

An Explainable, Robust, and Empirically-Validated Stacked Ensemble (RXTG-Stack) for Cardiovascular Risk Prediction Across Heterogeneous Datasets

Talha Farooq Khan

The University of Faisalabad

Manahel Ansari

The University of Faisalabad

Maryam Tanveer

The University of Faisalabad

Fahad Alturise

falturise@qu.edu.sa

Qassim University

Article

Keywords: Heart Disease Prediction, Explainable AI, Machine Learning, Empirical Verification, Conformal Prediction, SHAP, Stacked Ensemble, Feature Engineering

Posted Date: August 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10457811/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Scientific Reports

An Explainable, Robust, and Empirically-Validated Stacked Ensemble (RXTG-Stack) for Cardiovascular Risk Prediction Across Heterogeneous Datasets

Talha Farooq Khan¹, Manahel Ansari^{1*}, Maryam Tanveer¹, and Fahad Alturise^{2*}

¹Department of Computer Science, The University of Faisalabad, Pakistan

²Department of Cybersecurity, Qassim University, Saudi Arabia

*Fahad Alturise.falturise@qu.edu.sa, Manahel Ansari.manahelansari101@gmail.com

ABSTRACT

Cardiac disease is one of the main reason of death globally and it is best to catch it early. Risk prediction becomes relevant to support timely clinical interventions. Predicting heart disease is, however, difficult because there is a complex relationship between physiological, symptomatic and demographic factors. While predictive modeling approaches have been used to improve prediction power, many current models lack interpretability and are not routinely tested for robustness, fairness, and prediction reliability. This paper introduces the leakage-free five-fold Out-of-Fold (OOF) stacked ensemble model called RXTG-Stack by leveraging RandomForest (RF), XGBoost (XGB), TabNet (TN) and GradientBoosting (GB), with LogisticRegression (LR) as a meta-learner. The framework includes Shapley Additive exPlanations (SHAP) for both global and local explanations with the aim of increasing model interpretability. To enhance predictive capability, feature engineering was performed by deriving clinically meaningful variables, including Body Mass Index (BMI), Pulse Pressure (PP), Mean Arterial Pressure (MAP), heart-rate reserve (HRR), rate-pressure product (RPP), and the cholesterol-age ratio (Chol_Age_Ratio). In addition, an empirical verification framework assesses local robustness under input perturbation, feature monotonicity, SHAP consistency, demographic fairness, and split-conformal prediction coverage. The proposed approach was evaluated on two independent cardiovascular disease datasets. On the clinically curated dataset (CVD-1K), it achieved 98.5% accuracy, 0.9994 ROC-AUC, 0.969 Cohen's Kappa (κ), and 0.969 Matthews Correlation Coefficient (MCC). On the larger population-scaled data set (Cardio-68K), it reached an accuracy of 77.0%, ROC-AUC of 0.803, MCC of 0.539 and Cohen's Kappa score of 0.539. SHAP analysis consistently identified the slope of the peak exercise ST-segment, chest pain type, resting blood pressure, and serum cholesterol as influential predictors. The empirical verification analyses indicated prediction stability of 94.8–99.3% under the tested $\pm 3\%$ relative perturbation regime, a demographic parity difference of at most 0.026, and split-conformal coverage of 0.900 on the CVD-1K dataset and 0.899 on the Cardio-68K dataset against a target coverage of 0.90. Overall, the results suggest that RXTG-Stack not only achieves competitive prediction performance but also provides interpretable predictions and complementary verification analyses for cardiac disease prediction.

Keywords: Heart Disease Prediction, Explainable AI, Machine Learning, Empirical Verification, Conformal Prediction, SHAP, Stacked Ensemble, Feature Engineering

1 Introduction

Cardiovascular disease claims approximately 17.9 million lives annually, accounting for almost one-third of all deaths globally, and the latest global data from the World Health Organization indicates that the death rate has been rising over time, despite advances in diagnostic technology¹. Although traditional methods of diagnosis are popular like clinical assessment, ECG, and Echocardiography, they can be time-consuming, costly, and even prone to human error². In many cases, several people in the world will have died of a condition that could have been prevented. These were identified years ago based on a blood pressure cuff, a cholesterol test and an ECG strip. The persistent gap is not a lack of data; hospitals, wearables, and public datasets now provide more cardiovascular risk data than ever³. It is the difficulty of turning that data into predictions that are simultaneously accurate, trustworthy, and clinically actionable.

In the era of big and complex medical data, machine learning (ML) and deep learning (DL) techniques have emerged as valuable tools for handling such information. There are a variety of ML algorithms that are available. It has been successfully used for predicting heart disease. Support Vector Machines (SVMs), and In particular, the Radial Basis Function (RBF) kernel has been shown to perform well for modelling nonlinear decision boundaries⁴. Probabilistic models like Naïve Bayes (NB) are

computationally efficient and simple, while models like Nearest Centroid and Quadratic Discriminant Analysis (QDA) include more complex models, provide meaningful results for particular classification tasks⁵.

Several models combined together is called ensemble learning, which has been shown to improve the accuracy and power of prediction. Yet, there are a major difficulty in the fact that these however, models are not always interpretable by the clinicians who are required to be told how the model makes predictions in order to trust and use it in practice. The use of ensemble learning techniques, like RF, GB, and XGB, has also enhanced the predictive power and stability of the models, leveraging the strengths of several decision trees while minimizing the risk of overfitting⁶.

The potential of intelligent predictive models for more accurate and robust heart disease diagnosis has been further illustrated by emerging techniques like clustering-based SVM with Gaussian Mixture Models (GMM) and semi-supervised XG Boost is enhanced with the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance⁷. Current research tends to be mainly concerned with maximizing predictive accuracy with limited Understanding of how predictions are made. This opacity diminishes the trust clinicians have in AI-driven decision-making and hinders the actual implementation of such tools. models in health care settings. Thus, there is an increasing demand for predictive models that are not only accurate for diagnosis but also explainable ations and comprehensive validation to support trustworthy clinical decision-making.

The goal of this study is to create an accurate, interpretable, and trustworthy heart disease prediction system using advanced ensemble learning and Explainable Artificial Intelligence (XAI). Specifically, the research aims to integrate several complementary machine learning and deep learning models into a unified predictive model and to explain the model's decisions globally and individually with SHAP^{8,9}. The proposed system is also empirically assessed to evaluate its stability, fairness, and prediction-set reliability make it a valuable tool for early and clinically actionable prediction of heart disease¹⁰.

Two cardiovascular disease datasets are used to perform a comprehensive evaluation of the proposed framework to ensure reliable performance assessment. The Cardio-68 K dataset¹¹, which has been widely used in past Studies as a benchmark for heart disease prediction, contains approximately 70,000 patient records and also includes other patient information like BMI, smoking history, and physical activity. The CVD-1 K dataset¹² consists of clinical attributes, such as significant cardiovascular risk factors like age, bloodpressure, cholesterol level, and other diagnostic attributes that are typically linked to heart disease. Both benchmarks allow a thorough assessment of the predictive performance of both across a wide range of clinical features.

This study proposes RXTG-Stack, a four-base-learner Out-of-Fold (OOF) stacked ensemble architecture. The proposed framework integrates the predictive strengths of four diverse learning algorithms, including RF, GB, XGB, and TN. Instead of directly combining predictions, the proposed architecture employs OOF stacking, where predictions generated during cross-validation are used as inputs to a LR meta-learner. This strategy minimizes information leakage while enabling the meta-model to learn the optimal combination of the four base learners. The proposed ensemble is evaluated against eight baseline models implemented under the same preprocessing pipeline and experimental settings. The eight baseline models include LR, KNN, NB, SVM, RF, GB, XGB, and TN.

To enhance transparency, the study incorporates SHAP-based explainability, providing global feature importance analysis, local explanations for individual patient predictions using waterfall and force plots, and SHAP dependence plots to identify clinically meaningful relationships. The proposed framework undergoes a comprehensive five-part empirical verification protocol, assessing local robustness under input perturbation, monotonicity consistency, SHAP feature-ranking stability, demographic fairness, and split-conformal prediction coverage for uncertainty estimation. These evaluations ensure that the proposed model is not only accurate but also reliable, interpretable and appropriate for clinical decision support systems.

The rest of this study is structured as follows. In Section 2, related work is reviewed for the prediction of CVD, ensemble learning and explainable/verifiable AI. The eight baseline models are described in Section 3. In Section 4, the proposed OOF-Stacked ensemble is introduced and explained as a novel design compared to previous designs. In Section 5, the experimental setup is described, which consists of both datasets, preprocessing, evaluation metrics, cross-validation, and the SHAP methodology. Experimental results are reported in Section 6 for both datasets, covering cross validation, visual analysis and explanation of the patterns observed. It also presents the empirical verification suite. The Section 7 ends with limitations and future work.

2 Literature Review

In the last few years, there has been a lot of improvement in the prediction of heart disease, transitioning from classic machine learning to deep learning, hybrid, ensemble and explainable artificial intelligence (XAI). This trend is fuelled by the increasing amount of healthcare data, advancements in computing power, and the need for precise, understandable, and clinically actionable prediction models.

Early research for 2022 was mainly classical machine learning algorithms for cardiovascular disease prediction. Saboor et al.¹³ developed a machine learning approach to enhance the accuracy of the prediction, by using the effective feature selection and data preprocessing methods to achieve better results, and solve the problem of dataset imbalance. Similarly, in a study by Boukhatem et al.¹⁴ they have shown that the traditional classifiers like RF, SVM and KNN can successfully predict heart

disease. These studies established a solid foundation and emphasized the need for the selection of adequate algorithms and preprocessing for accurate cardiovascular risk prediction.

In 2023, research focused on improving the predictive performance using optimization techniques, feature engineering, and ensemble learning. Bhatt et al.¹⁵ noted that the optimized Support Vector Machines and Multi-Layer Perceptrons (MLPs) had higher prediction accuracy with the help of optimized feature engineering and parameter optimizations.

Yadav et al.¹⁶ and Arumugam et al.¹⁷ also found that feature selection and pre-processing greatly affects the classification accuracy. Further, Wong et al. [18] showed that using machine learning algorithms in combination with optimization techniques could help improve the image recognition process and enhanced the prediction accuracy and the stability of the model.

During that time, hybrid learning models became popular as well. Qadri et al.¹⁹ propose chaos game optimization using recurrent neural networks to improve sequential medical data prediction. In the same way, Alam and Muqem²⁰ also reported the same improvement using a hybrid optimization strategy, and improved sequential medical data prediction. In addition, Almazroi et al.²¹ applied deep neural network architectures to medical diagnosis systems and demonstrated the potential of AI-based automated heart disease diagnosis in real-world clinical settings. Conversely, these studies were a step towards the departure from single-model prediction systems to optimized and ensemble-based prediction systems.

