fairness must be evaluated before deployment. Li et al. showed that both conventional and machine-learning CVD risk models may demonstrate demographic bias, particularly across gender groups [39]. In our study, the deep neural network achieved the strongest balanced classification performance, and logistic regression showed the highest AUC and sensitivity, whereas XGBoost was not superior. Clinically, this reinforces a cautious message: model selection should be evidence-based, transparent, and validated, with special attention to calibration, explainability, fairness, and external generalizability.

Future research

Future research should validate these findings in larger, contemporary, and geographically diverse cardiovascular cohorts. External validation is particularly important because the current analysis was based on a small historical dataset. Future studies should also assess calibration, decision-curve performance, subgroup fairness, and model explainability using

methods such as SHAP or LIME. Because logistic regression and the deep neural network showed strong but different performance profiles, subsequent work should examine whether combining interpretable statistical models with neural approaches improves clinical usefulness. Prospective studies are ultimately needed to determine whether these tools improve diagnostic efficiency, patient outcomes, and clinician decision-making in real-world practice.

Conclusion

In this comparative analysis of the original UCI Heart Disease dataset, deep neural network and logistic regression models demonstrated the strongest overall performance for predicting angiographically defined heart disease. The deep neural network provided the most balanced classification profile, while logistic regression achieved the highest sensitivity and AUC. Random forest also showed robust performance. In contrast, XGBoost did not demonstrate superiority and performed less favorably in several AUC-based comparisons. These findings challenge the assumption that more complex ensemble methods necessarily outperform simpler or neural models in small tabular clinical datasets. More importantly, they highlight the need for transparent validation, confidence interval reporting, and formal paired model comparisons when evaluating predictive algorithms for clinical decision support. External validation in larger and more diverse cohorts remains essential.

Acknowledgements

The authors would like to express their sincere gratitude to the researchers whose invaluable work contributed to this study. Their dedicated efforts in developing and performing this research were essential for the completion of this article, which was conducted without external funding.

Disclosure of conflict of interest

None.

AI generative statement

The authors declare that Grammarly and ChatGPT were used during the preparation of the manuscript, only to improve grammar, clarity, and fluency of the language. No AI generative tools were used to generate, analyze, or interpret data. All content was written, and critically reviewed by the authors.

Author contributions

E.B., P.Z. (Pardis Zamani), Y.Kh., and D.Y. conceptualized the study and drafted the manuscript. R.B., A.B., P.Z. (Pariya Zibaei), and A.T. conducted the literature review and drafted the Discussion section. M.F. and Z.F. curated and analyzed the dataset. N.D. and M.D. critically appraised the manuscript and supervised the study. E.B. and Y.M. prepared the final draft. All authors confirmed the final draft.

Abbreviations

AB, Adaptive Boosting; AI, Artificial Intelligence; AUC-ROC, Area Under the Receiver Operating Characteristic curve; BP, Blood Pressure; Chol, Cholesterol; CNN, Convolutional Neural Network; CVD, Cardiovascular Diseases; DL, Deep Learning; DNN, Deep Neural Network; DT, Decision Tree; ECG, Electrocardiogram; EDA, Exploratory Data Analysis; ET, Extra Tree; FBS, Fasting Blood Sugar; FNN, Feed Forward Neural Network; HPC, High-Performance Computing; IQR, Interquartile Range; KNN, K-Nearest Neighbors; LR, Logistic Regression; LSTM, Long Short Term Memory; MHR, Maximum Heart Rate; ML, Machine Learning; NB, Naive Bayes; QOL, Quality of Life; RAM, Random Access Memory; RF, Random Forest; RNN, Recurrent Neural Network; ROC, Receiver Operating Characteristic; SVM, Support Vector Machine; WHO, World Health Organization; XGBoost, Extreme Gradient Boosting.

Address correspondence to: Danyal Yarahmadi, School of Medicine, Iran University of Medical Sciences (IUMS), Hemat Highway, next to Milad Tower, Tehran 14496-14535, Iran. ORCID: 0009-0003-4509-2447; E-mail: danial78y78@gmail.com; Zahra Farrokhi and Mobina Fathi, School of Medicine, Shahid Beheshti University of Medical Sciences, Arabi Ave, Daneshjoo Blvd, Velenjak, Tehran 19839-63113, Iran. E-mail: Zahra-Farrokhi@sbmu.ac.ir (ZF); mobina.fathi78@gmail.com (MF)

