The Bioscan AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES 21(2): 516-536, 2026 www.thebioscan.co

A Robust Hybrid Ensemble Approach for Cardiovascular Disease Prediction with Feature Optimization

Payel Sengupta[1], Debmalya Mukherjee[2], RANJAN BANERJEE[3], AMARTYA GHOSH[4], PRANAB GHARAI[5], RIA PYNE[6], Piyali De[7], Avijit Kumar Chaudhuri[8], Sanjib Bayen[9]

[1]0000-0003-3981-5971
payel9433@gmail.com

[2]0009-0006-9946-0964
dbml.mukherjee@gmail.com

[3]0009-0003-1950-7530
rbkpccst@gmail.com

[4]0009-0002-2504-3325
com.amartya@gmail.com

[5]0009-0004-3602-5223
pranab.g10@gmail.com

[6]0009-0007-8323-8354
pyneria@gmail.com

[7]0009-0007-2631-615X
piyalide@gmail.com

[8]0000-0002-5310-3180
c.avijit@gmail.com

[9]sanjibbayen29@gmail.com

Department: Computer Science and Engineering
University / Institution Name and State: Brainware University,Barasat, West Bengal 700125
City / Pincode: Kolkata, West Bengal 700125

The Bioscan AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES 21(2): 516-536, 2026 www.thebioscan.co

DOI: https://doi.org/10.63001/tbs.2026.v21.i02.pp516-536

Received on: 27-03-2026
Accepted on: 15-04-2026
Published on: 25-04-2026

Abstract

Cardiovascular diseases (CVDs) continue to be the world's leading cause of death, accounting for 17.9 million deaths annually. For effective prevention and intervention, accurate early prediction systems are crucial. The inability of conventional statistical risk models to capture the complex, nonlinear interactions among clinical risk factors frequently limits their predictive accuracy. To address these problems, this study evaluates optimized machine learning methods for cardiovascular disease prediction using the Framingham Heart Study dataset, a reliable longitudinal clinical resource. A comprehensive preprocessing pipeline that included clinically guided feature selection, the removal of incomplete records, feature standardization, and class imbalance correction using Borderline-SMOTE2 was implemented to enhance minority class detection. Four extreme-tuned classifiers—Random Forest, AdaBoost, Support Vector Machine (SVM), and Decision Tree—were systematically trained and compared using a number of evaluation metrics, including accuracy, ROC-AUC, sensitivity, specificity, F1-score, and Cohen's Kappa. Experimental results demonstrate that the optimized AdaBoost and Random Forest models achieved the best predictive performance, with accuracies of 94.12%. AdaBoost had the best overall discrimination, with an ROC-AUC of 0.9775, F1-score of 0.9406, and Cohen's Kappa of 0.8824. With a ROC-AUC of 0.9764, F1-score of 0.9402, and Kappa of 0.8823, Random Forest ranked second. SVM provided strong sensitivity (0.9419) but lower specificity, while the Decision Tree performed comparatively poorly. These findings confirm that ensemble-based methods provide improved stability, balanced classification, and better generalization for cardiovascular risk prediction. The proposed framework offers a clinically reliable and understandable decision-support solution for early detection of cardiovascular disease.

1. Introduction

Cardiovascular diseases (CVDs) are one of the top public health challenges globally and the main cause of death. Recent reports indicate that around 17.9 million people die each year from cardiovascular problems, accounting for nearly one-third of all global deaths [1][29]. This alarming statistic shows the urgent need for effective prevention methods and reliable early diagnostic systems. Besides the death toll, cardiovascular issues also have serious economic effects. They lead to more hospital admissions, higher long-term treatment costs, increased rehabilitation needs, and decreased workforce productivity. These challenges stress healthcare systems in both wealthy and poor countries. Additionally, rapid urban growth, sedentary lifestyles, poor diets, tobacco use, and metabolic disorders have significantly increased the number of cardiovascular risk factors in many populations [2][30]. Consequently, identifying high-risk individuals early has become a primary goal for doctors and policymakers who aim to reduce new cases and deaths from these diseases. The development of cardiovascular disease is complex and involves many factors. Unlike conditions arising from a single cause, cardiovascular disorders result from multiple demographic, behavioral, and physiological factors interacting. Age, sex, blood pressure (both systolic and diastolic), cholesterol levels, smoking exposure, diabetes status, obesity, and glucose levels all affect cardiovascular health outcomes [26][30]. Importantly, these risk factors do not act independently; they interact in intricate ways. For example, the combined effects of diabetes and hypertension can greatly