In 2024, there were significant developments in hybrid deep learning architectures and XAI. Mandava and Reddy Vinta²² used convolutional neural networks (CNNs) equipped with residual connections for enhancing feature extraction and cardiovascular disease (CVD) classification in the context of image analysis. Similarly, Khan et al.²³ introduced an ensemble deep learning framework called EnsCVDD Net and BICVDD-Net to attain high multiclass classification accuracy while retaining the model interpretability.

The use of personalized health indicators in cardiovascular disease prediction systems is more common. Patient-specific risk factors were shown to significantly enhance early detection of cardiovascular disease²⁴. Meanwhile, quantum-enhanced machine learning became an area of special interest. It has been proved by Babu et al.²⁵ that the prediction accuracy and computational efficiency could be enhanced using quantum-inspired learning methods. Similar enhancements have been observed for quantum convolutional neural networks, designed by Pitchal et al.²⁶ and Lenin and Venkatasalam²⁷.

One important development in the field of cardiovascular disease prediction was the use of Explainable Artificial Intelligence. Mienye and Jere²⁸ applied SHAP-based interpretation techniques to ensemble learning models to ensure the medical staff comprehended the contribution of the individual features in the prediction, and to boost their confidence in the AI diagnosis. Zhou et al.²⁹ also did a thorough survey of recent deep learning architectures and pinpointed research directions for future work on predicting cardiovascular diseases.

A study released in 2025 highlighted the trend towards hybrid intelligence, federated learning, wearable health monitoring devices, and lightweight deep learning models in cardiovascular disease prediction for real-world applications. Lohachab and Kumar³⁰ introduced federated deep learning for decentralized heart failure prediction. This method provided both distributed and scalable clinical decision making, while resolving privacy issues.

Hybrid approaches using a combination of traditional machine learning methods and quantum-inspired optimization algorithms also showed promising results, as demonstrated by Kumar et al.³. Likewise, Sekar and Aruchamy³¹ proposed the AITH2O-SANFIS hybrid classification approach which obtained close to the optimum classification accuracy and revealed the effectiveness of the integration of intelligent optimization methods with advanced neuro-fuzzy systems.

Another major field of research was the wearable healthcare systems. Sivasubramaniam and Balamurugan² combined wearable sensing with deep learning algorithm to enable continuous and real-time cardiovascular monitoring. These systems have the potential to provide significant benefits in detecting abnormal physiological patterns at an early stage and providing early clinical interventions. Dhandapani et al.³² created ensemble models for ECG image classification with low weight. They were able to improve the diagnostic accuracy and reduce the complexity of computation, making it more applicable in healthcare settings with limited resources and portability.

Rohan et al.³³ and Rabbi et al.³⁴ performed extensive benchmarking studies on a wide variety of artificial intelligence techniques and came up with the most reliable techniques in terms of accuracy, robustness and generalization power. Additionally, the combination of dimensionality reduction methods, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), with ensemble learning models was demonstrated to boost the classification accuracy while minimizing the number of redundant features and improving the generalizability of the models.

Several studies published in 2026 further reinforced the transition toward interpretable ensemble learning and trustworthy AI for heart disease prediction. Nahar et al.³⁵ suggested a two layered stacked ensemble framework that uses SVM, KNN, NB and Decision Tree (DT) as base learners and LR as the meta-learner. Five published cardiovascular datasets were pooled and used to evaluate the model and report classification accuracy of 93.7%.

The study not only focused on predicting performance, but also on transparency, interpretability, and auditability of models. Several explainability methods such as SHAP, LIME, ELI5, and counterfactual explanations were used to give the global and local explanations for the model's predictions. These techniques allowed the researchers and health care professionals to take a

closer look at the general behavior of the predictive model and the factors that led to the different diagnoses.

In the same way, the stacking-based ensemble was proposed by Chen et al.³⁶ along with the interpretability using SHAP. They determined that the most important predictors of cardiovascular disease were ST-segment slope and chest pain type. These observations corroborate the feature-importance rankings derived in the present study, which demonstrates the clinical importance of these variables. This consistency also reinforces the expansion of available research to develop accurate but also interpretable, transparent and clinically reliable predictive models.

However, a significant amount of the current research focuses on enhancing the classification results instead of in-depth model validation and clinical reliability. Relatively few works simultaneously address predictive accuracy, interpretability, robustness, fairness, and uncertainty estimation within a unified framework. These limitations motivate the development of more trustworthy and explainable prediction systems, such as the proposed RXTG-Stack framework, which combines high predictive performance with SHAP-based explainability and comprehensive empirical verification for practical clinical decision support.

3 Baseline Models

To compare the proposed ensemble’s performance, eight baseline classifiers spanning the major model families used in this literature were trained on an identical train/test split for each dataset.

Logistic Regression is a linear baseline valued for interpretability and computational efficiency, and a common benchmark throughout the cardiovascular ML literature³⁷.

K-Nearest Neighbors is a simple, non-parametric instance-based method included for its interpretability and its very different inductive bias from tree- and gradient-based models^{38,39}.

Naïve Bayes is a probabilistic model that assumes conditional feature independence; it is fast to train and provides a useful lower bound for comparison⁴⁰.

Support Vector Machine with an RBF kernel is effective at capturing non-linear decision boundaries in clinical data, at higher computational cost than linear models, and has been widely used in this domain⁴¹.

Random Forest is a bagged ensemble of decision trees⁴²; it is resistant to overfitting and a strong baseline on structured medical data⁴³.

Gradient Boosting builds trees sequentially, each correcting the errors of the last, and is historically one of the strongest tree-based methods on tabular clinical data^{44,45}.

XGBoost is a regularized, scalable implementation of gradient boosting⁴⁶, included both as a strong standalone baseline and as one of the four base learners in the proposed ensemble.

TabNet is an attention-based deep-learning architecture purpose-built for tabular data, using sequential decision steps and sparse (entmax) feature masks, and has previously demonstrated strong accuracy-interpretability trade-offs in structured prediction tasks such as insurance risk modeling^{47,48}.

All eight baselines were trained on identical scaled/encoded features and evaluated on an identical held-out test split within each dataset.

4 Proposed Model: RXTG-Stack

RXTG-Stack is a two-layer stacking architecture. Layer 1 consists of four base learners chosen for complementary inductive biases: RF (bagged decision trees, robust to noise), XGB (sequential boosting, strong on tabular structure), TN (attention-based sequential feature selection, captures non-linear deep interactions), and GB (an independent boosting family for redundancy). Each base learner is trained using 5-fold stratified cross-validation on the training set; the out-of-fold (OOF) predicted probabilities from each fold are collected so that every training patient receives a held-out prediction from each base learner, preventing leakage into the meta-model. Test-set predictions are averaged across the five per-fold models for each base learner.

Layer 2 consists of a Logistic Regression meta-learner trained on the OOF probability outputs of the four base learners, namely RF, XGB, TN, and GB, represented as P_{RF} , P_{XGB} , P_{TabNet} , and P_{GB} . Formally, for a patient i , the meta-learner produces the final probability of disease as

$$\hat{p}_i = \sigma \left(\beta_0 + \beta_1 P_{RF}^{(i)} + \beta_2 P_{XGB}^{(i)} + \beta_3 P_{TabNet}^{(i)} + \beta_4 P_{GB}^{(i)} \right), \quad (1)$$

where $\sigma(z) = 1/(1 + e^{-z})$ is the logistic (sigmoid) function and $\beta_0, \dots, \beta_4$ are the learned meta-coefficients. The stacked-generalization formulation of (1) follows Wolpert⁴⁹, with the logistic-regression meta-learner defined as in⁵⁰. The learned meta-model assigns positive coefficients to all four base learners, as summarized in Table 1. The positive contribution of each coefficient indicates that every base learner provides complementary and discriminative predictive information to the final decision. This demonstrates that the stacked ensemble effectively exploits the strengths of all constituent models, rather than

relying predominantly on a single classifier, thereby improving the diversity, robustness, and overall predictive performance of the proposed framework.

Table 1. Learned Logistic-Regression Meta-Learner Coefficients for the Four Base Learners on Both Datasets.

Meta-feature	CVD-1K	Cardio-68K
P_{RF} (Random Forest)	2.28	0.13
P_{XGB} (XGBoost)	1.42	1.35
P_{TabNet} (TabNet)	2.50	1.83
P_{GB} (Gradient Boosting)	2.86	1.47
Intercept (β_0)	-4.68	-2.36

4.1 Novelty of the Proposed Model

The proposed framework differs from prior cardiovascular prediction models in four principal respects.

1. **Leakage-free stacking:** Meta-features are generated exclusively from out-of-fold predictions, so the meta-learner never observes a base learner’s prediction on a sample that the same base learner was trained on. This eliminates the optimistic bias common to naive stacking implementations.
2. **Heterogeneous four-learner diversity:** By combining a bagging model (RF), two boosting families (XGB and GB), and an attention-based deep tabular model (TB), the ensemble draws on genuinely different inductive biases rather than variants of a single algorithm.
3. **Unified explainability and verification:** SHAP-based global and local explanations are coupled with a five-part empirical verification suite (robustness, monotonicity, SHAP consistency, fairness, and split-conformal coverage), addressing accuracy, transparency, and trustworthiness within a single pipeline.
4. **Cross-scale validation:** The identical pipeline is evaluated on a small clinically-curated dataset and a large population-scale survey dataset, demonstrating that the framework generalizes across substantially different data-quality regimes.

The complete architecture of the proposed RXTG-Stack framework, including the preprocessing stage, the four out-of-fold base learners, the Logistic Regression meta-learner, and the explainability and verification modules, is illustrated in Fig. 1.

4.2 Mathematical Formulation of the Proposed RXTG-Stack Framework

Let the cardiovascular disease dataset be represented as

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N, \quad (2)$$

where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the feature vector of the i^{th} patient after preprocessing and feature engineering, and $y_i \in \{0, 1\}$ represents the absence or presence of cardiovascular disease.

The proposed framework employs four complementary base learners,

$$f_1 = \text{RF}, \quad f_2 = \text{XGB}, \quad f_3 = \text{TN}, \quad f_4 = \text{GB}, \quad (3)$$

trained using five-fold stratified Out-of-Fold (OOF) stacking. The OOF probability generated by the m^{th} base learner for the i^{th} training sample is

$$p_{im}^{\text{OOF}} = f_m^{(-k(i))}(\mathbf{x}_i), \quad (4)$$

where $f_m^{(-k(i))}$ denotes the model trained without the fold containing sample i .

