ACCSCIENCE PUBLISHING

Brain & Heart

ORIGINAL RESEARCH ARTICLE

Evaluating machine learning models for cardiovascular risk prediction: A Shapley Additive Explanations-based approach with statistical testing

Shiv Kunwar, Swathi Ganesan*, and Sangita Pokhrel

Department of Data Science, York St. John University, London, United Kingdom

Abstract

Cardiovascular disease (CVD) remains the leading global cause of mortality, underscoring the need for accurate and interpretable prediction models to facilitate early diagnosis. Existing machine learning (ML) approaches often face challenges balancing predictive performance with clinical interpretability, limiting their adoption. This study introduces a structured evaluation framework combining A/B testing with statistical hypothesis validation to rigorously compare ML models for CVD risk prediction. Utilizing a dataset of 1,001 patient records, models including logistic regression, random forest (RF), artificial neural networks, and extreme gradient boosting (XGBoost) were trained and evaluated. Synthetic Minority Oversampling Technique was applied to address class imbalance, while Shapley Additive Explanations (SHAP) provided insights into feature contributions and guided the development of reduced-feature models. Results indicate that RF achieved the highest accuracy (98.5%) and area under the receiver operating characteristic curve (0.9991), whereas XGBoost coupled with SHAP enabled effective feature selection with minimal loss in predictive power. A/B testing demonstrated the trade-offs between model complexity and interpretability, while statistical testing confirmed the significance of performance differences. These findings suggest that interpretable, reduced-feature models may be viable for deployment in resource-limited clinical settings, advancing the integration of artificial intelligence in cardiovascular healthcare.

Keywords: Cardiovascular disease; Machine learning; Shapley Additive Explanations; A/B testing; Statistical validation; Feature selection

***Corresponding author:**
Swathi Ganesan
(s.ganesan@yorksj.ac.uk)

Citation: Kunwar S, Ganesan S, Pokhrel S. Evaluating machine learning models for cardiovascular risk prediction: A Shapley Additive Explanations-based approach with statistical testing. *Brain & Heart*. 2026;4(2):025260032. doi: 10.36922/BH025260032

Received: June 27, 2025

Revised: February 2, 2026

Accepted: February 10, 2026

Published online: March 6, 2026

1. Introduction

Cardiovascular disease (CVD) remains the leading cause of death globally, affecting both men and women across diverse populations. It includes conditions such as coronary artery disease (CAD), cerebrovascular diseases, congenital heart defects, and myocardial disorders. Among these, CAD is the most common, accounting for approximately 64% of all CVD cases, making it a major contributor to global premature mortality.[1] According to the World Health Organization,[2] CVD causes an estimated 17.9 million deaths annually,

a figure expected to rise beyond 2030. The primary risk factors for CVD include hypertension, high cholesterol, obesity, diabetes, smoking, lack of physical activity, and excessive alcohol consumption, many of which are linked to lifestyle and metabolic health issues. Due to its high prevalence and the significant strain it places on healthcare systems, improving early diagnosis and management of CVD is a pressing global need. Predictive analytics has significantly transformed healthcare by utilizing historical data, machine learning (ML) algorithms, and statistical techniques to forecast future health outcomes. The increasing reliance on electronic health records has enhanced predictive models, enabling healthcare providers to identify high-risk patients, optimize treatment plans, and improve overall patient outcomes.[1]

Machine learning and artificial intelligence (AI) offer promising approaches to early diagnosis by leveraging pattern recognition and predictive analytics. These technologies can analyze vast amounts of data to identify risk factors long before symptoms appear.[3,4] Early diagnosis also contributes to reduced healthcare costs by minimizing the need for intensive treatment and lowering hospital readmissions.[5] Predictive analytics, powered by AI, is reshaping personalized medicine and risk assessment. Using large datasets from electronic health records, genetic profiles, and patient histories, AI models can accurately assess CVD risk, enabling proactive healthcare strategies and better resource allocation.[6] Meanwhile, hospitals benefit by identifying high-risk individuals early, reducing emergency incidents, and enhancing patient care.[5] ML models offer significant advantages over traditional statistical methods by efficiently processing large, complex datasets and identifying non-linear relationships among risk factors. ML algorithms, such as k-nearest neighbors, support vector machines, and random forest (RF) classifiers, have demonstrated superior accuracy in predicting cardiovascular risk compared to traditional methods.[7] These models integrate diverse data sources, including imaging, patient records, and lab results, to uncover hidden correlations and provide comprehensive risk assessments.[5] Ensemble models such as RF and extreme gradient boosting (XGBoost) are particularly effective, as they reduce overfitting and can handle imbalanced datasets.[2,8-10] Recent advances in high-dimensional optimization and regularization techniques have further improved the scalability and robustness of machine learning models in complex predictive tasks.[11]

Logistic regression (LR) remains a popular baseline model due to its simplicity and interpretability, often enhanced with techniques such as principal component analysis (PCA) and K-means++ clustering for better performance.[12,13] Despite its advantages, LR struggles with capturing complex, non-linear relationships. Deep learning models, including neural networks, show promise in identifying intricate patterns in large datasets but are often criticized for their lack of interpretability—a critical issue in clinical decision-making.[1,6] Explainable AI methods, such as Shapley Additive Explanations (SHAP), have been introduced to enhance transparency by highlighting the contributions of individual features to model predictions.[3,14,15] However, SHAP can be computationally expensive, especially with high-dimensional data. Improvements like Kernel SHAP and TreeSHAP have been developed to address these issues.[3] One of the persistent challenges in medical predictive modeling is class imbalance, where the number of healthy individuals significantly outweighs those with CVD. This issue can bias models toward the majority class, leading to poor sensitivity. Techniques such as Synthetic Minority Oversampling Technique (SMOTE) and its variants (e.g., Adaptive-SMOTE) are employed to generate synthetic minority samples, improving model recall and sensitivity while reducing false negatives.[16-19]

Additionally, there is a lack of standardized protocols for applying SHAP-driven feature reduction in clinical settings. Most studies focus on model explainability rather than formalized, repeatable processes for selecting features based on SHAP values. This limits reproducibility and comparability across studies, hindering broader clinical adoption.

This study aims to advance CVD risk prediction using ML by evaluating current models, addressing data imbalance, and applying interpretability tools like SHAP. It also compares traditional models like LR with advanced neural networks, focusing on performance metrics such as accuracy, precision, recall, and F1-score. The ultimate goal is to support clinical decision-making, improve early detection, and reduce CVD-related mortality worldwide.