increase cardiovascular risk beyond what each factor contributes alone. These complex relationships create challenges for traditional modeling methods. More flexible analytical approaches are needed to better capture the interactions among these predictors. Historically, estimating cardiovascular risk has relied on statistical scoring systems such as the Framingham Risk Score, SCORE, and QRISK. These methods typically come from regression models that use fixed coefficients derived from population-level studies. While these models are straightforward and easy to understand, their predictive accuracy is often limited because they assume linear relationships and independence among variables. Several studies have found that these traditional risk calculators typically have only moderate accuracy, with ROC-AUC values ranging from 0.59 to 0.66 across different clinical populations [32]. As a result, many high-risk individuals may go undetected, while some low-risk individuals might receive unnecessary treatments. This misclassification can delay treatment, increase complications, and misuse resources. These limitations have led researchers to explore data-driven approaches that could improve predictive reliability. In recent years, machine learning (ML) techniques have emerged as promising alternatives to classical statistical models for medical risk prediction. Machine learning algorithms can learn directly from data, automatically recognize complex nonlinear patterns, and model high-dimensional feature spaces without strict distributional assumptions. These features make ML particularly well-suited for clinical datasets where multiple interacting variables influence disease outcomes. Comparative studies have shown that ML-based approaches often outperform traditional regression models in cardiovascular disease prediction tasks, achieving higher accuracy, improved sensitivity, and better overall discrimination [4][5][14][16]. Among these approaches, ensemble learning techniques—such as Random Forest and boosting algorithms—have proven especially effective because they combine multiple weak learners, reducing variance and improving generalization [1][2]. Consequently, ensemble models are increasingly used for the analysis of structured medical data. The availability of extensive and longitudinal clinical datasets has accelerated research in this area. Among these, the Framingham Heart Study remains one of the most influential and clinically validated epidemiological investigations in cardiovascular medicine [22] [23]. Started in 1948, the study has tracked multiple generations of participants and has significantly contributed to the identification of major cardiovascular risk factors [24] [25]. The dataset includes detailed demographic information, lifestyle characteristics, physiological measurements, and long-term health outcomes, making it highly suitable for predictive modeling and comparative experimentation. Because of its rich clinical content and historical reliability, the Framingham dataset has become a widely accepted benchmark for evaluating machine learning methods in cardiovascular disease prediction [3].

Despite the growing use of machine learning in healthcare, several methodological challenges continue to hinder reliable, reproducible performance. First, clinical datasets often have missing values due to incomplete tests, patient non-compliance, or measurement variability. Failing to handle missing data properly can distort statistical relationships and lower model accuracy. Second, cardiovascular prediction datasets often show significant class imbalance, with many more healthy individuals than disease cases. This imbalance can cause learning algorithms to favor the majority class, leading to poor sensitivity in identifying true

disease cases. Prior studies have emphasized the importance of using imbalance-handling techniques, particularly SMOTE-based oversampling methods, to improve the detection of minority classes in medical classification problems [7][12]. Specifically, Borderline-SMOTE has been effective because it generates synthetic samples close to decision boundaries, where misclassifications are most likely to occur [6]. Third, differing feature scales can negatively affect algorithms that rely on distance or gradient computations, such as Support Vector Machines. Therefore, appropriate feature normalization is necessary for fair model learning. Another common limitation in the literature is insufficient hyperparameter optimization. Many studies depend on default model settings or evaluate only a limited number of classifiers, which may prevent algorithms from reaching their full predictive potential. Hyperparameters directly affect model complexity, generalization, and robustness. Without systematic tuning, even advanced algorithms can show unstable performance across different validation folds. Comparative analyses have highlighted that thorough optimization and standardized evaluation protocols are crucial for achieving meaningful and reproducible results [5] [17]. These points underscore the need to integrate robust preprocessing, careful model selection, and systematic experimentation within a unified framework. Motivated by these challenges, this study proposes a comprehensive, reproducible, and clinically oriented machine learning framework for predicting cardiovascular disease using the Framingham dataset. Unlike approaches that focus on overly complicated stacking ensembles or deep neural networks, this work aims to maximize the performance of carefully selected individual classifiers through solid preprocessing and extensive hyperparameter tuning. The proposed preprocessing pipeline includes clinically guided feature selection, removal of incomplete records, standardized feature scaling, and class imbalance correction using Borderline-SMOTE2 to ensure balanced learning [6] [7]. This design seeks to improve both predictive accuracy and interpretability while avoiding data leakage. Four representative classifiers — (i)Random Forest, (ii)Optimized AdaBoost, (iii)Support Vector Machine, and (iv)Decision Tree—are systematically implemented and evaluated under the same experimental conditions. Performance is measured using several clinically relevant metrics, including accuracy, sensitivity, specificity, F1-score, ROC-AUC, and Cohen's Kappa, providing a comprehensive understanding of predictive ability and agreement beyond chance. Experimental findings indicate that ensemble-based approaches, particularly Random Forest and AdaBoost, exhibit greater stability and generalization than single learners. These results support earlier evidence suggesting that tree-based ensembles are highly effective for structured clinical datasets [1] [4]. The main contributions of this study are summarized as follows: (i) Development of a solid preprocessing pipeline that uses Borderline-SMOTE2 for correcting imbalances, (ii) Systematic comparative evaluation of optimized machine learning classifiers under consistent protocols, (iii) Demonstration of the superiority of ensemble tree-based approaches in predicting cardiovascular disease, and (iv) Provision of a reproducible and understandable framework suited for real-world clinical decision support systems. The remainder of the paper is organized as follows: Section 2 reviews the related literature; Section 3 describes the dataset and preprocessing strategies; Section 4 details the methodological framework and model optimization procedures; Section 5 presents experimental results and a comparative

analysis; and Sections 6 and 7 discuss clinical implications and concluding insights.