The resulting meta-feature vector is

$$\mathbf{p}_i = [p_{i1}^{\text{OOF}}, p_{i2}^{\text{OOF}}, p_{i3}^{\text{OOF}}, p_{i4}^{\text{OOF}}]. \quad (5)$$

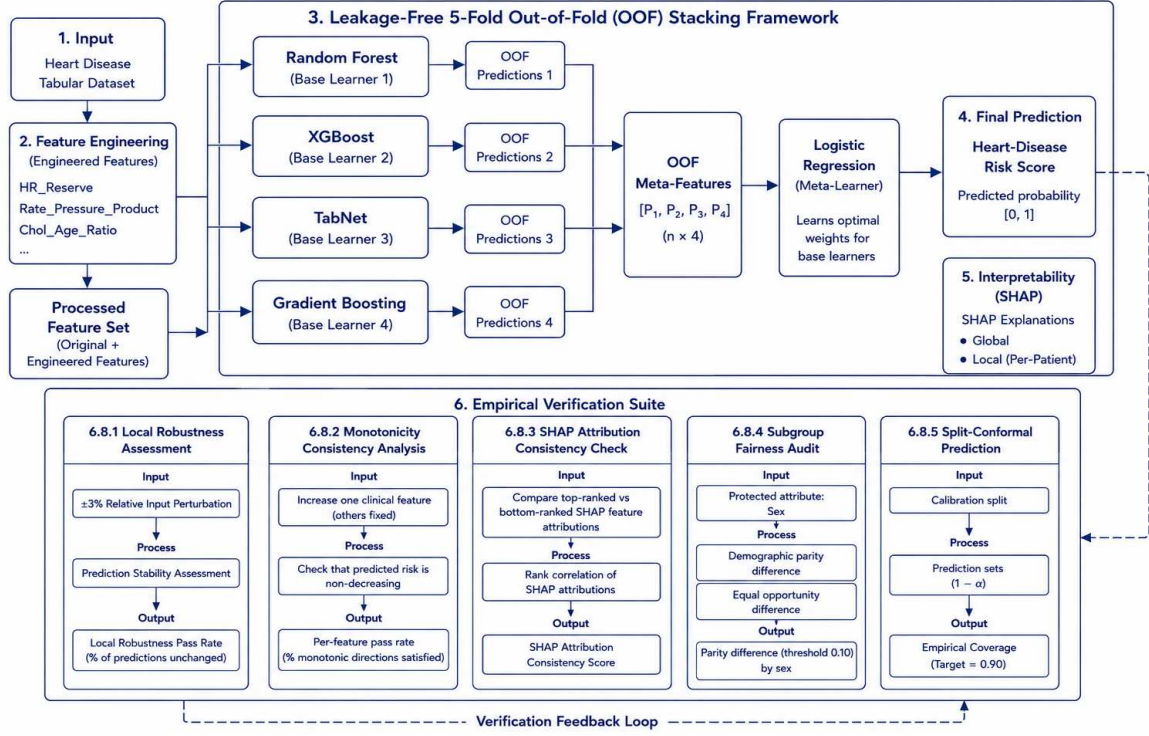


Figure 1. Overall architecture of the proposed RXTG-Stack framework, showing the preprocessing stage, the four out-of-fold base learners (Random Forest, XGBoost, TabNet, and Gradient Boosting), the Logistic Regression meta-learner, and the SHAP explainability and empirical verification modules.

The LR meta-learner combines the four OOF probabilities to estimate the final probability of cardiovascular disease:

$$\hat{p}_i = \sigma \left(\beta_0 + \sum_{m=1}^4 \beta_m p_{im}^{\text{OOF}} \right), \quad (6)$$

where

$$\sigma(a) = \frac{1}{1 + e^{-a}} \quad (7)$$

is the sigmoid activation function.

The final class prediction is obtained as

$$\hat{y}_i = \begin{cases} 1, & \hat{p}_i \geq \tau, \\ 0, & \hat{p}_i < \tau, \end{cases} \quad (8)$$

where $\tau = 0.5$ is the decision threshold.

To provide interpretable predictions, SHAP values from the individual base learners are aggregated according to the meta-learner coefficients. If $\phi_{jm}^{(i)}$ denotes the SHAP value of feature j for base learner m , the ensemble SHAP contribution is

$$\Phi_j^{(i)} = \sum_{m=1}^4 \beta_m \phi_{jm}^{(i)}. \quad (9)$$

The local prediction explanation is expressed as

$$g(\hat{p}_i) = \Phi_0 + \sum_{j=1}^d \Phi_j^{(i)}, \quad (10)$$

where $g(\cdot)$ denotes the logit function.

To evaluate robustness, perturbed samples are generated as

$$\mathbf{x}'_i = \mathbf{x}_i + \boldsymbol{\delta}, \quad |\delta_j| \leq \varepsilon |x_{ij}|, \quad (11)$$

where $\varepsilon = 0.03$. Prediction stability is defined as

$$R_i = \mathbb{I} [\hat{y}(\mathbf{x}'_i) = \hat{y}(\mathbf{x}_i)], \quad (12)$$

and the overall robustness rate is

$$R = \frac{1}{N} \sum_{i=1}^N R_i. \quad (13)$$

For clinically established risk factors, monotonicity requires

$$x'_{ij} \geq x_{ij} \implies \hat{p}(\mathbf{x}'_i) \geq \hat{p}(\mathbf{x}_i), \quad (14)$$

and the corresponding monotonicity pass rate is

$$M_j = \frac{1}{N} \sum_{i=1}^N \mathbb{I} [\hat{p}(\mathbf{x}'_i) \geq \hat{p}(\mathbf{x}_i)]. \quad (15)$$

Demographic fairness is measured using demographic parity difference

$$\Delta_{\text{DP}} = |P(\hat{Y} = 1 | A = 0) - P(\hat{Y} = 1 | A = 1)|, \quad (16)$$

and equal opportunity difference

$$\Delta_{\text{EO}} = |P(\hat{Y} = 1 | Y = 1, A = 0) - P(\hat{Y} = 1 | Y = 1, A = 1)|. \quad (17)$$

Prediction uncertainty is quantified using split-conformal prediction. For a target coverage of $1 - \alpha$, the conformal prediction set is

$$C(\mathbf{x}) = \{y \in \{0, 1\} : s(\mathbf{x}, y) \leq q_{1-\alpha}\}, \quad (18)$$

which satisfies

$$P(Y \in C(\mathbf{X})) \geq 1 - \alpha. \quad (19)$$

Algorithm 1 Proposed RXTG-Stack Framework

Require: Preprocessed training set $(X_{\text{train}}, y_{\text{train}})$, test set $(X_{\text{test}}, y_{\text{test}})$, number of stratified folds $K = 5$, and decision threshold $\tau = 0.5$.

Ensure: Final prediction probabilities $\hat{\mathbf{p}}_{\text{final}}$, predicted class labels $\hat{\mathbf{y}}_{\text{final}}$, SHAP explanations, verification outcomes, and performance metrics.

- 1: Initialize the base-learner set $\mathcal{F} = \{f_{\text{RF}}, f_{\text{XGB}}, f_{\text{TN}}, f_{\text{GB}}\}$, corresponding to Random Forest, XGBoost, TabNet, and Gradient Boosting.
- 2: Partition the training set into K stratified folds while preserving the class distribution.
- 3: Initialize the OOF training meta-feature matrix Z_{train} and the test meta-feature matrix Z_{test} .
- 4: **for** each base learner $f_m \in \mathcal{F}$ **do**
- 5: Initialize fold-wise test probability predictions for f_m .
- 6: **for** each fold $k = 1, \dots, K$ **do**
- 7: Use fold k as the validation subset and the remaining $K - 1$ folds as the fold-training subset.
- 8: Train f_m on the fold-training subset.
- 9: Generate probability predictions for the held-out validation subset.
- 10: Store the validation probabilities in the corresponding column of Z_{train} .
- 11: Generate probability predictions for X_{test} and retain the fold-level outputs.
- 12: **end for**
- 13: Average the K fold-level test probability vectors.
- 14: Store the averaged probabilities in the corresponding column of Z_{test} .
- 15: **end for**
- 16: Train the Logistic Regression meta-learner f_{LR} using $(Z_{\text{train}}, y_{\text{train}})$.
- 17: Apply f_{LR} to Z_{test} to obtain the final probability vector $\hat{\mathbf{p}}_{\text{final}}$.
- 18: Assign $\hat{y}_i = 1$ if $\hat{p}_i \geq \tau$; otherwise, assign $\hat{y}_i = 0$.
- 19: Generate global SHAP explanations to rank the overall importance of the clinical features.
- 20: Generate local SHAP explanations to identify the positive and negative feature contributions for individual patient predictions.
- 21: Evaluate local robustness, feature monotonicity, SHAP consistency, demographic fairness, and split-conformal prediction coverage.
- 22: Compute Accuracy, Precision, Recall, F1-score, ROC-AUC, Cohen’s Kappa, and Matthews Correlation Coefficient.
- 23: **Return** $\hat{\mathbf{p}}_{\text{final}}$, $\hat{\mathbf{y}}_{\text{final}}$, global and local SHAP explanations, verification outcomes, and performance metrics.

5 Experimental Setup

Two datasets were selected to evaluate the proposed framework across different scales while maintaining the same binary classification objective, namely the prediction of the presence or absence of cardiovascular disease. The first dataset, referred to as CVD-1K¹², consists of 1,000 patient records with clinically measured attributes. The second dataset, Cardio-68K¹¹, is a large-scale Kaggle cardiovascular disease dataset containing 70,000 patient records. This larger dataset was included to evaluate the scalability, robustness, and generalizability of the proposed model across a substantially larger and more diverse patient population. Together, these datasets provide a comprehensive assessment of the proposed framework. A summary of both datasets after preprocessing is given in Table 2, and a full description of their features is provided in Table 3.

Table 2. Summary of the Two Cardiovascular Datasets After Cleaning.

Characteristic	CVD-1K	Cardio-68K
Source reference	¹²	¹¹
Raw records	1,000	70,000
Records after cleaning	1,000	68,589
Records removed	0 (0%)	1,411 (2.0%)
Features (after eng.)	15	14
Positive class rate	0.58	0.495
Training samples	800	54,871
Test samples	200	13,718
Data type	Clinical	Survey

Table 3. Description of the Features Used in the CVD-1K and Cardio-68K Datasets, Including Clinically-Motivated Engineered Features.