This research is designed primarily for screening and early triage of individuals at potential cardiovascular risk, prior to imaging or specialist referral. Unlike models tailored for intra-operative or acute care prediction, the present framework aims to support primary prevention and outpatient decision-support, helping clinicians identify high-risk individuals during routine health assessments. Such positioning aligns the model with the goals of early detection and population-level risk management rather than disease-specific prognostication.

2. Materials and methods

2.1. Materials

The dataset used in this study was obtained from a publicly accessible medical repository hosted on Kaggle and is intended for research purposes. It comprises 1,001 anonymized patient records with 13 clinical predictor variables, 1 non-predictive identifier (patient identification), and 1 binary target variable related to cardiovascular health. The predictor variables include age, gender, resting blood pressure, serum cholesterol, and maximum heart rate achieved. The target variable indicates the presence (1) or absence (0) of CVD. All data were provided in a structured CSV format, facilitating seamless integration with ML tools and reproducible experimentation. Table 1 provides a detailed overview of all clinical attributes used in the model, including assigned codes, units, and data types.

2.2. Data preprocessing

Initial data cleaning involved handling missing and zero values. Categorical variables with missing data, such as the slope of the peak exercise ST segment, were imputed using the mode, while continuous variables like serum cholesterol were imputed with the median to reduce the impact of outliers. The patient identification column was excluded due to its non-predictive nature. Categorical features (chest pain type, resting electrocardiogram results, slope, and number of major vessels) were one-hot encoded to improve compatibility with ML models, particularly artificial neural networks (ANNs). Numerical features (age, resting blood pressure, serum cholesterol, maximum heart rate, and ST depression induced by exercise, and "oldpeak") were standardized using the StandardScaler function to normalize their scales. The processed dataset was then split into training (80%) and testing (20%) subsets, corresponding to 800 and 201 records, respectively, ensuring unbiased model evaluation. The overall preprocessing workflow, including data cleaning, encoding, scaling, and train–test splitting, is illustrated in Figure 1. Exploratory data analysis was performed to visualize the distribution of key variables and relationships among features. Histograms illustrating the distributions of age, resting blood pressure, serum cholesterol, maximum heart rate, and oldpeak are shown in Figure 2. Additional visualizations of categorical variables and their relations to the target variable are presented in Figure 3.

2.3. Model development and evaluation

Four ML algorithms were employed to predict CVD: LR,

Table 1. Attribute description and data type

Attribute	Assigned code	Unit	Type of the data
Patient identification number	patientid	Number	Numeric
Age	age	Years	Numeric
Gender	gender	0 (Female), 1 (Male)	Binary
Chest pain type	chestpain	0 (Typical angina), 1 (Atypical angina), 2 (Non-anginal pain), 3 (Asymptomatic)	Nominal
Resting blood pressure	restingBP	94–200 mmHg	Numeric
Serum cholesterol	serumcholestrol	126–564 mg/dL	Numeric
Fasting blood sugar	fastingbloodsugar	0 (False), 1 (True) > 120 mg/dL	Binary
Resting electrocardiogram results	restingelectro	0 (Normal), 1 (Having ST–T wave abnormality [T-wave inversions and/or ST elevation or depression of >0.05 mV]), 2 (Showing potential or definite left ventricular hypertrophy by Estes' criteria)	Nominal
Maximum heart rate achieved	maxheartrate	71–202 beats/min	Numeric
Exercise-induced angina	exerciseangia	0 (No), 1 (Yes)	Binary
Oldpeak = ST	oldpeak	0–6.2	Numeric
Slope of the peak exercise ST segment	slope	1 (Upsloping), 2 (Flat), 3 (Downsloping)	Nominal
Number of major vessels	noofmajorvessels	0, 1, 2, 3	Numeric
Classification	target	0 (Absence of heart disease), 1 (Presence of heart disease)	Binary