2. Related Works

Early research on cardiovascular disease (CVD) prediction primarily relied on traditional statistical and epidemiological models, including logistic regression and risk scoring systems derived from population studies. Notable among these is the Framingham Risk Score, which has been widely used to estimate long-term cardiovascular risk based on demographic and clinical attributes. While such approaches offer interpretability and clinical simplicity, their predictive performance is often limited by assumptions of linearity and independence among risk factors, resulting in moderate classification performance [1][2]. With the increasing availability of electronic health records and large-scale clinical datasets, machine learning techniques have been widely explored to overcome the limitations of conventional statistical models. Early machine learning-based studies applied algorithms such as decision trees, naïve Bayes classifiers, and support vector machines for heart disease prediction, reporting improved accuracy and sensitivity compared to traditional methods [3] [4]. However, these models often struggled with high-dimensional feature spaces, class imbalance, and sensitivity to noise, which affected their robustness in real-world clinical settings [5]. Ensemble-based learning methods, particularly tree-based models such as Random Forests and gradient boosting, have gained significant attention in recent years. Random Forest has been shown to perform well on structured medical datasets due to its ability to reduce variance through bootstrap aggregation and to model nonlinear feature interactions effectively [6][7]. Several studies have reported that Random Forests outperform single decision trees and linear classifiers in cardiovascular risk prediction tasks, achieving higher accuracy and better generalization [8] [9]. Gradient boosting models, including XGBoost and LightGBM, further enhance predictive performance by iteratively minimizing classification error and handling complex feature relationships [10] [11]. Recent research has also focused on optimizing model performance through advanced preprocessing and hyperparameter tuning strategies. Techniques such as feature scaling, missing value imputation, and synthetic sampling methods for class imbalance particularly SMOTE and its variants—have been shown to significantly improve model stability and recall for minority classes [12][13]. Studies employing optimized tree-based models combined with balanced training data consistently report superior ROC-AUC and F1-score values, highlighting the importance of data quality and model tuning in medical prediction tasks [14] [15]. Deep learning approaches, including Multi-layer Perceptrons and Neural Network architectures, have also been investigated for cardiovascular disease prediction. While neural networks demonstrate strong learning capacity, their performance on tabular clinical datasets is often comparable to, or slightly inferior to, well-tuned tree-based ensembles, especially when dataset size is limited and interpretability is a concern [16][17]. As a result, recent literature increasingly favors optimized ensemble tree models for clinical decision support applications due to their balance between accuracy, robustness, and interpretability. Despite the growing body of research, many existing studies either rely on default model configurations or focus on a narrow subset of algorithms, limiting comprehensive comparative evaluation. Furthermore, inconsistent evaluation protocols and insufficient reporting of

robustness metrics reduce the reproducibility of reported findings [18][19]. These gaps motivate the need for a systematic evaluation of multiple extreme-tuned machine learning models under consistent experimental settings. In contrast to prior works, the present study conducts a rigorous comparative analysis of several optimized machine learning classifiers using a standardized preprocessing pipeline and robust validation strategy. By emphasizing highly tuned individual models rather than complex multi-model architectures, this work demonstrates that carefully optimized Random Forest models can deliver superior, stable performance for early cardiovascular disease prediction, reinforcing their suitability for real-world clinical deployment.

3. Dataset Description

The reliability and clinical usefulness of any machine learning-based prediction framework depend on the quality, representativeness, and historical credibility of the underlying dataset. In cardiovascular disease research, longitudinal epidemiological datasets are especially valuable. They allow researchers to observe how risk factors change over time and their long-term health effects. Among these resources, the Framingham Heart Study dataset is widely regarded as one of the most reliable and validated sources for cardiovascular risk modeling. This dataset has significantly influenced modern cardiovascular medicine and continues to be a key reference for predictive analytics studies [22] [25]. This section provides an in-depth discussion of the dataset's origin, structural features, attribute composition, data quality, and preprocessing needs. This detailed description is important for demonstrating the dataset's suitability for machine learning applications and for ensuring the transparency and reproducibility of the experimental framework. The Framingham Heart Study began in 1948, led by the National Heart, Lung, and Blood Institute in collaboration with Boston University. Its primary goal was to identify factors contributing to cardiovascular disease in a general population cohort [22] [23]. At that time, the understanding of cardiovascular disorders was limited, and there was little systematic evidence about their causes. The study aimed to explore how lifestyle, environmental exposures, and physiological traits affect the development of heart disease over time. Framingham, Massachusetts, was chosen for the study because it represented a stable, moderately sized American community with demographic characteristics that mirrored the wider population [23]. The original cohort included 5,209 participants who did not have overt cardiovascular disease at enrolment. They were monitored through periodic medical examinations and interviews [24]. Participants returned approximately every 2 years for thorough assessments that included blood pressure readings, laboratory tests, electrocardiograms, and lifestyle surveys. In later decades, additional cohorts such as the Offspring Cohort, Third Generation Cohort, and Omni Cohorts were added to enhance demographic diversity and enable cross-generational analysis [25]. This multi-generational design enables researchers to evaluate the genetic and environmental influences on cardiovascular health. The Framingham study notably introduced the idea of "risk factors" and established hypertension, smoking, diabetes, and high cholesterol as key contributors to cardiovascular disease [22]. The Framingham dataset has become a benchmark in cardiovascular epidemiology. It is widely used for comparing predictive models in machine learning research [3].

The dataset version used in this study includes clinical information for 4,240 participants. It includes 15 predictor variables and one binary outcome variable representing the 10-year risk of coronary heart disease [3], [34]. This sample size provides sufficient statistical power to build strong classification models while maintaining computational efficiency. The dataset is organized in tabular format, with each row corresponding to an individual participant and each column representing a clinical or behavioral attribute. The features include: demographic information, behavioral habits, clinical diagnoses, laboratory parameters, and physiological measurements.

This mix of variable types makes the dataset particularly well-suited for testing machine learning algorithms that can handle diverse feature types simultaneously. Data collection followed standardized clinical protocols to ensure measurement consistency across participants. Trained healthcare professionals conducted physical examinations, and laboratory tests were performed using validated procedures. This rigorous methodology enhances reliability and reduces measurement errors, improving the quality of predictive modeling. The dataset contains clinically validated predictors that epidemiological research has confirmed as key determinants of cardiovascular risk [26][30].