Dataset	Feature	Description	Type / Unit
CVD-1K	age	Patient age	Numeric (years)
	gender	Biological sex (0/1)	Categorical
	chestpain	Type of chest pain (0–3)	Categorical
	restingBP	Resting blood pressure	Numeric (mmHg)
	serumcholesterol	Serum cholesterol	Numeric (mg/dL)
	fastingbloodsugar	Fasting blood sugar >120 mg/dL	Binary (0/1)
	restingelectro	Result of resting electrocardiographic	Categorical (0–2)
	maxheartrate	Maximum heart rate achieved	Numeric (bpm)
	exerciseangia	Exercise-induced angina	Binary (0/1)
	oldpeak	ST depression induced by exercise	Numeric
	slope	Slope of the peak exercise ST segment	Categorical (1–3)
	noofmajorvessels	Number of major vessels (0–3)	Numeric
	hr_reserve [†]	Heart-rate reserve,	Numeric (engineered)
	rate_pressure_product [†]	Rate–pressure product,	Numeric (engineered)
	chol_age_ratio [†]	Cholesterol-to-age ratio	Numeric (engineered)
Cardio-68K	age	age of patient (converted from days)	Numeric (years)
	gender	Male and Female (1/2)	Categorical
	height	Body height	Numeric (cm)
	weight	Body weight	Numeric (kg)
	ap_hi	Systolic blood pressure	Numeric (mmHg)
	ap_lo	Diastolic blood pressure	Numeric (mmHg)
	cholesterol	Cholesterol level (1–3)	Ordinal
	gluc	Glucose level (1–3)	Ordinal
	smoke	Status of smoking	Binary (0/1)
	alco	Alcohol intake	Binary (0/1)
	active	Physical activity	Binary (0/1)
	bmi [†]	Body Mass Index,	Numeric (engineered)
	map [†]	Mean Arterial Pressure,	Numeric (engineered)
	pulse_pressure [†]	Pulse pressure, ap_hi – ap_lo	Numeric (engineered)

[†] Clinically-motivated engineered feature derived deterministically from existing columns.

5.1 Data Preprocessing

5.1.1 Data Cleaning

For Cardio-68K¹¹, physiologically implausible records were removed, including ap_hi outside 80–250 mmHg, ap_lo outside 40–200 mmHg, records where systolic pressure was not $\geq$ diastolic pressure, height outside 120–220 cm, weight outside 30–200 kg, and exact duplicate rows, together removing 1,411 of 70,000 rows (2.0%). Age, originally recorded in days, was converted to years. For CVD-1K¹², rows were checked for missing values and duplicates; none were found, so all 1,000 records were retained⁵¹.

5.1.2 Clinically-Motivated Feature Engineering

Clinically standard derived features were added to each dataset from existing columns. These add no leakage risk, since each is a deterministic transformation of variables that are already observed at prediction time⁵². For Cardio-68K, BMI, MAP, and PP are derived as follows, where (20), (21), and (22) follow the standard clinical definitions reported in^{53, 54, and 55}, respectively:

$$\text{BMI} = \frac{\text{weight (kg)}}{(\text{height (m)})^2}, \quad (20)$$

$$\text{MAP} = \text{ap_lo} + \frac{\text{ap_hi} - \text{ap_lo}}{3}, \quad (21)$$

$$PP = ap_hi - ap_lo. \quad (22)$$

For CVD-1K, HRR, RPP, and the cholesterol-to-age ratio are derived as follows, where (23) uses the age-predicted maximum heart rate and heart-rate reserve concept of^{56,57} and (24) follows the rate–pressure product definition of⁵⁸:

$$HRR = (220 - age) - maxHR, \quad (23)$$

$$RPP = \frac{restingBP \times maxHR}{100}, \quad (24)$$

$$Chol/Age = \frac{serum\ cholesterol}{age}. \quad (25)$$

5.1.3 Scaling and Splitting

Numerical features were standardized via a z -score transformation using `StandardScaler`, fit on the training split only to prevent information leakage⁵⁹:

$$x' = \frac{x - \mu_{train}}{\sigma_{train}}, \quad (26)$$

where μ_{train} and σ_{train} are the mean and standard deviation of training sets' features. Using stratified sampling on the target label, the dataset was divided 80:20 into training and test sets. Cardio-68K was split into 54,871 train / 13,718 test and CVD-1K was split into 800 train / 200 test⁶⁰. This data split allows the model to be trained effectively while providing an independent test set for evaluating its generalization capability. For the split-conformal analysis, a further calibration subset was held out from the training split, so that the calibration and evaluation samples contained disjoint patients. Cardio-68K was divided into 43,896 conformal-fitting and 10,975 calibration records, and CVD-1K into 600 conformal-fitting and 200 calibration records; in both cases the test set defined above remained untouched.

5.1.4 Evaluation Metrics

Model performance was assessed using widely adopted classification metrics to comprehensively evaluate predictive accuracy and overall classification capability^{61,62}.

1. **Accuracy:** The total percentage of correctly identified cases among all predictions is known as accuracy. Equation (27) follows the standard confusion-matrix definition of accuracy⁶¹.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (27)$$

where TP stands for True Positives, TN for True Negatives, FP for False Positives, and FN for False Negatives.

2. **Precision:** Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive, as defined in (28)⁶¹.

$$Precision (P) = \frac{TP}{TP + FP} \quad (28)$$

3. **Recall (Sensitivity):** Recall calculates the percentage of real positive cases that the model accurately detected, as defined in (29)⁶¹.

$$Recall (R) = \frac{TP}{TP + FN} \quad (29)$$

4. **F1-Score:** The F1-score, given in (30), is a fair assessment of the model that is the harmonic mean of precision and recall^{61,66}.

$$\text{F1-Score} = 2 \times \frac{P \times R}{P + R} \quad (30)$$

5. **Receiver Operating Characteristic (ROC) Curve and AUC:** The ROC curve evaluates the model by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds, following the formulation of Fawcett⁶³. The True Positive Rate is computed using (31) as

$$\text{TPR} = \frac{TP}{TP + FN} \quad (31)$$

while the False Positive Rate is calculated using (32) as

$$\text{FPR} = \frac{FP}{FP + TN}. \quad (32)$$

The AUC offers an extensive measure of the model's performance in distinguishing between positive and negative classes. Better classification is indicated by an AUC value that is closer to 1.⁶³.

6. **Confusion Matrix:** This makes a summary of the classifying outcomes through a comparison of predicted labels with the actual labels. The confusion matrix forms the basis for calculating Accuracy, Precision, Recall, and F1-score.
7. **Cohen's Kappa (κ):** Cohen's Kappa, defined in (33), measures the level of consent between the predicted and actual class labels while considering agreement occurring by chance⁶⁴.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (33)$$

where P_e is the expected agreement by chance, and P_o is the observed agreement defined for a binary problem as

$$P_e = \frac{(TP + FP)(TP + FN) + (TN + FN)(TN + FP)}{(TP + TN + FP + FN)^2} \quad (34)$$

A κ value of 1 indicates perfect agreement, whereas 0 indicates agreement equivalent to random chance.

8. **Matthews Correlation Coefficient (MCC):** The MCC is a fair evaluation metric that takes into account all four component of the confusion matrix. It is particularly suitable for binary classification problems with imbalanced class distributions. The formulation in (35) follows the original coefficient of Matthews⁶⁵ and its use as a balanced classification metric⁶⁶.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (35)$$

The MCC is between -1 to $+1$, where $+1$ shows perfect classification, 0 represents random prediction, and -1 shows complete contradiction between predicted and actual labels.

Table 4. Performance Comparison of the Proposed RXTG-Stack Against Eight Baseline Models on the CVD-1K Dataset.

Model	Acc	Prec	Rec	F1	AUC	κ	MCC
Proposed	.985	.983	.991	.987	.999	.969	.969
Grad. Boosting	.978	.976	.984	.980	.998	.956	.957
XGBoost	.977	.974	.985	.979	.998	.954	.955
Log. Regression	.975	.966	.991	.979	.998	.948	.949
SVM (RBF)	.973	.965	.987	.976	.997	.944	.945
Random Forest	.974	.973	.979	.976	.996	.948	.949
TabNet	.970	.968	.976	.972	.995	.940	.941
Naïve Bayes	.945	.934	.974	.954	.991	.886	.887
KNN	.925	.947	.922	.935	.987	.847	.847

Table 5. Performance Comparison of the Proposed RXTG-Stack Against Eight Baseline Models on the Cardio-68K Dataset

Model	Acc	Prec	Rec	F1	AUC	κ	MCC
Proposed	.770	.778	.748	.763	.803	.539	.539
XGBoost	.735	.748	.701	.724	.802	.471	.471
Grad. Boosting	.735	.747	.701	.723	.802	.469	.470
TabNet	.734	.745	.704	.724	.800	.467	.468
Log. Regression	.727	.748	.675	.710	.790	.453	.455
Random Forest	.727	.731	.709	.720	.789	.453	.453
SVM (RBF)	.732	.755	.678	.715	.788	.464	.466
Naïve Bayes	.721	.765	.629	.691	.781	.441	.448
KNN	.720	.726	.697	.711	.779	.439	.439

6 Results and Verification

6.1 The Proposed Model's comparison with baseline models

The proposed framework combines RF, XGB, GB, and TN through an OOF stacked ensemble with a LR meta-learner, allowing it to leverage the complementary strengths of multiple classifiers.

As shown in Tables 4 and 5, RXTG-Stack attained the maximum overall datasets' performance. It achieved 98.5% accuracy and an AUC of 0.999 on the CVD-1 K dataset and 77.0% accuracy and an AUC of 0.803 on the Cardio-68K dataset, outperforming the strongest baseline models on both. The stacked ensemble that was proposed achieved the best results in all of the reported metrics on both datasets.

6.2 Performance Evaluation Across Datasets

The results obtained by the proposed RXTG-Stack framework on both datasets are enumerated in Table 6. The results demonstrate a substantial difference in both datasets. The accuracy difference is around 21% between the CVD-1K and Cardio-68K datasets. This difference is primarily attributed to the nature of the datasets. The CVD-1K dataset has attributes that are clinically measured and include very discriminative information, which are very useful for accurate rate classification. By comparison, the Cardio-68K dataset is primarily based on self-report of demographic and lifestyle information, which is more variable and less informative for disease prediction, leading to lower predictive performance. Reporting results for both datasets illustrates the generalizability of the proposed framework and a realistic assessment of its performance in the real world, under varying data quality situations.

6.3 Cross-Validation Analysis

A five-fold stratified cross-validation method^{67,68} was used for further evaluation of the stability and generalization capability of the proposed RXTG-Stack framework shown in Table 7. The mean cross-validation accuracy on the CVD-1K dataset was $97.5\% \pm 1.3\%$, with each fold demonstrating high predictive accuracy. The close agreement between the cross-validation accuracy (97.5%) and the independent test accuracy (98.5%) indicates that the proposed framework is not dependent on a particular training-test split and generalizes well to unseen data.

Similarly, the proposed RXTG-Stack framework achieved a mean cross-validation accuracy of $77.0\% \pm 0.3\%$ on the Cardio-68K dataset, demonstrating consistent performance across all five folds. Fold-to-fold variation was small (± 0.3 percentage

Table 6. Performance of the Proposed RXTG-Stack Framework Across Both Datasets.