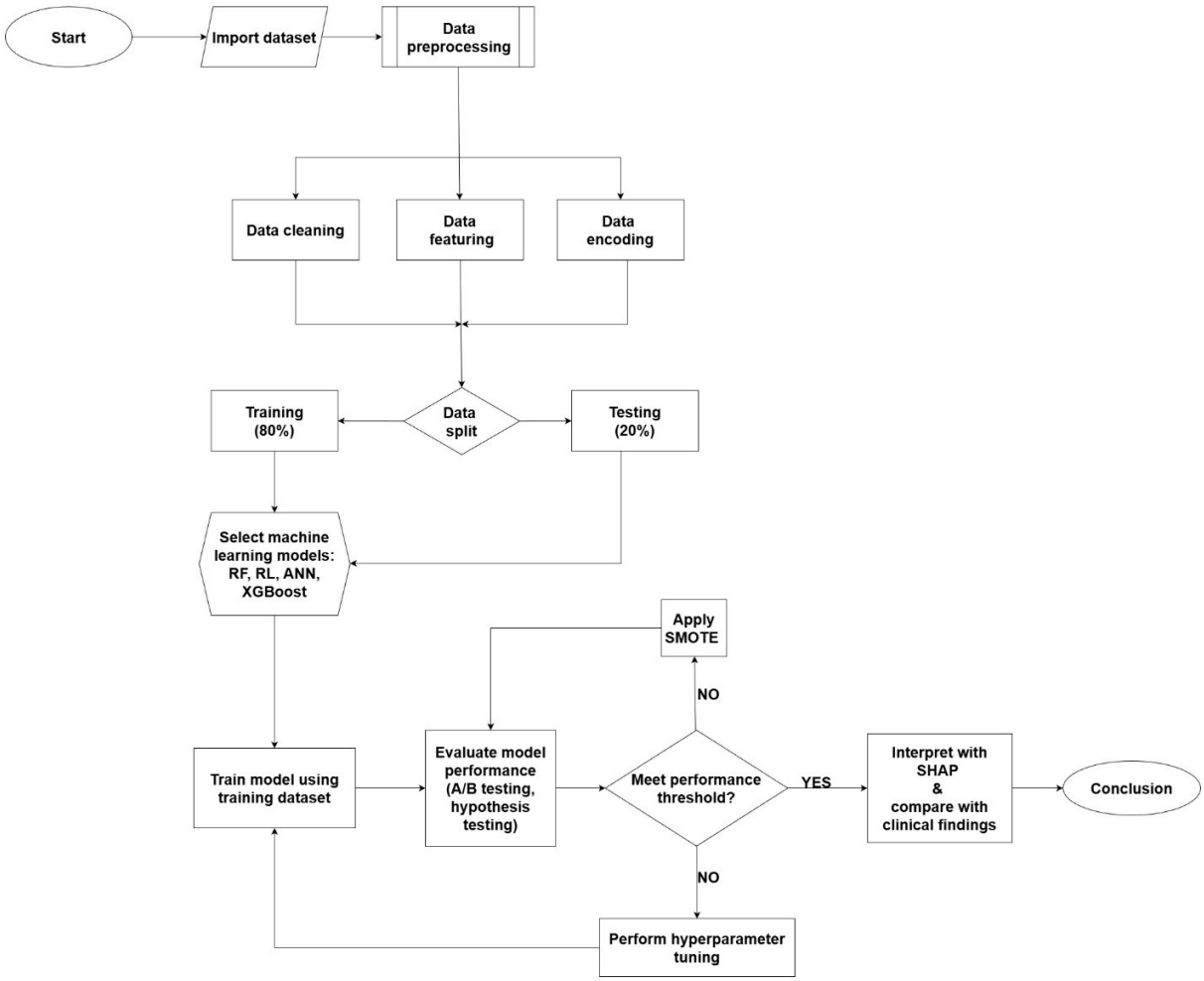

Figure 1. Workflow of the methodology for cardiovascular disease prediction
Abbreviations: ANN: Artificial neural network; LR: Logistic regression; RF: Random forest; SHAP: Shapley Additive Explanations; SMOTE: Synthetic Minority Oversampling Technique; XGBoost: Extreme gradient boosting.

RF, ANN, and XGBoost. These models were selected for their proven effectiveness in classification tasks involving clinical data. The training set was used for model fitting, while the testing set was reserved for independent evaluation.

Class imbalance was addressed using SMOTE to generate synthetic samples of the minority class, improving sensitivity without compromising specificity. Hyperparameter tuning was performed using grid search, exploring a comprehensive range of parameters to optimize each model's performance. This process was informed by recent advances in high-dimensional model optimization, which emphasize efficient variable selection and computational scalability in large predictor spaces.[11] Although the present study did not employ support vector machines directly, these optimization principles guided the design of the hyperparameter tuning framework, particularly in balancing model complexity with computational feasibility.

For LR, the regularization strength was tuned over $C \in \{0.01, 0.1, 1, 10\}$ with both L1 and L2 penalties evaluated. For RF, the number of trees (n_estimators $\in \{50, 100, 200\}$), maximum tree depth (max_depth $\in \{\text{None}, 5, 10, 20\}$), and minimum samples per split (min_samples_split $\in \{2, 5, 10\}$) were explored over predefined ranges. For XGBoost, learning rate (learning_rate $\in \{0.01, 0.1, 0.2\}$), maximum depth (max_depth $\in \{3, 5, 7\}$), and number of estimators (n_estimators $\in \{50, 100, 200\}$) were tuned. ANN training hyperparameters, including learning rate and batch size, were also adjusted during preliminary tuning. For all models, optimal configurations were selected based on mean cross-validation performance on the training data.

For the ANN model, architectural and training parameters were selected based on preliminary tuning and

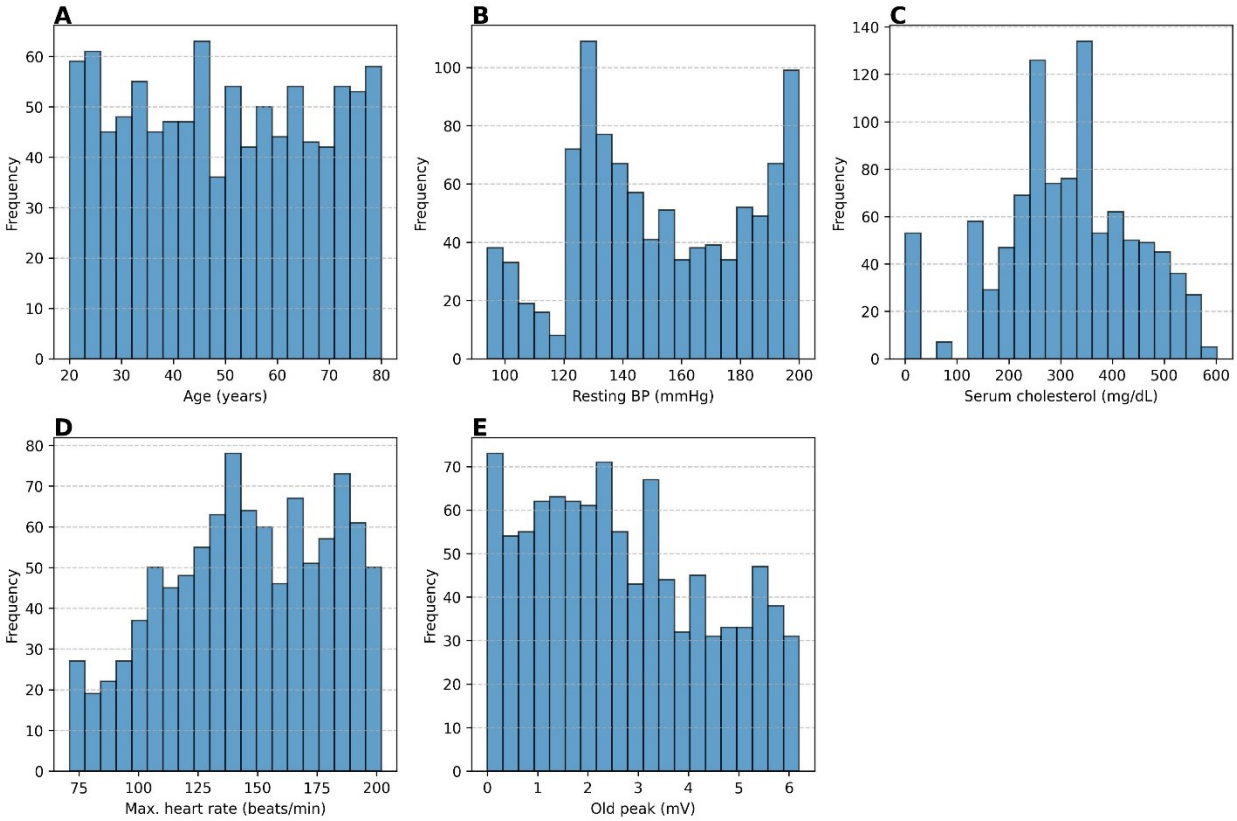

Figure 2. Histogram analysis of numeric features: (A) Age, (B) resting blood pressure (BP), (C) serum cholesterol, (D) maximum heart rate, and (E) old peak