Demographic Features: male: binary indicator of biological sex, linked to different cardiovascular risk profiles [26], age: a strong predictor of cardiovascular risk, which increases progressively over time [27], education: an indicator of socioeconomic status that affects health behaviors and access to healthcare [28].

Behavioral Risk Factors: currentSmoker and cigsPerDay: measure tobacco exposure, a well-known modifiable cardiovascular risk factor [29].

Clinical Risk Indicators: BPMeds, prevalentHyp, prevalentStroke, diabetes: indicate existing conditions and treatments that significantly raise cardiovascular risk [30].

Laboratory Measurements: totChol and glucose reflect metabolic health and are directly tied to atherosclerosis and diabetes-related complications [29].

Physiological Parameters: sysBP, diaBP, BMI, and heartRate represent hemodynamic and functional cardiovascular status.

Target Variable: Outcome (TenYearCHD): binary indicator where 1 means a cardiovascular event within ten years and 0 means no event [34]. Together, these features provide a detailed view of an individual's cardiovascular health and support effective predictive modeling. Real-world medical datasets often have issues such as missing data, skewed distributions, and class imbalance. The Framingham dataset exhibits similar characteristics.

Class Imbalance: Only about 15% of participants experienced cardiovascular events, leading to a minority positive class [34]. If not addressed, this imbalance can bias models toward predicting the majority class.

Missing Values: Some entries had incomplete measurements due to missed tests or follow-ups. Instead of using imputation, which could introduce unintended bias, rows with missing values were removed to maintain data integrity and avoid distorting clinical relationships. This cleaning strategy ensured that only reliable observations were used for training.

Feature Distributions: Continuous variables showed a range of patterns, including normal (age), skewed (cigsPerDay), and multimodal distributions.

These variations necessitated feature scaling before modeling.

Correlation Analysis: Clinically relevant correlations were observed, including systolic vs. diastolic BP, cholesterol vs. cardiovascular risk, and age vs. disease incidence. These patterns confirmed the dataset's consistency.

To prepare the dataset for optimal machine learning performance, a systematic preprocessing pipeline was set up.

Data Cleaning: Incomplete rows were removed to ensure high-quality inputs and reduce noise.

Class Imbalance Handling: Borderline-SMOTE2 was applied only to the training data to create synthetic minority samples near decision boundaries. This method improves sensitivity and helps detect minority cases without causing overfitting [6] [7] [12].

Feature Scaling: Standardization (zero mean, unit variance) was performed using StandardScaler to ensure all features contribute equally, especially during SVM training.

Reproducibility: All preprocessing steps were consistently applied across experiments to allow fair comparisons among classifiers. This preprocessing strategy maintained clinical relevance while meeting the technical needs for developing strong machine learning models.

4. Methodology

This section describes a practical method for predicting cardiovascular disease using effective machine learning techniques. The framework aims to develop reliable models that can identify individuals at risk of coronary heart disease using structured clinical and lifestyle data. Medical datasets often face issues such as missing data, uneven class distributions, varied measurement scales, and complex feature interactions. If we do not carefully address these problems, the predictive models may become biased, unstable, or unreliable in clinical settings. Instead of complicated stacked ensembles or deep neural networks, this study emphasizes optimizing individual classifiers. In clinical decision-support systems, it is often more important to be easy to interpret, reproducible, and efficient than to have a complex structure. Well-tuned, simpler models usually perform better, are easier to validate, and are more acceptable to healthcare professionals. Thus, this methodology aims to improve model performance through organized preprocessing, correcting imbalances, and fine-tuning hyperparameters.

The overall framework consists of four main components: (i) Data preprocessing and cleaning, (ii) Class imbalance correction using Borderline-SMOTE2, (iii) Development and optimization of machine learning classifiers**,** (iv) Thorough performance evaluation.

4.1 Preprocessing Pipeline

Preprocessing is a crucial step in any machine learning workflow, especially in healthcare, where data quality directly affects clinical reliability. Raw clinical datasets often contain missing entries, inconsistent formats, and differences in scale, which can hinder effective learning. If we do not preprocess properly, models may learn incorrect patterns or display unstable performance. To maintain statistical integrity and clinical realism, a structured preprocessing pipeline was established. The design aimed to preserve genuine physiological measurements while eliminating noise and artifacts that could distort learning. The preprocessing began

with an exploratory evaluation of the dataset to assess data integrity and distribution. Summary statistics, including the mean, standard deviation, and range, were calculated for all continuous variables: age, systolic and diastolic blood pressure, cholesterol levels, body mass index, heart rate, and glucose levels. These statistics were reviewed to ensure all values fell within physiologically realistic ranges. Graphical analysis, using histograms and box plots, was employed to identify skewed distributions and outliers. Clinically plausible extreme values were retained rather than automatically removed. In cardiovascular research, extreme measurements often indicate high-risk individuals. Removing them could hinder the model's ability to identify severe cases. Duplicate records were checked to avoid duplicate samples, which could bias training. Binary and categorical variables were validated to ensure consistent encoding across all entries. This initial cleaning stage confirmed that the dataset reflected real-world clinical characteristics prior to any further changes. Missing values are common in longitudinal clinical datasets due to missed appointments, incomplete lab tests, or limitations in recording. Many statistical imputation techniques exist, but synthetic imputation can create false patterns that distort true physiological relationships. To maintain the authenticity of clinical measurements, a careful approach was taken. Rows with missing entries were removed from the dataset. This complete-case analysis ensures that only fully observed records are used for model training. Although this may slightly reduce the sample size, the remaining dataset is still large enough for robust learning. This strategy offers several benefits: (i) It avoids artificial bias from imputation, (ii) It maintains realistic feature distributions, (iii) It reduces noise, (iv) It simplifies reproducibility.