Metric	CVD-1K	Cardio-68K
Accuracy	.985	.770
Precision	.983	.778
Recall	.991	.748
F1-Score	.987	.763
ROC-AUC	.999	.803
Cohen's Kappa	.969	.539
MCC	.969	.539

points) for the evaluation set of 13,718 samples, signifying stable behaviour across resampling. These results demonstrate that the proposed RXTG-Stack framework exhibits stable learning behaviour and robust predictive performance across both benchmark-scale and large-scale cardiovascular disease datasets.

Table 7. Performance Summary of RXTG-Stack Across Both Datasets.

Metric	CVD-1K	Cardio-68K
Test Accuracy	.985	.770
Test ROC-AUC	.999	.803
one fold	.969	.766
two fold	1.000	.771
three fold 3	.975	.773
four fold	.969	.769
five fold	.963	.771
Mean $\pm$ Std.	.975 $\pm$.013	.770 $\pm$.003

6.4 Visual Analysis

Figure 2 compares Accuracy, Precision, Recall, F1-score, and ROC-AUC on the CVD-1K and Cardio-68K datasets. On the CVD-1K dataset, the proposed framework achieved 98.5% accuracy, 98.3% precision, 99.1% recall, 98.7% F1-score, and an AUC of 0.999, exceeding the best-performing baseline models while maintaining excellent overall classification performance. On the Cardio-68K dataset, the proposed framework performed most effectively, with 77.0% accuracy, 77.8% precision, 74.8% recall, 76.3% F1-score, and an AUC of 0.803, outperforming the strongest individual classifiers. These findings show that the proposed stacked ensemble consistently provides balanced and reliable predictions across both datasets.

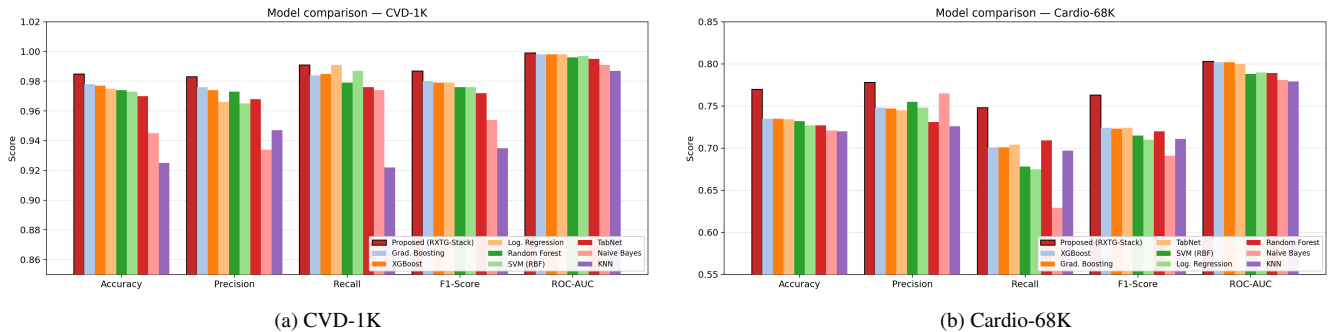
**Figure 2.** Comparisons of all models in terms of performance on (a) the CVD-1K dataset and (b) the Cardio-68K dataset. The proposed RXTG-Stack achieved the best possible score for each reported measure.

Figure 3 presents the confusion matrices for the CVD-1K and Cardio-68K datasets. On the CVD-1K dataset, the proposed framework classified 82 true negatives and 115 true positives correctly, while producing only 2 false positives and 1 false negative, indicating excellent classification performance with minimal diagnostic errors. The model accurately detected 5,481

true negatives and 5,077 true positives, while producing 1,449 false positives and 1,711 false negatives on the Cardio-68 K dataset. The prediction performance of the model was balanced across the two classes, although the number of misclassifications increased with the complexity and variability of the survey-based dataset.

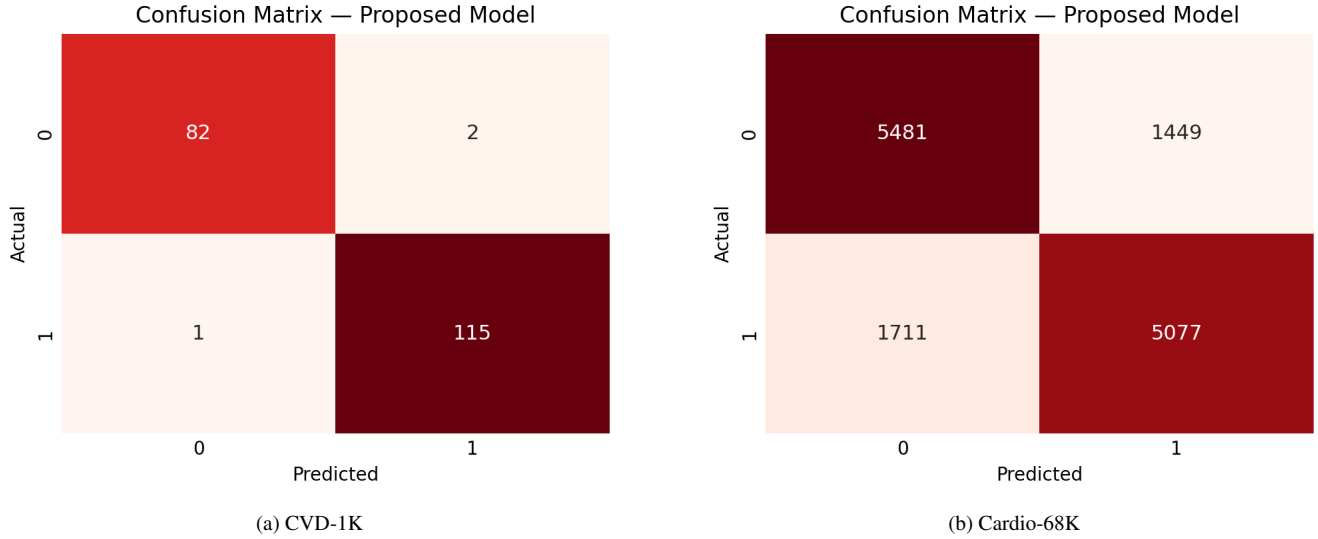


Figure 3. The proposed RXTG-Stack framework’s confusion matrices on (a) CVD-1K and (b) Cardio-68K datasets.

Figure 4 depicts the ROC curve of the proposed RXTG-Stack framework and baseline model. The proposed framework attained an AUC of 0.999 on the CVD-1K dataset, which means the model had almost perfect power to discriminate individuals with and without heart disease. On the Cardio-68K dataset, the proposed framework outperformed the other frameworks with the highest AUC score of 0.803, which is marginally higher than XGB(0.802), GB(0.802), TN (0.800), LR (0.790), RF (0.789), SVM (0.788), NB (0.781), and KNN (0.779) models. The results shown here demonstrate that the proposed model had the highest AUC value in both the datasets. It offers better discriminative power and exhibits good prediction performance on both the benchmark and population-scale cardiovascular datasets..

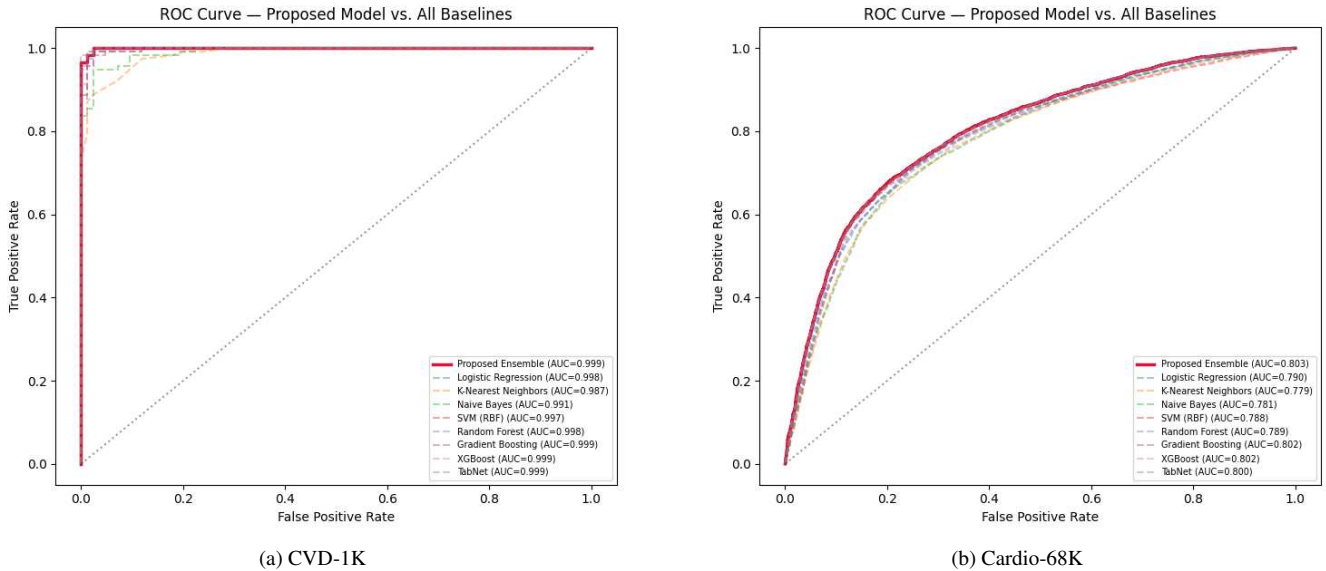


Figure 4. ROC curves and AUC values of the proposed RXTG-Stack framework and the baseline models on (a) the CVD-1K dataset and (b) the Cardio-68K dataset.

6.5 Agreement and Correlation Analysis

The baseline models were compared with the proposed RXTG-Stack model in terms of Accuracy, Matthews Correlation Coefficient (MCC), and Cohen's Kappa (κ), as shown in Table 8. On the CVD-1K dataset, RXTG-Stack achieved the highest Accuracy (98.5%), MCC (0.969), and κ (0.969). Likewise, for Cardio-68K it achieved the highest values of Accuracy (77.0%), MCC (0.539), and κ (0.539), which were higher than any of the other models. The consistently higher MCC values suggest that the proposed and actual class labels are more correlated and the higher Cohen's Kappa values indicate that the agreement between the proposed and actual class labels is more than chance. The proposed framework achieved the highest scores on all three measures, including Accuracy, MCC, and Cohen's Kappa, meaning that the ranking is not based on any one measure. This reflects successful, balanced and reliable classification performance across both cardiovascular datasets.

Table 8. Comparison of Accuracy, MCC, and Cohen's Kappa of the Proposed Model and Baselines on Both Datasets.