are reported explicitly below to ensure reproducibility. The ANN implemented in this study followed a feedforward architecture designed for binary classification. The network consisted of two fully connected hidden layers, with 64 and 32 neurons, respectively. Each hidden layer employed the rectified linear unit (ReLU) activation function, while the output layer used a sigmoid function to generate probability estimates for the presence of CVD. The model was trained using the Adam optimizer with a learning rate of 0.001 and binary cross-entropy as the objective function. No explicit regularization techniques, such as dropout or L2 regularization, were applied. Training was conducted for a maximum of 100 epochs with a batch size of 32, and model convergence was controlled using a fixed-epoch training strategy without early stopping. This configuration enabled effective learning of non-linear feature interactions while maintaining comparability with the other evaluated ML models.

Model evaluation metrics included accuracy, precision, recall, F1-score, and receiver operating characteristic area under the curve (ROC-AUC), providing a balanced assessment of classification quality, as illustrated in Figure 4. To enhance interpretability, SHAP was applied to the XGBoost model, quantifying the contribution of each feature to model predictions. SHAP values also guided the construction of a reduced-feature model using the top-ranked predictors, tested to evaluate the trade-off between performance and simplicity. The SHAP summary plot and feature importance ranking are shown in Figure 5.

2.4. Statistical analysis

To prevent data leakage, SMOTE was applied exclusively to the training data within each fold during cross-validation, while validation and test subsets were kept strictly independent and unaltered. SMOTE was implemented using its standard nearest-neighbor–based sampling strategy with a fixed random seed to ensure reproducibility. Following an initial 80/20 train–test split, 10-fold cross-validation was conducted exclusively on the training subset to generate performance estimates. Paired *t*-tests were conducted with a significance threshold of $p < 0.05$ to assess whether the observed performance differences between models (RF and XGBoost) were statistically significant. The paired *t*-test assumes approximate normality of performance differences across folds and independence between cross-validation estimates. Given

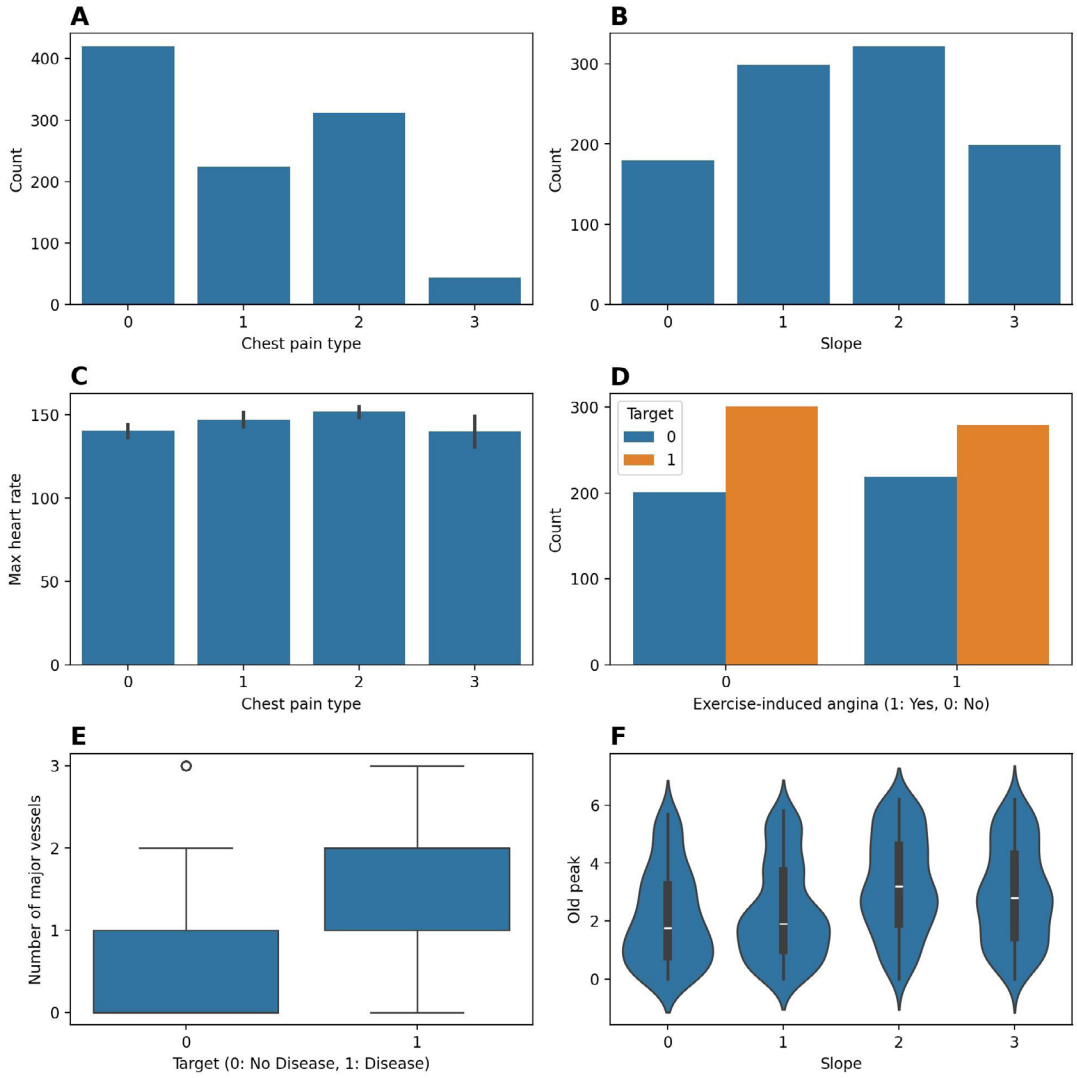

Figure 3. Distribution and relation between features. (A) Chest pain type distribution. (B) Slope distribution. (C) Maximum heart rate by chest pain type. (D) Exercise-induced angina vs. target. (E) Number of major vessels by target. (F) Distribution of old peak by slope.