References

[1] Townsend N, Kazakiewicz D, Lucy Wright F, Timmis A, Huculeci R, Torbica A, Gale CP, Achenbach S, Weidinger F and Vardas P. Epide-

miology of cardiovascular disease in Europe. Nat Rev Cardiol 2022; 19: 133-143.
[2] Benjamin EJ, Virani SS, Callaway CW, Chamberlain AM, Chang AR, Cheng S, Chiuve SE, Cushman M, Delling FN, Deo R, de Ferranti SD, Ferguson JF, Fornage M, Gillespie C, Isasi CR, Jiménez MC, Jordan LC, Judd SE, Lackland D, Lichtman JH, Lisabeth L, Liu S, Longenecker CT, Lutsey PL, Mackey JS, Matchar DB, Matsushita K, Mussolino ME, Nasir K, O'Flaherty M, Palaniappan LP, Pandey A, Pandey DK, Reeves MJ, Ritchey MD, Rodriguez CJ, Roth GA, Rosamond WD, Sampson UKA, Satou GM, Shah SH, Spartano NL, Tirschwell DL, Tsao CW, Voeks JH, Willey JZ, Wilkins JT, Wu JH, Alger HM, Wong SS and Muntner P; American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics-2018 update: a report from the American Heart Association. Circulation 2018; 137: e67-e492.
[3] Rahman AU, Alsenani Y, Zafar A, Ullah K, Rabie K and Shongwe T. Enhancing heart disease prediction using a self-attention-based transformer model. Sci Rep 2024; 14: 514.
[4] World Health Organization. Global cardiovascular disease statistics and impacts: Key findings and challenges. World Health Statistics Report 2024.
[5] Atimbire SA, Appati JK and Owusu E. Empirical exploration of whale optimisation algorithm for heart disease prediction. Sci Rep 2024; 14: 4530.
[6] El-Sofany H, Bouallegue B and El-Latif YMA. A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Sci Rep 2024; 14: 23277.
[7] Pal P, Singh HV, Grover V, Manikandan R, Karimi R and Khishe M. Interactive cardiovascular disease prediction system using learning techniques: insights from extensive experiments. Results Control Optim 2025; 19: 100560.
[8] Artinian NT, Fletcher GF, Mozaffarian D, Kris-Etherton P, Van Horn L, Lichtenstein AH, Kumanyika S, Kraus WE, Fleg JL, Redeker NS, Meininger JC, Banks J, Stuart-Shor EM, Fletcher BJ, Miller TD, Hughes S, Braun LT, Kopin LA, Berra K, Hayman LL, Ewing LJ, Ades PA, Durstine JL, Houston-Miller N and Burke LE; American Heart Association Prevention Committee of the Council on Cardiovascular Nursing. Interventions to promote physical activity and dietary lifestyle changes for cardiovascular risk factor reduction in adults: a scientific statement from the American Heart Association. Circulation 2010; 122: 406-441.
[9] Karwa V, Wanjari A, Kumar S, Dhondge RH, Patil R and Kothari M. Optimizing cardiovascular health: a comprehensive review of risk assessment strategies for primary prevention. Cureus 2024; 16: e66341.
[10] Perone F, Bernardi M, Spadafora L, Betti M, Cacciatore S, Saia F, Fogacci F, Jaiswal V, Asher E, Paneni F, De Rosa S, Banach M, Biondi Zoccai G and Sabouret P. Non-traditional cardiovascular risk factors: tailored assessment and clinical implications. J Cardiovasc Dev Dis 2025; 12: 171.
[11] Bouqentar MA, Terrada O, Hamida S, Saleh S, Lamrani D, Cherradi B and Raihani A. Early heart disease prediction using feature engineering and machine learning algorithms. Heliyon 2024; 10: e38731.
[12] Shorewala V. Early detection of coronary heart disease using ensemble techniques. Inform Med Unlocked 2021; 26: 100655.
[13] Noori Mirtaheri P, Akhbari M, Najafi F, Mehrabi H, Babapour A, Rahimian Z, Rigi A, Rahbarbaghbani S, Mobaraki H, Masoumi S, Nouri D, Mirzohreh ST, Sadat Rafiei SK, Asadi Anar M, Golkar Z, Asadollah Salmanpour Y, Vesali Mahmoud A, Gholami Chahkand MS and Khodaei M. Performance of deep learning models for automatic histopathological grading of meningiomas: a systematic review and meta-analysis. Front Neurol 2025; 16: 1536751.
[14] Gholami Chahkand MS, Karimi MA, Aghazadeh-Habashi K, Esmaeilpour Moallem F, Mehrabanpour R, Dadkhah PA, Esmailinia R, Esfandiari N, Azarm E, Rafiei SKS, Asadi Anar M and Shahriari A. Machine learning-based detection of EGFR mutation and HER2 overexpression in metastatic brain adenocarcinoma: systematic review and meta-analysis. Top Magn Reson Imaging 2025; 34: e0320.
[15] Saeed MH and Hama JI. Cardiac disease prediction using AI algorithms with SelectKBest. Med Biol Eng Comput 2023; 61: 3397-3408.
[16] Chen T and Guestrin C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: Association for Computing Machinery; 2016. pp. 785-794.
[17] Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande S and Singh P. Prediction of heart disease using a combination of machine learning and deep learning. Comput Intell Neurosci 2021; 2021: 8387680.
[18] Bani Hani SH and Ahmad MM. Machine-learning algorithms for ischemic heart disease prediction: a systematic review. Curr Cardiol Rev 2023; 19: e090622205797.
[19] Teja MD and Rayalu GM. Optimizing heart disease diagnosis with advanced machine learning models: a comparison of predictive performance. BMC Cardiovasc Disord 2025; 25: 212.