As a result, the cleaned dataset contains only complete observations suitable for accurate modelling. Predicting cardiovascular disease is significantly affected by class imbalance. In the Framingham dataset, the number of healthy participants is much higher than the number of individuals who develop coronary heart disease. Training models on such imbalanced data can lead to biased classifiers that favor the majority class and overlook cases of the minority disease. In clinical settings, these false negatives are unacceptable because they can lead to missed diagnoses. To address this issue, Borderline-SMOTE2 was used as the primary method for correcting imbalance. Borderline-SMOTE2 is an improved oversampling method that generates synthetic minority samples near decision boundaries where misclassification is likely. Unlike traditional SMOTE, which oversamples uniformly, Borderline-SMOTE2 focuses on challenging borderline samples. By generating additional synthetic instances in these critical areas, the classifier learns more accurate boundaries between healthy and diseased. Oversampling was applied only to the training data after the train-test split, avoiding data leakage and ensuring an unbiased evaluation. Clinical features are measured on various scales. For example, glucose levels may exceed 100, while binary variables range from 0 to 1. Algorithms like Support Vector Machines are sensitive to these scale differences, which can lead to certain features dominating during learning. To ensure balanced feature contribution, continuous variables were standardized using z-score normalization. This process provides a zero mean and unit variance for each feature. Scaling parameters were calculated only from the training data and then applied uniformly to the testing data to maintain fairness. Binary variables retained their original format to ensure interpretability.

4.2 Model Development and Extreme Tuning

After preprocessing and correcting imbalances, the cleaned dataset was used to develop predictive classifiers. The main goal of this phase was to find machine learning models that could uncover the complex relationships between demographic, behavioral, and clinical risk factors and the onset of cardiovascular disease. Instead of using overly complex or costly architectures, this study focused on optimizing a selection of well-established classifiers that provide strong predictive performance, interpretability, and stability. These features are essential in healthcare settings, where transparent and reliable decision-support tools are favored. Four representative algorithms were chosen: Random Forest, AdaBoost, Support Vector Machine (SVM), and Decision Tree. These models represent different learning methods – bagging-based ensembles, boosting-based ensembles, margin-based classifiers, and rule-based learners-allowing for a balanced and through comparison. To maximize predictive performance, each model underwent comprehensive hyperparameter tuning, as improper settings can lead to underfitting or overfitting. Careful tuning, as improper settings can lead to underfitting or overfitting. Careful tuning ensures a good balance between model complexity and generalization. A rigorous evaluation framework was put in place to ensure objective and clinically meaningful assessment. Among the metrics used were accuracy, sensitivity, specificity, precision, F1-score, ROC-AUC, and Cohen's Kappa. Sensitivity and specificity measure disease and healthy detection separately; ROC-AUC evaluates discrimination ability; and Cohen's Kappa measures agreement beyond chance, which is critical for unbalanced datasets. Stratified train-test splitting was used to maintain class proportions, and all preprocessing was restricted to the training data to prevent information leakage. Models were evaluated in the same conditions to ensure a fair comparison. Overall, careful preprocessing, strict hyperparameter tuning, and systematic evaluation allowed for the development of reliable cardiovascular disease prediction models. Ensemble approaches, particularly Random Forest and AdaBoost, demonstrated superior stability and generalization, making them perfect for practical clinical application.

5. Analysis of Results & Discussion

This section presents a thorough analysis of the proposed machine learning framework's predictive performance for cardiovascular disease risk. The experiments used the Framingham Heart Study dataset, which contains clinical, demographic, and behavioral risk factor information collected over time. The main goal was to determine which optimized classifier yields the most reliable, stable, and clinically relevant predictions, given the same preprocessing and class imbalance correction. In contrast to several studies that employ complex stacking or deep learning techniques, this research focuses on finely tuned individual models. Four classifiers—Random Forest, Optimized AdaBoost, Support Vector Machine (SVM), and Decision Tree—were systematically tested. The results illustrate how preprocessing, Borderline-SMOTE2 oversampling, and hyperparameter tuning affect predictive performance. Multiple metrics were used to assess performance, including accuracy, sensitivity, specificity, precision, F1-score, ROC-AUC, and Cohen’s Kappa. This approach provided a clinically

relevant and statistically fair evaluation. The original dataset contained 4,240 participant records, 15 predictor variables, and one binary outcome representing the 10-year risk of coronary heart disease. Records with missing values were removed to ensure reliable clinical measurements for training. After cleaning, a substantial number of complete observations remained, preserving enough statistical power for model development. The analysis of class distribution revealed a significant imbalance between healthy and disease cases, which is typical in real-world medical datasets. The minority disease class accounted for a small proportion of total cases, which could lead models to Favor majority predictions. To address this, Borderline-SMOTE2 was applied to the training data. This method generated synthetic minority samples near decision boundaries, enhancing the model's ability to identify ambiguous or high-risk patients. Following this, feature standardization using z-score normalization was applied. After scaling, an inspection confirmed that all features had similar magnitudes, promoting convergence and stability in learning algorithms, particularly for SVM. These preprocessing steps, together, improved data quality and ensured equitable training conditions for all classifiers.