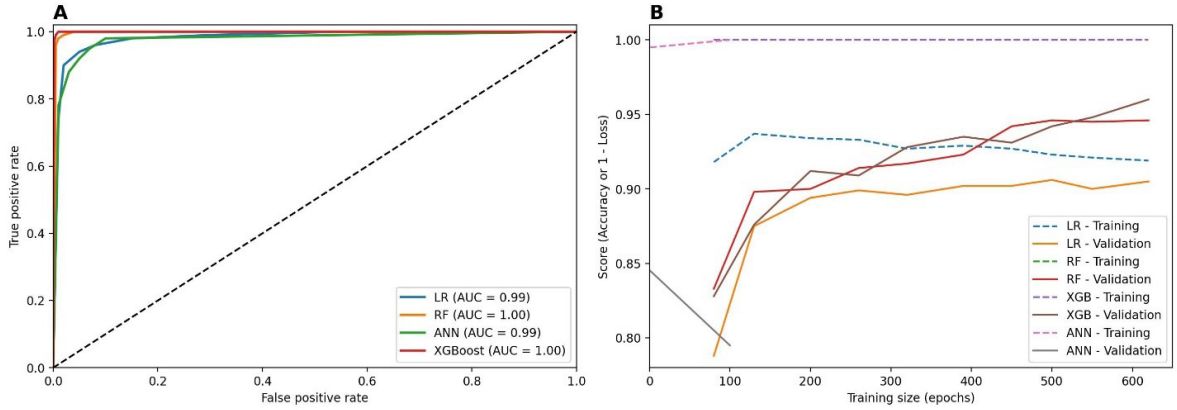

Figure 4. Comparison of (A) receiver operating characteristic and (B) learning curves across the tested models: artificial neural network (ANN), logistic regression (LR), random forest (RF), and extreme gradient boosting (XGBoost)
Abbreviation: AUC: Area under the curve.

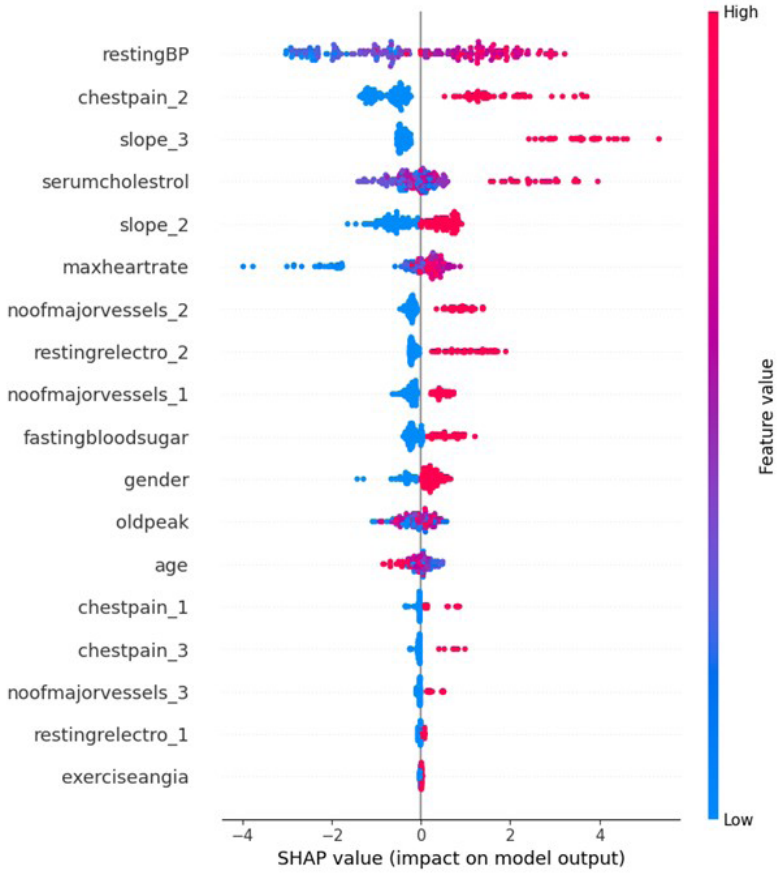

Figure 5. Shapley Additive Explanations (SHAP) summary plot

the repeated-measures design and commonly reported use of this test in comparative ML studies, the paired *t*-test was considered appropriate. Non-parametric alternatives such as the Wilcoxon signed-rank test may be explored in future work to further assess robustness. Furthermore, A/B testing compared the full-feature model against the SHAP-driven reduced-feature model on the same test data, evaluating the impact of feature reduction on predictive performance and interpretability. All analyses were implemented using Python libraries, including Scikit-learn, PyTorch, and XGBoost. Statistical testing was performed with SciPy.

3. Results

3.1. Model performance comparison

Table 2 summarizes the performance metrics of the ML models evaluated for CVD prediction. RF achieved the highest accuracy of 98.5% and an ROC-AUC of 0.9991. The consistently high ROC-AUC values observed for tree-based models indicate strong separability of cardiovascular risk patterns within the available feature space, while the comparatively lower performance of LR highlights the limitations of linear decision boundaries for capturing complex, non-linear clinical relationships. LR showed the lowest accuracy and ROC-AUC, while ANN and XGBoost produced competitive results with 98% accuracy and ROC-AUC values above 0.97. The ROC curves and learning curves for each model are presented in **Figure 4**. RF and XGBoost demonstrated near-perfect ROC-AUC scores, suggesting strong classification capabilities. Learning curves indicate that both models generalize well, with training and validation performances closely aligned. LR appears to underfit, whereas the ANN model benefits from increased training data, showing improved performance as the sample size increases. The close alignment between training and validation curves for RF and XGBoost suggests stable generalization under cross-validation; however, the near-saturation of performance also implies that these results may reflect dataset-specific characteristics rather than universally generalizable clinical behavior.

3.2. Shapley Additive Explanations analysis and feature selection

Shapley Additive Explanations was applied to the XGBoost model to interpret feature contributions. The SHAP summary plot (Figure 5) highlights the most influential

predictors, with chest pain type 2 (non-anginal pain) emerging as a dominant feature. This finding aligns with clinical literature suggesting that atypical chest pain patterns may be underrecognized yet indicative of underlying cardiovascular risk.[20] Similarly, resting blood pressure and a downsloping ST-segment slope, both well-established markers of cardiac strain, showed high SHAP values, reinforcing their diagnostic importance.