[20] Janosi A, Steinbrunn W, Pfisterer M and Detrano R. Heart disease [Dataset]. UCI Machine Learning Repository, Hungarian Institute of Cardiology. Budapest, University Hospital, Zurich, 1989.
[21] Al-Alshaikh HA, P P, Poonia RC, Saudagar AKJ, Yadav M, AlSagri HS and AlSanad AA. Comprehensive evaluation and performance analysis of machine learning in heart disease prediction. Sci Rep 2024; 14: 7819.
[22] Xia B, Innab N, Kandasamy V, Ahmadian A and Ferrara M. Intelligent cardiovascular disease diagnosis using deep learning enhanced neural network with ant colony optimization. Sci Rep 2024; 14: 21777.
[23] Cao K, Liu C, Yang S, Zhang Y, Li L, Jung H and Zhang S. Prediction of cardiovascular disease based on multiple feature selection and improved PSO-XGBoost model. Sci Rep 2025; 15: 12406.
[24] Darolia A, Chhillar RS, Alhussein M, Dalal S, Aurangzeb K and Lilhore UK. Enhanced cardiovascular disease prediction through self-improved aquila optimized feature selection in quantum neural network & LSTM model. Front Med (Lausanne) 2024; 11: 1414637.
[25] Bhagawati M, Paul S, Mantella L, Johri AM, Laird JR, Singh IM, Singh R, Garg D, Fouda MM, Khanna NN, Cau R, Abraham A, Al-Maini M, Isenovic ER, Sharma AM, Fernandes JFE, Chaturvedi S, Karla MK, Nicolaides A, Saba L and Suri JS. Deep learning approach for cardiovascular disease risk stratification and survival analysis on a Canadian cohort. Int J Cardiovasc Imaging 2024; 40: 1283-1303.
[26] Deng Y, Liu L, Jiang H, Peng Y, Wei Y, Zhou Z, Zhong Y, Zhao Y, Yang X, Yu J, Lu Z, Kho A, Ning H, Allen NB, Wilkins JT, Liu K, Lloyd-Jones DM and Zhao L. Comparison of state-of-the-art neural network survival models with the pooled cohort equations for cardiovascular disease risk prediction. BMC Med Res Methodol 2023; 23: 22.
[27] Meng LB, Cui K, Song Y, Li B, Zhang Y and Dou K. The association between sleep disorders and the risk of atherosclerotic cardiovascular disease: regression analysis and neural network prediction based on NHANES data. J Cardiothorac Surg 2026; 21: 164.
[28] Ejiyi CJ, Qin Z, Nneji GU, Monday HN, Agbesi VK, Ejiyi MB, Ejiyi TU and Bamisile OO. Enhanced cardiovascular disease prediction modelling using machine learning techniques: a focus on CardioVitalnet. Network 2025; 36: 716-748.
[29] Sadr H, Salari A, Ashoobi MT and Nazari M. Cardiovascular disease diagnosis: a holistic approach using the integration of machine learning and deep learning models. Eur J Med Res 2024; 29: 455.
[30] Heo J, Yoon JG, Park H, Kim YD, Nam HS and Heo JH. Machine learning-based model for prediction of outcomes in acute stroke. Stroke 2019; 50: 1263-1265.
[31] Holm PC, Haue AD, Westergaard D, Röder T, Banasik K, Tragante V, Johansen CH, Christensen AH, Thomas L, Nøst TH, Skogholt AH, Iversen KK, Pedersen F, Høfsten DE, Pedersen OB, Ostrowski SR, Ullum H, Svendsen MN, Gjødsbøl IM, Gudnason T, Guðbjartsson DF, Helgadottir A, Hveem K, Køber LV, Holm H, Stefansson K, Brunak S and Bundgaard H. Development and validation of a neural network survival prediction model for ischemic heart disease. Cardiovasc Diabetol 2026; 25: 59.
[32] Oikonomou EK and Khera R. Machine learning in precision diabetes care and cardiovascular risk prediction. Cardiovasc Diabetol 2023; 22: 259.
[33] Luo X, Li B, Zhu R, Tai Y, Wang Z, He Q, Zhao Y, Bi X and Wu C. Development and validation of an interpretable machine learning model for predicting in-hospital mortality for ischemic stroke patients in ICU. Int J Med Inform 2025; 198: 105874.
[34] Ordikhani M, Saniee Abadeh M, Prugger C, Hassannejad R, Mohammadifard N and Sarrafzadegan N. An evolutionary machine learning algorithm for cardiovascular disease risk prediction. PLoS One 2022; 17: e0271723.
[35] Wang Z, Gu Y, Huang L, Liu S, Chen Q, Yang Y, Hong G and Ning W. Construction of machine learning diagnostic models for cardiovascular pan-disease based on blood routine and biochemical detection data. Cardiovasc Diabetol 2024; 23: 351.
[36] Nusinovici S, Rim TH, Li H, Yu M, Deshmukh M, Quek TC, Lee G, Chong CCY, Peng Q, Xue CC, Zhu Z, Chew EY, Sabanayagam C, Wong TY, Tham YC and Cheng CY. Application of a deep-learning marker for morbidity and mortality prediction derived from retinal photographs: a cohort development and validation study. Lancet Healthy Longev 2024; 5: 100593.
[37] Hughes JW, Olgin JE, Avram R, Abreau SA, Sittler T, Radia K, Hsia H, Walters T, Lee B, Gonzalez JE and Tison GH. Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation. JAMA Cardiol 2021; 6: 1285-1295.
[38] Triantafyllidis A, Kondylakis H, Katehakis D, Kouroubali A, Koumakis L, Marias K, Alexiadis A, Votis K and Tzovaras D. Deep learning in mHealth for cardiovascular disease, diabetes, and cancer: systematic review. JMIR Mhealth Uhealth 2022; 10: e32344.
[39] Li F, Wu P, Ong HH, Peterson JF, Wei WQ and Zhao J. Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction. J Biomed Inform 2023; 138: 104294.