Model	CVD-1K			Cardio-68K		
	Acc	MCC	κ	Acc	MCC	κ
Proposed	.985	.969	.969	.770	.539	.539
XGBoost	.977	.955	.954	.735	.471	.471
Grad. Boosting	.978	.957	.956	.735	.470	.469
TabNet	.970	.941	.940	.734	.468	.467
Random Forest	.974	.949	.948	.727	.453	.453
Log. Regression	.975	.949	.948	.727	.455	.453
SVM (RBF)	.973	.945	.944	.732	.466	.464
Naïve Bayes	.945	.887	.886	.721	.448	.441
KNN	.925	.847	.847	.720	.439	.439

6.6 Explainability Analysis (SHAP)

The global feature importance for the proposed RXTG-Stack framework on the CVD-1K and Cardio-68K dataset is shown in Figure 5 and Figure 6. The mean absolute SHAP value is used to rank the features. They demonstrated that the average absolute SHAP values of every feature in the stacked ensemble learning model can reflect the magnitude and direction of the feature's influence on the model prediction of heart diseases, which makes SHAP a powerful and reliable framework for heart disease prediction from both verification and explainability perspectives⁶⁹.

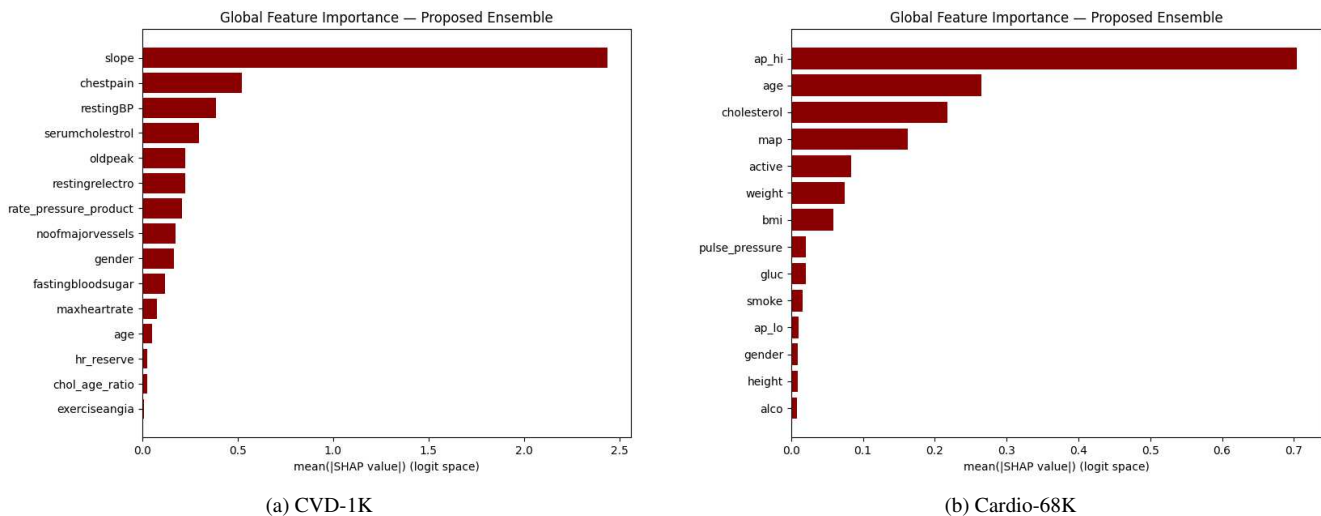


Figure 5. Global SHAP feature importance of the proposed RXTG-Stack framework on (a) the CVD-1K dataset and (b) the Cardio-68K dataset.

The most influential features for the CVD-1K dataset were Slope, Chest Pain Type, Resting Blood Pressure, Serum Cholesterol, and Old peak. These variables have the highest SHAP values, which suggests that they have a substantial influence on the prediction of heart disease. Other factors like Age, Exercise-Induced Angina and HRR were less important to the final

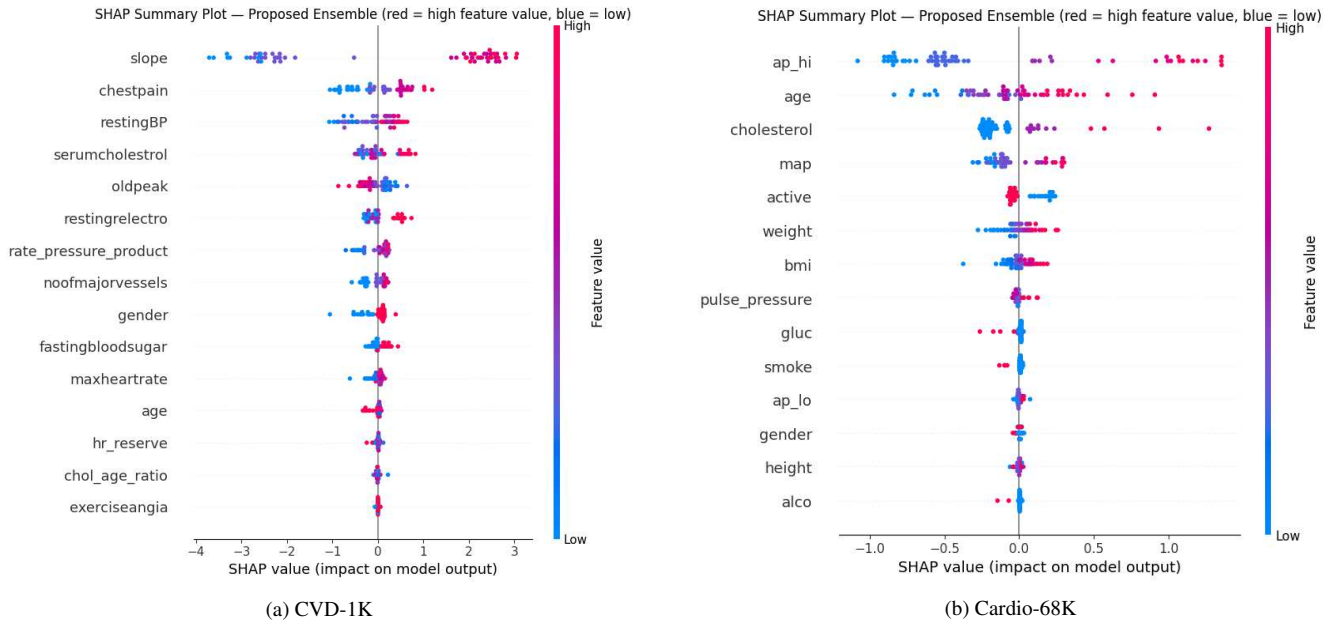


Figure 6. SHAP summary plots for the proposed RXTG-Stack framework on the (a) CVD-1K and (b) Cardio-68K datasets, illustrating global feature importance and feature contributions to model predictions

prediction. The most important variable in the Cardio-68K dataset was Systolic Blood Pressure (ap_hi), followed by Age, Cholesterol, MAP, Physical Activity, Weight, and BMI. SHAP values for variables that are related to lifestyle, such as Smoking, Alcohol Consumption, and Height, were relatively small, suggesting their relatively small contribution to the predictions of the model.

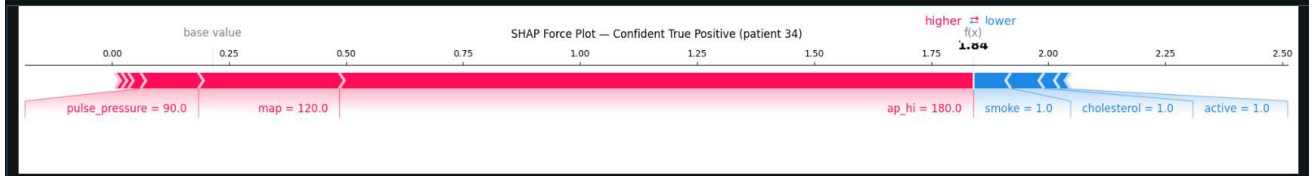
Feature rankings for the two datasets illustrate the effect of the characteristics of the datasets and differences in the underlying populations on cardiovascular disease prediction. The clinical variables are more influential in the CVD-1K dataset, such as Slope and Chest Pain Type, while physiological and demographic factors, in particular, Systolic Blood Pressure and Age, have a stronger influence in the Cardio-68K dataset. These results highlight the context-dependency of cardiovascular risk factors and illustrate the need for validation of predictive models on different datasets.

Overall, the SHAP visualizations provide both global and instance-level interpretability by demonstrating how particular features positively or negatively impact the predictions of the model. This level of transparency enhances the explainability of the proposed RXTG-Stack framework and increases its suitability for trustworthy clinical decision-support applications.

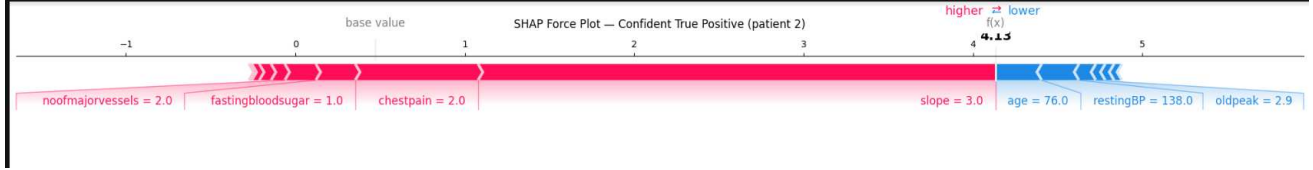
6.7 Local Interpretability Analysis

In Figure 7, SHAP force plots explain the contribution of distinct characteristics to each prediction^{9,68}. In the confident true-positive case from the Cardio-68K dataset (Patient 34), the model correctly predicts cardiovascular disease, with pulse pressure (90 mmHg), mean arterial pressure (MAP = 120 mmHg), and systolic blood pressure (ap_hi = 180 mmHg) making the largest positive contributions toward the prediction. Although smoking status (1), cholesterol level (1), and physical activity (1) contribute negatively to the model output for this instance, their effects are smaller than the positive contributions of the blood-pressure-related features, resulting in a correct positive prediction. Similarly, in the confident true-positive case from the CVD-1K dataset (Patient 2), the prediction is primarily influenced by number of major vessels (ca = 2), fasting blood sugar (fbs = 1), chest pain type (cp = 2), and ST-segment slope (slope = 3). Although age (76 years), resting blood pressure (138 mmHg), and ST depression (oldpeak = 2.9) exhibit negative SHAP contributions for this specific patient, their influence is outweighed by the stronger positive contributions of the remaining features, leading to a correct positive classification.