Table 2. Model performance comparison

Model	Accuracy (%)	ROC-AUC
Random forest	98.5	0.9991
Logistic regression	95.0	0.9874
Artificial neural network	98.0	0.9886
XGBoost (full-feature model)	98.0	0.9987
XGBoost (SHAP-reduced feature model)	93.5	0.9965

Abbreviations: AUC: Area under the curve; ROC: Receiver operating characteristic; SHAP: Shapley Additive Explanations; XGBoost: Extreme gradient boosting.

Based on the SHAP-derived feature importance rankings, a reduced-feature XGBoost model was developed with the top 10 features. The selection of 10 features represents a pragmatic design choice aimed at balancing model interpretability, computational efficiency, and predictive performance, rather than a universally optimal threshold. Although the reduced-feature model exhibited a modest decrease in accuracy (from 98% to 93.5%), it retained strong discriminative capability, demonstrating that substantial dimensionality reduction can be achieved with limited loss of performance.

It is important to note that SHAP-based feature selection is inherently context-dependent and sensitive to dataset characteristics. No formal sensitivity analysis was conducted in this study to evaluate alternative feature-set sizes, and features with low global SHAP importance may still hold clinical relevance for specific patient subgroups. Consequently, SHAP-driven feature selection should be viewed as a decision-support mechanism that benefits from clinical oversight rather than an automated elimination rule. These findings nonetheless highlight the potential of simplified, interpretable models for deployment in resource-limited healthcare settings, where transparency and computational efficiency are essential. The concentration of SHAP importance among a limited subset of features further supports the feasibility of dimensionality reduction, while reinforcing the need for cautious interpretation in the absence of external validation.

3.3. A/B testing: Full vs. Reduced-feature models

An A/B testing framework was implemented to compare the full-feature XGBoost model with the SHAP-based reduced-feature model. Both models were evaluated on the same test set to ensure fair comparison. The full-feature model achieved a higher accuracy (98%) compared to the SHAP-reduced model's 93.5%, reflecting the trade-off between model complexity and interpretability (Table 2). The full-feature model attained a slightly higher ROC-AUC (0.9987) compared with the SHAP-reduced model (0.9965), indicating that feature reduction resulted in only a marginal decrease in predictive performance. This observation is further supported by the ROC curves in Figure 6, where both models demonstrated strong discriminative capability with closely aligned performance profiles.

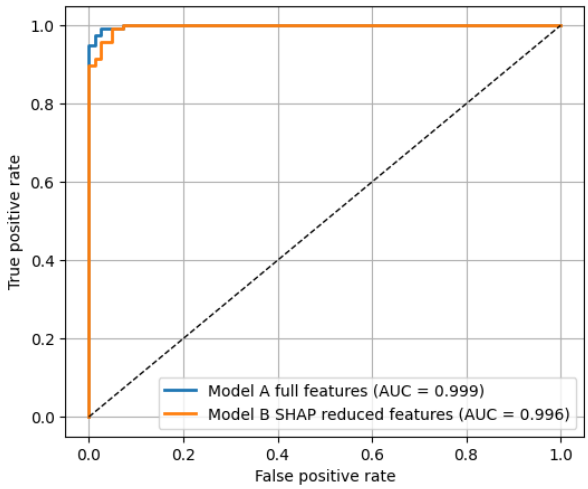

Figure 6. Receiver operating characteristic (ROC) for full feature model (Model A) and SHAP-reduced feature model (Model B)
Abbreviations: AUC: Area under the curve; SHAP: Shapley Additive Explanations.

These findings highlight the practical clinical value of reduced-feature models. Despite using fewer input variables, the SHAP-based model maintains high predictive performance while improving interpretability and computational efficiency. Such characteristics are particularly beneficial in resource-constrained healthcare environments, where transparent and lightweight decision-support systems are preferred.

3.4. Hypothesis testing

Paired *t*-tests were conducted on 10-fold cross-validation accuracy scores to statistically validate performance differences between RF and XGBoost. The results indicated a significant difference ($p < 0.05$), confirming that RF outperforms XGBoost on this dataset. Similar tests comparing RF with LR also yielded significant differences.

3.5. Model calibration and overfitting discussion

While RF achieved very high performance metrics (accuracy = 98.5%, ROC-AUC = 0.9991), such values may indicate potential model optimism or partial overfitting, particularly in single-source datasets. Cross-validation results showed stable mean accuracy with low variance, suggesting reasonable robustness; however, external validation remains essential. In a similar cardiovascular prediction study,[21] Brier scores have been used to evaluate the agreement between predicted probabilities and observed outcomes. Although formal calibration plots were not generated in this version, future analyses will include these measures to ensure that predicted risk levels correspond accurately to real-world event probabilities. Regularization and nested cross-validation will be incorporated in future work to further reduce model optimism and improve the robustness of performance estimation.

4. Discussion

This study demonstrates the effective use of ML models for CVD prediction, with RF achieving superior accuracy and ROC-AUC compared to other evaluated algorithms. The application of SHAP provided valuable interpretability, identifying clinically relevant predictors such as resting blood pressure, chest pain type, and serum cholesterol, consistent with established cardiovascular research.[22,23] Importantly, SHAP-guided feature reduction enabled the creation of a simplified XGBoost model that maintained strong predictive performance. This suggests that reduced-feature, interpretable models may be particularly advantageous in resource-limited settings where computational resources and clinical explainability are critical. Resource-limited settings may include hospitals or clinics lacking GPU-enabled infrastructure, cloud computing access, or trained data science personnel, where model simplicity and low computational overhead are essential.

To support broader clinical adoption, future implementations could benefit from alignment with established explainability frameworks, such as the Defense Advanced Research Projects Agency's explainable AI program[24] or the European Union's Ethics Guidelines for Trustworthy AI,[25] which emphasize transparency, user trust, and ethical deployment of AI systems. In this context, recent work has demonstrated that optimized feature selection using SHAP can improve both predictive performance and interpretability in CVD models, while maintaining computational efficiency and robustness across feature-reduced settings.[26] Furthermore, broader studies have emphasized that the integration of ML into healthcare must be guided by ethical governance, transparency, and clinician trust to ensure responsible and sustainable real-world adoption.[27] The A/B testing framework highlighted the trade-off between model complexity and interpretability, supporting the adoption of streamlined models without substantial loss of accuracy. Statistical hypothesis testing confirmed the robustness of the RF model over others, underscoring the need for rigorous performance validation alongside interpretability.