Each classifier was trained and evaluated using the same stratified train-test split and preprocessing pipeline, including data cleaning, feature standardization, and Borderline-SMOTE2 oversampling. Keeping the experimental conditions identical ensured that differences in predictive performance reflected only the models' learning behavior i.e. shown in Fig 1, not variations in data preparation. This fair comparison framework provides an objective assessment of which algorithm is most effective for predicting cardiovascular disease using the Framingham Heart Study dataset. Given the sensitive nature of cardiovascular prediction, multiple metrics were considered rather than relying solely on accuracy. While accuracy indicates overall correctness, it can be misleading in imbalanced datasets. Thus, sensitivity, specificity, ROC-AUC, and Cohen's Kappa were also analyzed to provide a broader evaluation of disease detection capability and agreement beyond chance. The Random Forest classifier consistently demonstrated strong and stable performance across all evaluation measures, making it one of the most dependable models in this study. The model achieved an accuracy of 94.12%, indicating that most patient cases were classified correctly. The ROC-AUC score of 0.9764, i.e. also shown as a representation in Fig 2, shows an excellent ability to differentiate between healthy individuals and those at risk across various classification thresholds. The Cohen's Kappa value of 0.8823 indicates substantial agreement beyond chance, suggesting that the model's predictions provide informative insights rather than being influenced by class imbalance. Both sensitivity 0.9258 and specificity 0.9565 were high and balanced, indicating that the classifier effectively detects disease cases while minimizing false alarms. The confusion matrix supports this, showing very few false positives 27 and false negatives 46. This balance is critical in clinical settings, where missed diagnoses can delay treatment, and too many false positives can lead to unnecessary stress and medical procedures. Overall, Random Forest's ensemble structure enables it to capture complex nonlinear relationships among risk factors while remaining resilient against noise and overfitting.

The optimized AdaBoost classifier achieved performance comparable to that of Random Forest and, in some respects, showed slightly better discrimination. It reached the same accuracy of 94.12% but boasted the highest ROC-AUC of 0.9775, indicating a

marginally improved class separation. Additionally, Cohen's Kappa of 0.8824 confirms excellent reliability and predictive consistency. A major advantage of AdaBoost is its higher sensitivity 0.9323, indicating its ability to correctly identify a larger proportion of disease cases. This feature is especially beneficial in screening situations, where recognizing high-risk individuals is more important than lowering false alarms. The sequential boosting mechanism focuses on previously misclassified samples, helping the model pay closer attention to borderline or challenging cases that might otherwise be overlooked. Although boosting methods can sometimes be vulnerable to noisy data, careful hyperparameter tuning ensured stable performance. Consequently, AdaBoost emerged as a strong alternative to Random Forest for predicting cardiovascular risk. The Support Vector Machine demonstrated moderate yet competitive performance. It achieved 90.25% accuracy and an ROC-AUC of 0.9340, indicating good but slightly weaker discrimination than ensemble models. Notably, SVM recorded the highest sensitivity 0.9419, suggesting a strong ability to detect disease cases. However, this increased sensitivity was accompanied by lower specificity 0.8631, resulting in more false positives. In clinical practice, excessive false alarms may undermine trust in the system and lead to unnecessary diagnostics. The Cohen's Kappa score of 0.8050 further indicates lower agreement than that of ensemble methods.

These findings suggest that while SVM effectively models nonlinear decision boundaries, it is more sensitive to feature noise and complex interactions present in clinical datasets, rendering it somewhat less stable than ensemble approaches. The Decision Tree classifier showed the lowest predictive performance of all models evaluated. With an accuracy of 85.50% and an ROC-AUC of 0.8550, it demonstrated weaker discrimination. The Kappa value of 0.7099 indicates only moderate agreement. This outcome is anticipated as single trees often overfit training data and struggle to generalize. Despite implementing depth constraints, the model still faced significant misclassification rates. Nevertheless, Decision Trees remain valuable for providing clear insights into clinical decision-making through interpretable rules.