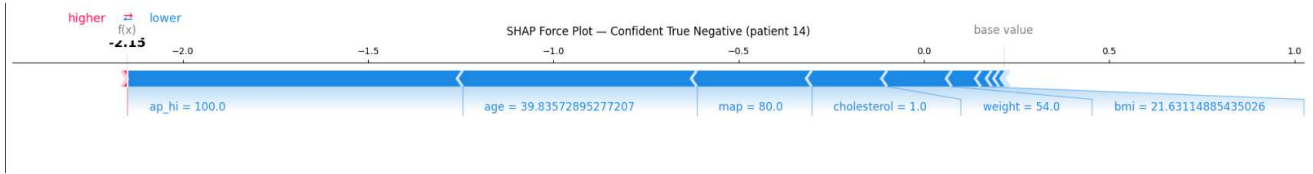
The remaining force plots illustrate the model's behavior for negative and borderline predictions. In the confident true-negative case (Patient 14), features including systolic blood pressure (100 mmHg), MAP (80 mmHg), age (39.84 years), weight (54 kg), BMI (21.63 kg/m²), and cholesterol level (1) contribute toward the negative class, resulting in a confident prediction of no cardiovascular disease. The comparable magnitude of these positive and negative SHAP values places the prediction close to the decision boundary, demonstrating that the proposed RXTG-Stack framework integrates the combined influence of multiple clinical features to produce transparent and interpretable patient-specific predictions.



(a) Patient 34 - Cardio-68K



(b) Patient 2 - CVD-1K



(c) Patient 14 - Cardio-68K

Figure 7. Representative SHAP force plots demonstrating patient-level interpretability. (a) Confident true-positive prediction from the Cardio-68K dataset. (b) Confident true-positive prediction from the CVD-1K dataset. (c) Confident true-negative prediction from the Cardio-68K dataset.

6.8 Empirical Verification Suite

Table 9 summarizes the empirical verification results of the proposed RXTG-Stack framework across the CVD-1K and Cardio-68K datasets using five complementary properties: local robustness, monotonicity consistency, SHAP attribution consistency, subgroup fairness, and split-conformal prediction^{69,70}. Except for the split-conformal component, which carries a distribution-free marginal coverage guarantee, these properties are assessed empirically on finite samples of test patients and under a specific perturbation budget; they therefore provide evidence of the stated behaviour rather than a proof over the whole input space. For the split-conformal property, the marginal coverage guarantee is satisfied by the prediction sets $C(\mathbf{x})$ ^{71,72}:

$$\Pr(y \in C(\mathbf{x})) \geq 1 - \alpha, \quad (36)$$

where $1 - \alpha$ is the target coverage level (here $1 - \alpha = 0.90$).

6.8.1 Local Robustness Assessment (Perturbation-Based)

Local robustness was assessed by perturbing each of a sample of test patients repeatedly within a relative budget of $\epsilon = 0.03$ and recording how often the predicted label changed. Under this regime the model remained stable for 99.3% of the 3,000 perturbed predictions evaluated on CVD-1K (200 patients $\times$ 15 trials) and 94.8% of the 1,500 perturbed predictions evaluated on Cardio-68K (100 patients $\times$ 15 trials), indicating resilience to measurement noise of the magnitude tested. These rates are empirical estimates over the sampled patients and the $\pm 3\%$ budget and are not claims of certified robustness over the entire input domain.

6.8.2 Monotonicity Consistency Analysis

Monotonicity consistency was examined for clinically established risk factors by increasing one feature at a time and checking whether the predicted risk was non-decreasing. Pass rates exceeded 92% for every feature examined on both datasets, and reached 1.000 for cholesterol on Cardio-68K. This is consistent with the expectation that predicted risk increases with blood pressure, cholesterol, age, BMI, and ST depression (oldpeak), in accordance with clinical knowledge. The residual violations indicate that the learned response surface is not strictly monotone, which is expected for an unconstrained ensemble.

6.8.3 SHAP Attribution Consistency Check

SHAP attribution consistency was measured as the proportion of sampled patients for which perturbing the top-ranked feature changed the predicted probability by at least as much as perturbing the bottom-ranked feature. The proposed framework obtained consistency scores of 75.0% on CVD-1K and 96.7% on Cardio-68K over 60 patients per dataset. The lower score on CVD-1K is expected because several highly informative clinical variables contribute similarly to the prediction, making their SHAP rankings interchangeable rather than indicating unreliable explanations.

6.8.4 Subgroup Fairness Audit

A subgroup fairness audit was conducted using sex as the protected attribute. The demographic parity difference was 0.026 on CVD-1K and 0.004 on Cardio-68K, and the equal opportunity difference was 0.011 and 0.001 respectively, so both remained well below the commonly used screening threshold of 0.10 on both datasets. This audit covers a single protected attribute and the available subgroups only, and does not establish fairness with respect to attributes that are not recorded in these datasets.

6.8.5 Split-Conformal Prediction

Split-conformal prediction was applied with a target coverage of 90%, using the calibration subset held out from the training split described in Section 5.1.3, which is disjoint from both the conformal-fitting data and the test set. The empirical coverage on the test set was 0.900 on CVD-1K and 0.899 on Cardio-68K, both matching the target level. Unlike the preceding four properties, the marginal coverage property of split-conformal prediction in (36) holds by construction under exchangeability, so this component provides a distribution-free finite-sample guarantee rather than an empirical estimate alone.

Collectively, these results provide evidence that the proposed RXTG-Stack framework behaved robustly under the tested perturbation regime and remained clinically consistent, fair with respect to the audited subgroups, and produced prediction sets whose empirical coverage matched the target level, in addition to being accurate.

The performance gain of RXTG-Stack over the best baseline models is relatively small, especially on the CVD-1K dataset, where multiple models reach close to the top. This behavior implies that the capacity of the model is not the limiting factor, but that the information carried by the features available is more important. The CVD-1K database has highly discriminative clinical attributes while the Cardio-68K database is mostly self-reported and demographic/lifestyle related, leading to a more similar performance in the evaluated models.

Moreover, stacking is quite generally considered as a variance reduction method, and it is not surprising that if, say, the base learners are strong, and their predictions are very strongly correlated, then there is not much to be gained from stacking. The Logistic Regression meta-learner gives positive coefficients to all the base learners, which means that each of the base learners provides complementary predictive information. This isn't all that RXTG-Stack offers, though. The framework is accurate, it is interpretable using SHAP and it is stable in prediction with the tested perturbation regime of $\pm 3\%$ demonstrating prediction stability of 94.8–99.3% on the tested datasets, the pass rates for monotonicity are excellent (more than 92%) and the difference in performance between different subgroups is at most 0.026, and the split-conformal performance on the tested datasets is 0.900 on CVD-1K and 0.899 on Cardio-68K at a target value of 0.90. These attributes show the capability to be competitive in terms of predictive performance, being transparent and providing reliable clinical decision support for datasets of significant scale and data characteristics.

Table 9. Empirical Verification Results of the Proposed RXTG-Stack Framework on Both Datasets, Under the Tested Settings ($\epsilon = 0.03$ Relative Perturbation; Target Conformal Coverage 0.90).

Verification Property	CVD-1K	Cardio-68K
Local robustness (pass rate, $\pm 3\%$)	.993	.948
Monotonicity (min. pass rate)	.943	.923
SHAP attribution consistency	.750	.967
Demographic parity difference	.026	.004
Equal opportunity difference	.011	.001
Split-conformal coverage (target 0.90)	.900	.899

7 Conclusion and Future Work

This paper proposed RXTG-Stack, a OOF stacked ensemble that combines RF, XGB, TN, and GB through a LR meta-learner for explainable heart disease prediction. The proposed framework achieved 98.5% accuracy, 98.7% F1-score, and 0.999 ROC-AUC on the CVD-1K dataset, with a 5-fold cross-validation accuracy of $97.5\% \pm 1.3\%$, while obtaining 77.0% accuracy and 0.803 ROC-AUC on the Cardio-68K dataset. Apart from its robust predictive capabilities, the framework provides SHAP-based

explainability and is supported by a comprehensive empirical verification suite comprising local robustness, monotonicity consistency, SHAP attribution consistency, subgroup fairness, and split-conformal prediction, indicating that it is accurate, interpretable, robust under the tested perturbation regime, and suitable for further evaluation in clinical decision-support applications.

Although the proposed framework shows promising results, To verify its generalizability, additional validation on bigger multi-center and real-world clinical datasets is required. It was assessed using two publicly available datasets. Future work will focus on integrating electronic health records and wearable sensor data, incorporating additional multimodal clinical features, and extending the present empirical analyses towards SMT-based formal verification together with enhanced uncertainty quantification to further improve the reliability and practical deployment of AI-assisted cardiovascular disease prediction systems.

Funding declaration

This research is funded by the Deanship of Graduate Studies and Scientific Research at Qassim University (www.qu.edu.sa) (QU-APC-2026).

Author contributions statement

T.F.K. conceived the study, designed the RXTG-Stack framework, implemented the models, performed the experiments and analysis, and wrote the main manuscript. M.A. contributed to the methodology, data preprocessing and clinically-motivated feature engineering, and prepared the figures and tables. M.T. contributed to the literature review, data curation, and validation of the results. F.A. supervised the project, provided resources, and critically revised the manuscript for important intellectual content. All authors reviewed and approved the final version of the manuscript.

Additional information

References

1. G. A. Roth *et al.*, “Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015,” *J. Amer. Coll. Cardiol.*, vol. 70, no. 1, pp. 1–25, 2017.
2. S. Sivasubramaniam and S. P. Balamurugan, “Early detection and prediction of heart disease using wearable devices and deep learning algorithms,” *Multimedia Tools Appl.*, vol. 84, no. 9, pp. 6187–6201, 2025.
3. A. Kumar, S. Dhanka, A. Sharma, R. Bansal, M. Fahlevi, F. Rabby, and M. Aljuaid, “A hybrid framework for heart disease prediction using classical and quantum-inspired machine learning techniques,” *Sci. Rep.*, vol. 15, no. 1, p. 25040, 2025.
4. J. A. K. Suykens and J. Vandewalle, *Support Vector Machines and Kernel Based Learning*. Cham, Switzerland: Springer, 2019.
5. A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Frontiers Neurobot.*, vol. 7, p. 21, Jan. 2013.
6. Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2019.
7. M. S. Kumar *et al.*, “Advancements in heart disease prediction: A comprehensive review of ML and DL algorithms,” *IEEE Access*, vol. 9, pp. 30345–30370, 2021.
8. E. Tjoa and C. Guan, “A survey on explainable artificial intelligence (XAI): Toward medical XAI,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.
9. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017, pp. 4765–4774.
10. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, 2021.
11. Sulianova. (Aug. 2024). *Cardiovascular Disease Dataset*. Kaggle. [Online]. Available: <https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>
12. B. P. Doppala and D. Bhattacharyya, “Cardiovascular disease dataset,” Mendeley Data, V1, 2021, doi: 10.17632/dzz48mvjht.1.
13. A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, “A method for improving prediction of human heart disease using machine learning algorithms,” *Mobile Inf. Syst.*, vol. 2022, p. 1410169, 2022.