Comparative research in cardiology has demonstrated that ML can be effectively tailored to distinct clinical applications. For instance, Cicek *et al.*[28] developed a pre-operative risk model for predicting myocardial injury in elderly patients undergoing non-elective surgery, achieving superior performance compared with the Revised Cardiac Risk Index. In acute care, another study proposed a multimodal deep-learning framework that integrated imaging and clinical variables to predict short-term mortality in patients with pulmonary embolism, reaching an AUC of 0.98 and outperforming the standard Pulmonary Embolism Severity Index score.[29] Similarly, Yilmaz *et al.*[30] employed electrocardiogram waveform features from treadmill exercise tests to detect obstructive CAD, reporting an AUC of 0.78.

In contrast, the present study introduced a SHAP-based framework for comprehensive cardiovascular risk screening rather than disease-specific modeling. While the aforementioned studies focused on narrow, high-acuity conditions using specialized imaging or signal data, our approach prioritizes transparency, interpretability, and scalability across broader outpatient and preventive care populations. This distinction underscores the complementary relationship between potentially generalizable screening frameworks and specialized diagnostic models in advancing modern cardiology.

Despite promising accuracy and interpretability,

the models were trained solely on a secondary dataset containing 1,001 records. A single-source dataset may inherit selection bias and limited population diversity, which can influence learned decision boundaries and inflate performance metrics under controlled experimental conditions. In particular, homogeneity in demographic composition, clinical measurement protocols, or disease prevalence may reduce the variability encountered during training. This leads to overly optimistic estimates of generalization performance. Consequently, the near-perfect accuracy and ROC-AUC values reported in this study should be interpreted as upper-bound estimates rather than indicators of real-world clinical performance. Without external validation on independent, multi-center cohorts representing diverse populations and care settings, these findings cannot be confidently extrapolated to broader clinical contexts.

Future work will therefore focus on external validation using multi-center clinical datasets and exploration of federated learning frameworks to enable model generalization without compromising patient privacy. The exclusive use of SMOTE for class balancing may further introduce bias by synthetically amplifying minority patterns that do not exist in real clinical populations; alternative strategies such as cost-sensitive learning and adaptive resampling will be explored to improve ecological validity.

Furthermore, SHAP calculations for complex models, such as ANN and XGBoost, can be computationally demanding. Implementing optimized variants like TreeSHAP or distributed GPU computation could reduce latency, enabling near-real-time inference in resource-limited healthcare environments. Collaborating with cardiology experts to interpret SHAP outputs will also ensure that clinically relevant yet low-ranked features are not discarded during feature selection. Although SHAP provided valuable insights into feature contributions and enhanced model transparency, the interpretation of these explanations was not formally validated by clinical experts in the present study. While the identified predictors align with established cardiovascular risk factors reported in the literature, expert clinical review is essential to ensure that model explanations are consistent with real-world diagnostic reasoning and clinical workflows. In particular, clinician involvement is critical to contextualize SHAP outputs, mitigate the risk of overinterpreting data-driven patterns, and ensure that clinically meaningful variables, especially those relevant to specific patient subgroups, are not inadvertently overlooked. Future work will therefore prioritize close collaboration with cardiology specialists to clinically validate and refine model explanations prior to any prospective deployment.

In addition, federated learning approaches will be explored to leverage larger and more diverse datasets while preserving patient privacy. Real-time validation within clinical workflows, along with the integration of wearable sensor data, may further enhance model adaptability and support continuous cardiovascular risk monitoring.

Limitations include the modest sample size, single-source data, and absence of expert-validated clinical inputs. SHAP-based feature selection lacks a standardized clinical protocol, which may affect reproducibility. Additionally, the lack of calibration and real-time testing limits deployment feasibility. Future research will integrate clinical expert feedback, external datasets, and live validation in healthcare workflows to confirm model reliability and fairness across demographic subgroups.

5. Conclusion

In summary, this research integrates interpretable ML techniques with robust statistical validation to improve CVD prediction. RF emerged as the best-performing model, while SHAP analysis facilitated the development of reduced-feature models balancing accuracy and transparency. The findings support deploying interpretable and efficient predictive tools in clinical practice, particularly in environments with limited resources. Future work should focus on real-world clinical validation and prospective trials to evaluate model performance in operational healthcare workflows. Incorporating dynamic and demographic data sources, such as wearable devices and real-time monitoring systems from diverse people, could enhance the adaptability of predictive models. This will contribute to continuous and enhanced cardiovascular risk assessment and early diagnosis.

Acknowledgments

The authors would like to acknowledge the London Campus, York St. John University, for providing institutional support and a conducive research environment. Special thanks are extended to Associate Professor Dr. Nalinda Somasiri, Head of Department, for his valuable guidance and leadership throughout this study.

Funding

None.

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Shiv Kunwar

Formal analysis: Shiv Kunwar, Swathi Ganesan
Investigation: Sangita Pokhrel
Methodology: Shiv Kunwar
Supervision: Swathi Ganesan
Writing–original draft: Shiv Kunwar
Writing–review & editing: Swathi Ganesan, Sangita Pokhrel

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The dataset used in this study is openly available and can be accessed at https://www.kaggle.com/datasets/jocelyndumlao/cardiovascular-disease-dataset. It is shared under the CC0 (Public Domain) license, meaning there are no restrictions on reuse. This ensures the data are reliable and fully permissible for academic research, reproducibility, and machine learning applications. The data that support the findings of this study are available from the authors upon reasonable request.

References

1. Ogunpola A, Saeed F, Basurra S, Albarrak AM, Qasem SN. Machine learning-based predictive models for detection of cardiovascular diseases. *Diagnostics (Basel)*. 2024;14(2):144.

   doi: 10.3390/diagnostics14020144

2. World Health Organization. Cardiovascular Diseases. World Health Organization. Published 2025. Available from: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 [Last accessed on 2026 March 03].