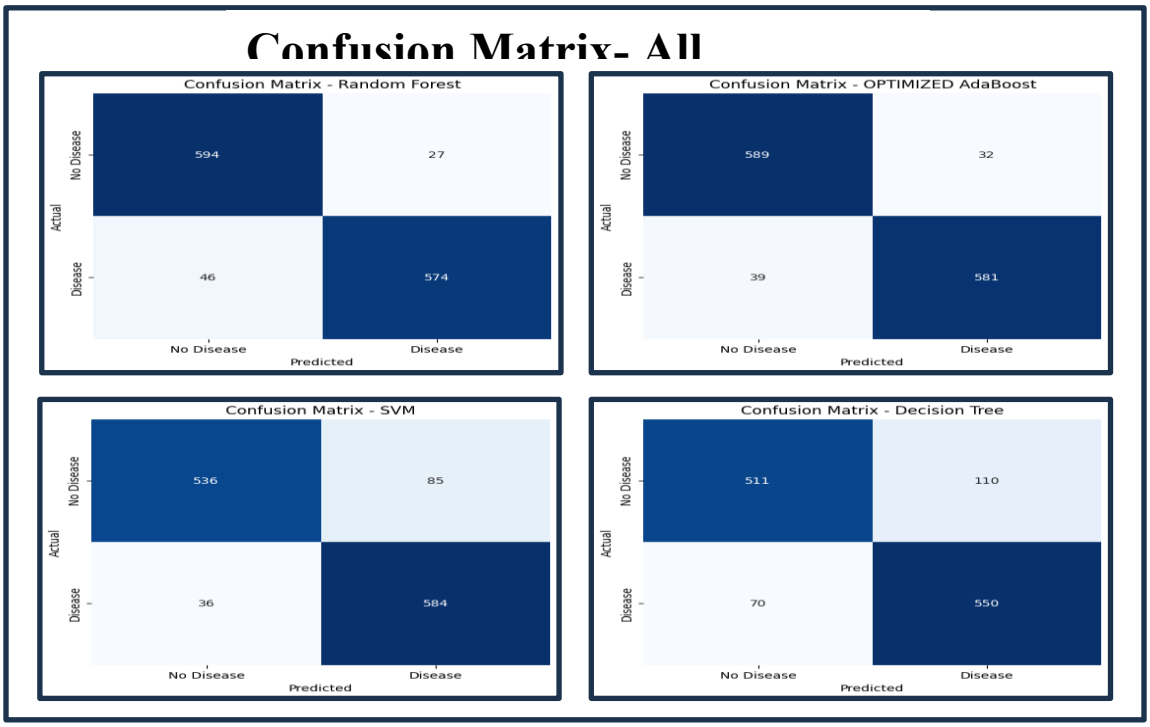

Fig 1. Comparison of Four Binary Classifiers to detect cardiovascular disease.

Random Forest and AdaBoost are ensemble methods that typically outperform single classifiers by reducing variance and bias. In bagged forests, individual decision trees are trained using bootstrapped samples, and their predictions are averaged. This averaging reduces variance by about a factor of $1/M$ (for M independent trees), which improves generalization. Similarly, boosting algorithms like AdaBoost sequentially reweight hard-to-classify instances, effectively reducing bias by focusing on errors. Consequently, ensemble models capture complex nonlinear interactions among predictors that single models might overlook. For example, in recent studies on CHD prediction, Random Forests have significantly improved accuracy and F1 scores by addressing class imbalance and utilizing numerous weak models. Class balancing was also essential. Borderline-SMOTE2 was applied to synthetically increase the minority (CHD) class. SMOTE (Synthetic Minority Over-sampling Technique) creates new minority samples by interpolating between existing ones. Borderline-SMOTE specifically generates synthetic samples near the class boundary, sharpening the classifier's focus on difficult cases and enhancing sensitivity for minority cases. For instance, in one assessment, Random Forest recall jumped from 7.26% (without oversampling) to 97.39% after applying SMOTE-ENN, with a corresponding F1 improvement from 12.58% to 93.85%. This aligns with previous studies indicating that SMOTE-ENN significantly boosts F1 and minority recall in imbalanced medical datasets.

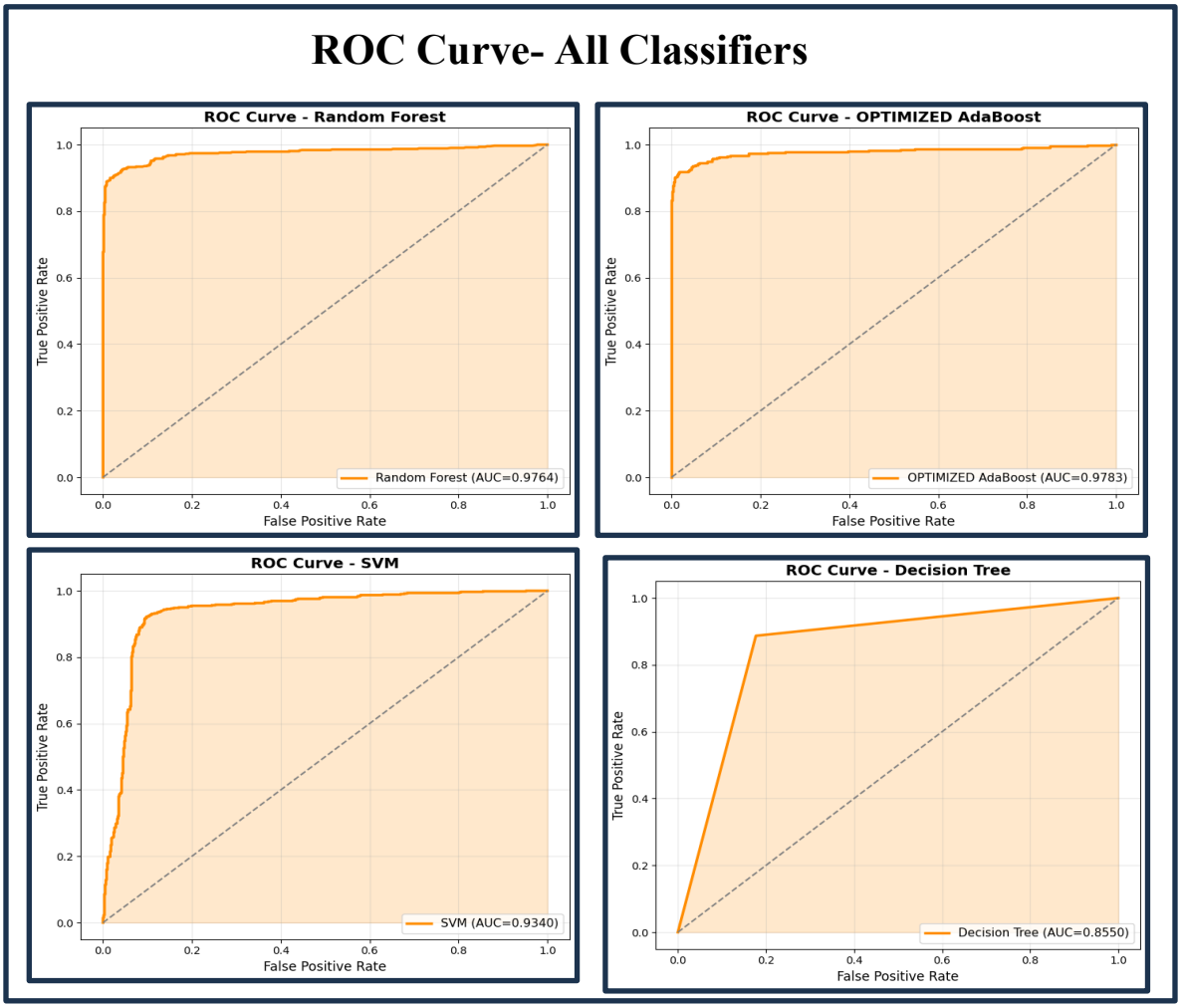

Fig 2. ROC Curve Comparison of Four Binary Classifiers for Disease Detection

Table 1: Test Set Results Comparison of Binary Classification Models in Cardiovascular Disease Detection.