Table S1. Train-validation-test split

Dataset	N	Heart disease present, N	Heart disease absent, N	Heart disease present, %
Training set	212	97	115	45.75
Validation set	45	21	24	46.67
Test set	46	21	25	45.65
Total	303	139	164	45.87

Table S2. Feature importance from XGBoost and random forest

XGBoost		Random Forest	
Feature	Importance	Feature	Importance
Thallium_Normal	0.3370	Max HR	0.1291
Thallium_Reversible defect	0.1723	ST depression	0.1014
Chest pain type_Typical angina	0.0725	Age	0.0978
ST depression	0.0589	Thallium_Normal	0.0968
Exercise angina_No exercise-induced angina	0.0561	Thallium_Reversible defect	0.0855
Max HR	0.0422	BP	0.0819
Age	0.0398	Cholesterol	0.0817
Chest pain type_Non-anginal pain	0.0371	Exercise angina_No exercise-induced angina	0.0562
Sex_Male	0.0353	Number of vessels fluro_1 major vessel	0.0420
Slope of ST_Flat	0.0225	Chest pain type_Non-anginal pain	0.0398
BP	0.0217	Sex_Male	0.0365
Number of vessels fluro_1 major vessel	0.0208	Slope of ST_Upsloping	0.0257
Slope of ST_Upsloping	0.0196	Chest pain type_Atypical angina	0.0241
Cholesterol	0.0191	Slope of ST_Flat	0.0211
Chest pain type_Atypical angina	0.0168	EKG results_Normal	0.0205
Number of vessels fluro_2 major vessels	0.0165	Chest pain type_Typical angina	0.0185
EKG results_Normal	0.0106	Number of vessels fluro_2 major vessels	0.0175
FBS over 120_Fasting blood sugar > 120 mg/dl	0.0012	Number of vessels fluro_3 major vessels	0.0143
EKG results_ST-T wave abnormality	0.0000	FBS over 120_Fasting blood sugar > 120 mg/dl	0.0095
Number of vessels fluro_3 major vessels	0.0000	EKG results_ST-T wave abnormality	0.0001