3. Das R, Turkoglu I, Sengur A. Effective diagnosis of heart disease through neural networks ensembles. *Expert Syst Appl*. 2009;36(4):7675-7680.

   doi: 10.1016/j.eswa.2008.09.013

4. Huang G, Li Y, Jameel S, Long Y, Papanastasiou G. From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality? *Comput Struct Biotechnol J*. 2024;24:362-373.

   doi: 10.1016/j.csbj.2024.05.004

5. TR R, Lilhore UK, M P, Simaiya S, Kaur A, Hamdi M. Predictive analysis of heart diseases with machine learning approaches. *Malays J Comput Sci*. 2022;2022(spec iss 1):132-148.

   doi: 10.22452/mjcs.sp2022no1.10

6. Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective heart disease prediction using machine learning techniques. *Algorithms*. 2023;16(2):88.

   doi: 10.3390/a16020088

7. Hagan R, Gillan CJ, Mallett F. Comparison of machine learning methods for the classification of cardiovascular disease. *Inform Med Unlocked*. 2021;24:100606.

   doi: 10.1016/j.imu.2021.100606

8. Yang J, Guan J. A heart disease prediction model based on feature optimization and Smote-Xgboost algorithm. *Information*. 2022;13(10):475.

   doi: 10.3390/info13100475

9. Sena R, Hamida SB. ACTIVE SMOTE for imbalanced medical data classification. *Lect Notes Bus Inf Process*. 2024;486:81-97.

   doi: 10.1007/978-3-031-51664-1_6

10. Srivastava S, Upreti K, Shanbhog M. Analysis of cardiovascular diseases prediction using machine learning classification algorithms. In: *Proceedings of the 2024 IEEE Conference*. IEEE; 2024.

    doi: 10.1109/accai61061.2024.10601806

11. Wang K, *et al*. Convolution smoothing and non-convex regularization for SVM in high dimensions. *Appl Soft Comput*. 2024;155:111433.

    doi: 10.1016/j.asoc.2024.111433

12. Flores-Araiza D, Villegas-Jimenez A, Lopez-Tiro F, Gonzalez-Mendoza M. On the link between model performance and causal scoring of medical image explanations. In: *2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE; 2024:1-8.

    doi: 10.1109/cbms61543.2024.00009

13. Zhao R. Predicting cardiovascular disease using simple machine learning techniques. *Highlights Sci Eng Technol*. 2024;81:356-362.

    doi: 10.54097/3f973x84

14. Miao H. Logistic regression for cardiovascular diseases prediction by integrating PCA and K-means++. *Theor Nat Sci*. 2024;38(1):126-132.

    doi: 10.54254/2753-8818/38/20240569

15. Sakho A, Malherbe E, Scornet E. Do we need rebalancing strategies? A theoretical and empirical study around SMOTE and its variants. *arXiv*. Preprint posted online February 7, 2024.

    doi: 10.48550/arXiv.2402.03819

16. Newaz A, Ahmed N, Haq FS. Survival prediction of heart failure patients using machine learning techniques. *Inform Med Unlocked*. 2021;26:100772.

    doi: 10.1016/j.imu.2021.100772

17. Salman HA, Kalakech A, Steiti A. Random Forest Algorithm Overview. *Babylon J Mach Learn*. 2024;2024:69-79.

doi: 10.58496/BJML/2024/007

18. Thakur R. Explainable AI: developing interpretable deep learning models for medical diagnosis. *Int J Multidiscip Res*. 2024;6(4).

    doi: 10.36948/ijfmr.2024.v06i04.25281

19. Sadeghi Z, Alizadehsani R, *et al*. A review of explainable artificial intelligence in healthcare. *Comput Electr Eng*. 2024;118:109370.

    doi: 10.1016/j.compeleceng.2024.109370

20. Canto JG. Prevalence, Clinical Characteristics, and Mortality Among Patients With Myocardial Infarction Presenting Without Chest Pain. *JAMA*. 2000;283(24):3223.

    doi: 10.1001/jama.283.24.3223

21. Khera R, Haimovich J, Hurley NC, *et al*. Use of Machine Learning Models to Predict Death After Acute Myocardial Infarction. *JAMA Cardiol*. 2021;6(6):633.

    doi: 10.1001/jamacardio.2021.0122

22. Marketou ME, Vlachopoulos C, Hahalis G, *et al*. Clinical characteristics and management of patients with diabetes mellitus and stable coronary artery disease in daily clinical practice. The SCAD-DM Registry. *Hellenic J Cardiol*. 2021;62(6):408-415.

    doi: 10.1016/j.hjc.2020.12.006

23. Rim AJ, *et al*. Concussions are associated with increases in blood pressure and cardiovascular risk in American-style football athletes. *JACC Adv*. 2025;4(5):101717.

    doi: 10.1016/j.jacadv.2025.101717

24. Gunning D. *Explainable Artificial Intelligence (XAI)*. Defense Advanced Research Projects Agency (DARPA); 2017. Available from: https://www.darpa.mil/program/explainable-artificial-intelligence [Last accessed on 2026 March 03].

25. European Commission, High-Level Expert Group on AI. *Ethics guidelines for trustworthy AI*. Publications Office of the European Union; 2019. Available from: https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf [Last accessed on 2026 March 03].

26. Ganesan S, Somasiri N. Enhancing Cardiovascular Disease Prediction: Optimised Feature Selection and Machine Learning Techniques for Improved Accuracy. In: *Proceedings of the 10th International Conference on Machine Learning Technologies (ICMLT)*. IEEE; 2025:55-62.

    doi: 10.1109/icmlt65785.2025.11193317

27. Ganesan S, Somasiri N. Navigating the integration of machine learning in healthcare: challenges, strategies, and ethical considerations. *J Comput Cogn Eng*. 2025;4(1):8-23.

    doi: 10.47852/bonviewJCCE42023600

28. Cicek V, Babaoglu M, Saylik F, *et al*. A New Risk Prediction Model for the Assessment of Myocardial Injury in Elderly Patients Undergoing Non-Elective Surgery. *J Cardiovasc Dev Dis*. 2024;12(1):6.

    doi: 10.3390/jcdd12010006

29. Cicek V, Orhan AL, Saylik F, *et al*. Predicting short-term mortality in patients with acute pulmonary embolism with deep learning. *Circ J*. 2025;89(5):602-611.

    doi: 10.1253/circj.CJ-24-0630

30. Yilmaz A, Hayiroglu MI, Salturk S, *et al*. Machine learning approach on high-risk treadmill exercise test to predict obstructive coronary artery disease by using P, QRS, and T-wave features. *Curr Probl Cardiol*. 2023;48(2):101482.

    doi: 10.1016/j.cpcardiol.2022.101482