14. C. Boukhatem, H. Y. Youssef, and A. B. Nassif, "Heart disease prediction using machine learning," in *Proc. Adv. Sci. Eng. Technol. Int. Conf. (ASET)*, 2022.
15. C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," *Algorithms*, vol. 16, no. 2, p. 88, 2023.
16. A. L. Yadav, K. Soni, and S. Khare, "Heart diseases prediction using machine learning," in *Proc. 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, 2023.
17. K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using machine learning algorithms," *Mater. Today, Proc.*, vol. 80, pp. 3682–3685, 2023.
18. K. K. Wong, D. N. Ghista, A. W. Ip, W. Zhang, N. Chandrasekhar, and S. Peddakrishna, "Enhancing heart disease prediction accuracy through machine learning techniques and optimization," *Processes*, vol. 11, no. 4, p. 1210, 2023.
19. A. M. Qadri, A. Raza, K. Munir, and M. S. Almutairi, "Effective feature engineering technique for heart disease prediction with machine learning," *IEEE Access*, vol. 11, pp. 56214–56224, 2023.
20. A. Alam and M. Muqeem, "An optimal heart disease prediction using chaos game optimization-based recurrent neural model," *Int. J. Inf. Technol.*, vol. 16, no. 5, pp. 3359–3366, 2023.
21. A. A. Almazroi, E. A. Aldhahri, S. Bashir, and S. Ashfaq, "A clinical decision support system for heart disease prediction using deep learning," *IEEE Access*, vol. 11, pp. 61646–61659, 2023.
22. M. Mandava and S. Reddy Vinta, "MDensNet201-IDRSRNet: Efficient cardiovascular disease prediction system using hybrid deep learning," *Biomed. Signal Process. Control*, vol. 93, p. 106147, 2024.
23. H. Khan, N. Javaid, T. Bashir, M. Akbar, N. Alrajeh, and S. Aslam, "Heart disease prediction using novel ensemble and blending based cardiovascular disease detection networks: EnsCVDD-Net and BICVDD-Net," *IEEE Access*, vol. 12, pp. 109230–109254, 2024.
24. F. M. Talaat, "Revolutionizing cardiovascular health: Integrating deep learning techniques for predictive analysis of personal key indicators in heart disease," *Neural Comput. Appl.*, vol. 37, no. 1, pp. 1–24, 2024.
25. S. V. Babu, P. Ramya, and J. Gracewell, "Revolutionizing heart disease prediction with quantum-enhanced machine learning," *Sci. Rep.*, vol. 14, no. 1, p. 7453, 2024.
26. P. Pitchal, S. Ponnusamy, and V. Soundararajan, "Heart disease prediction: Improved quantum convolutional neural network and enhanced features," *Expert Syst. Appl.*, vol. 249, p. 123534, 2024.
27. S. T. Lenin and K. Venkatasalam, "Effective classification of heart disease using convolutional neural networks," *Circuits Syst. Signal Process.*, vol. 44, no. 2, pp. 911–935, 2024.
28. I. D. Mienye and N. Jere, "Optimized ensemble learning approach with explainable AI for improved heart disease prediction," *Information*, vol. 15, no. 7, p. 394, 2024.
29. C. Zhou, P. Dai, A. Hou, Z. Zhang, L. Liu, A. Li, and F. Wang, "A comprehensive review of deep learning-based models for heart disease prediction," *Artif. Intell. Rev.*, vol. 57, no. 10, p. 263, 2024.
30. A. Lohachab and K. Kumar, "FedHFP: A federated deep learning framework for heart failure prediction," *IETE J. Res.*, vol. 71, no. 2, pp. 479–491, 2025.
31. J. Sekar and P. Aruchamy, "A novel approach for heart disease prediction using hybridized AITH2O algorithm and SANFIS classifier," *Netw., Comput. Neural Syst.*, vol. 36, no. 1, pp. 109–147, 2025.
32. S. Dhandapani, H. Somasundaram, and T. Angamuthu, "Hybrid deep learning framework for heart disease prediction using ECG signal images," *Sci. Rep.*, vol. 15, no. 1, p. 33922, 2025.
33. D. Rohan, G. P. Reddy, Y. V. Kumar, K. P. Prakash, and C. P. Reddy, "An extensive experimental analysis for heart disease prediction using artificial intelligence techniques," *Sci. Rep.*, vol. 15, no. 1, p. 6132, 2025.
34. M. S. H. Rabbi, M. M. Bari, T. Debnath, A. Rahman, A. K. Das, M. P. Hossain, and G. Muhammad, "Performance evaluation of optimal ensemble learning approaches with PCA and LDA-based feature extraction for heart disease prediction," *Biomed. Signal Process. Control*, vol. 101, p. 107138, 2025.
35. N. Nahar *et al.*, "Stacked ensemble model with explainable AI for early detection of heart disease," *Anal. Sci. Adv.*, vol. 7, no. 1, p. e70072, 2026.
36. Y. Chen, L. Chong, Z. Bao, S. Wang, Y. Wang, and Y. Feng, "An interpretability heart disease prediction model based on stacking ensemble with SHAP," *Frontiers Molecular Biosci.*, vol. 12, p. 1763157, 2026.

37. G. Ambrish, B. Ganesh, A. Ganesh, C. Srinivas, and K. Mensinkal, "Logistic regression technique for prediction of cardiovascular disease," *Global Transitions Proc.*, vol. 3, no. 1, pp. 127–130, 2022.
38. M. Shouman, T. Turner, and R. Stocker, "Applying k-nearest neighbour in diagnosing heart disease patients," *Int. J. Inf. Educ. Technol.*, vol. 2, no. 3, pp. 220–223, 2012.
39. H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 1022, no. 1, p. 012072, 2021.
40. G. Subbalakshmi, K. Ramesh, and M. C. Rao, "Decision support in heart disease prediction system using naive Bayes," *Indian J. Comput. Sci. Eng.*, vol. 2, no. 2, pp. 170–176, 2011.
41. V. V. Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: A survey," *Int. J. Eng. Technol.*, vol. 7, no. 2.8, pp. 684–687, 2018.
42. L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
43. Y. Rimal, N. Sharma, S. Paudel, A. Alsadoon, M. P. Koirala, and S. Gill, "Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy," *Sci. Rep.*, vol. 15, no. 1, p. 13444, 2025.
44. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.
45. N. L. Fitriyani *et al.*, "A novel approach utilizing bagging, histogram gradient boosting, and advanced feature selection for predicting the onset of cardiovascular diseases," *Mathematics*, vol. 13, no. 13, p. 2194, 2025.
46. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.
47. S. Ö. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 8, 2021, pp. 6679–6687.
48. K. McDonnell, F. Murphy, B. Sheehan, L. Masello, and G. Castignani, "Deep learning in insurance: Accuracy and model interpretability using TabNet," *Expert Syst. Appl.*, vol. 217, p. 119543, 2023.
49. D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.
50. S. Sperandei, "Understanding logistic regression analysis," *Biochemia Medica*, vol. 24, no. 1, pp. 12–18, 2014.
51. H. Lattar, A. B. Salem, and H. H. B. Ghezala, "Does data cleaning improve heart disease prediction?" *Proc. Comput. Sci.*, vol. 176, pp. 1131–1140, 2020.
52. I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *Social Netw. Comput. Sci.*, vol. 2, no. 3, p. 160, May 2021.
53. A. Keys, F. Fidanza, M. J. Karvonen, N. Kimura, and H. L. Taylor, "Indices of relative weight and obesity," *J. Chronic Diseases*, vol. 25, nos. 6–7, pp. 329–343, 1972.
54. E. Meaney, F. Alva, R. Moguel, A. Meaney, J. Alva, and R. Webel, "Formula and nomogram for the sphygmomanometric calculation of the mean arterial pressure," *Heart*, vol. 84, no. 1, p. 64, 2000.
55. S. S. Franklin, S. A. Khan, N. D. Wong, M. G. Larson, and D. Levy, "Is pulse pressure useful in predicting risk for coronary heart disease? The Framingham Heart Study," *Circulation*, vol. 100, no. 4, pp. 354–360, 1999.
56. S. M. Fox, J. P. Naughton, and W. L. Haskell, "Physical activity and the prevention of coronary heart disease," *Annals Clin. Res.*, vol. 3, no. 6, pp. 404–432, 1971.
57. M. J. Karvonen, E. Kentala, and O. Mustala, "The effects of training on heart rate: A longitudinal study," *Annales Medicinae Experimentalis et Biologiae Fenniae*, vol. 35, no. 3, pp. 307–315, 1957.
58. F. L. Gobel, L. A. Norstrom, R. R. Nelson, C. R. Jorgensen, and Y. Wang, "The rate–pressure product as an index of myocardial oxygen consumption during exercise in patients with angina pectoris," *Circulation*, vol. 57, no. 3, pp. 549–556, 1978.
59. F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
60. S. Chippalakatti, C. H. Renumadhavi, and A. Pallavi, "Comparative review on the machine learning algorithms for medical data," in *Proc. 6th Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solutions (CSITSS)*, Bangalore, India, Dec. 2022, pp. 1–6.
61. D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

62. K. Suzuki and T. Matsuzawa, “Model soups for various training and validation data,” *AI*, vol. 3, no. 4, pp. 796–808, Sep. 2022.
63. T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.
64. J. Cohen, “A coefficient of agreement for nominal scales,” *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
65. B. W. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” *Biochimica et Biophysica Acta (BBA)—Protein Struct.*, vol. 405, no. 2, pp. 442–451, 1975.
66. D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 1, p. 6, 2020.
67. M. Stone, “Cross-validatory choice and assessment of statistical predictions,” *J. Roy. Statist. Soc. B, Methodological*, vol. 36, no. 2, pp. 111–133, 1974.
68. S. M. Ganie, P. K. D. Pramanik, and Z. Zhao, “Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets,” *Sci. Rep.*, vol. 15, no. 1, p. 13912, 2025.
69. N. G. Rezk, S. Alshathri, A. Sayed, E. El-Din Hemdan, and H. El-Behery, “XAI-augmented voting ensemble models for heart disease prediction: A SHAP and LIME-based approach,” *Bioengineering*, vol. 11, no. 10, p. 1016, 2024.
70. M. Elmasry, M. G. Gafar, and M. A. Elsabagh, “Trustworthy educational risk modeling with calibrated probabilities, conformal uncertainty, explainable AI, and graph-based refinement,” *Inventions*, vol. 11, no. 3, p. 65, 2026.
71. V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. New York, NY, USA: Springer, 2005.
72. A. N. Angelopoulos and S. Bates, “Conformal prediction: A gentle introduction,” *Found. Trends Mach. Learn.*, vol. 16, no. 4, pp. 494–591, 2023.