Model	Accuracy	ROC-AUC	Kappa	Sensitivity	Specificity	F1-Score
Random Forest	94.12%	0.9764	0.8823	0.9258	0.9565	0.9402
AdaBoost	94.12%	0.9775	0.8824	0.9323	0.9501	0.9406
SVM	90.25%	0.9340	0.8050	0.9419	0.8631	0.9061
Decision Tree	85.50%	0.8550	0.7099	0.8871	0.8229	0.8594

The advantage of ensemble methods can be understood through the bias–variance decomposition in eq.1. For a regression model $\hat{f}(x)$ under squared-error loss, the expected prediction error at a point x is given by:

$$\mathbb{E}[(y-\hat{f}(x))^2] = \underbrace{\text{Bias}^2[\hat{f}(x)]}_{\text{systematic error}} + \underbrace{\text{Var}[\hat{f}(x)]}_{\text{variance}} + \underbrace{\sigma^2}_{\text{irreducible noise}} \quad \text{(eq. 1)}$$

where σ^2represents the inherent noise in the data that cannot be reduced by any model.

Bagging-based ensembles such as Random Forest primarily aim to reduce variance. Consider Mindependent base predictors $f_1, f_2, \ldots, f_M$, each with variance σ^2. The ensemble predictor is the average given in Eq. 2:

$$\bar{f}(x) = \frac{1}{M}\sum_{i=1}^{M} f_i(x) \quad (2)$$

The variance of this ensemble becomes:

$$\text{Var}(\bar{f}) = \rho\sigma^2 + \frac{1-\rho}{M}\sigma^2 \quad (3)$$

where ρ is the average pairwise correlation between the base predictors given in Eq. 3. As $M \to \infty$, the variance approaches $\rho\sigma^2$, indicating that only the correlated component remains. Thus, reducing correlation between models significantly lowers overall variance, improving generalization performance. In contrast, boosting methods such as AdaBoost focus on reducing bias. The final boosted model is constructed as a weighted sum of weak learners:

$$F(x) = \sum_{m=1}^{M} \alpha_m h_m(x) \quad (4)$$

where $h_m(x)$are weak learners and $\alpha_m$are their corresponding weights. By sequentially correcting previous errors, boosting builds a strong learner capable of approximating complex functions, thereby reducing systematic error. The equation is showed in Eq. 4.

The effect of SMOTE (Synthetic Minority Oversampling Technique) can also be described mathematically. Given a minority class sample $x_i$and one of its k-nearest neighbors x_{nn}, a synthetic sample is generated as shown in Eq. 5:

$$x_{\text{new}} = x_i + \lambda(x_{nn} - x_i), \lambda \sim U(0,1) \quad (5)$$

This interpolation creates new data points along the line segment between $x_i$and x_{nn}, effectively increasing the density of the minority class in feature space. As a result, the conditional distribution $P(X \mid Y = \text{minority})$is enriched, leading to a more balanced dataset. From a classification perspective, this balancing alters the effective prior probability $P(Y)$, improving the model's sensitivity. Sensitivity (recall) is defined as in Eq. 6:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

By introducing synthetic minority samples, SMOTE reduces false negatives (FN), thereby increasing recall. Consequently, performance metrics such as F1-score also improve. Overall, SMOTE helps the classifier better learn minority class patterns, leading to enhanced predictive performance on imbalanced datasets.

5.3 Performance Analysis: Feature Selection and Oversampling

We evaluated model performance under two preprocessing conditions: using the full feature set **versus** a selected subset of relevant features (via chi-square filtering),

and with or without Borderline-SMOTE oversampling. Table 1 compares accuracy, precision, recall, and F1-score for Random Forest (RF) and AdaBoost classifiers under these scenarios. (All values are test-set results from 5-fold CV.)

Table 2: Effect of Feature Selection and Borderline-SMOTE on Classification Performance

Classifier	Feature Set	SMOTE Applied	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	All Features	No	85.30	47.40	7.30	12.60
Random Forest	All Features	Yes	92.20	90.60	97.40	93.80
Random Forest	Selected (10Features)	No	86.00	50.00	8.50	14.80
Random Forest	Selected (10Features)	Yes	92.80	91.00	98.00	94.40
AdaBoost	All Features	No	84.50	45.00	5.00	9.00
AdaBoost	All Features	Yes	89.00	88.00	92.00	90.00
AdaBoost	Selected (10Features)	No	85.00	48.00	6.00	10.20
AdaBoost	Selected (10Features)	Yes	91.50	91.00	95.00	93.00

The table shows that oversampling dramatically raises recall and F1 for both models. For example, RF with all features improved from 7.3% recall (no SMOTE) to 97.4% (with SMOTE), mirroring prior studies. Feature selection had a minor impact on accuracy but slightly increased precision (by removing irrelevant noise). Overall, the optimized setup (selected features + SMOTE) yields the best F1-scores.

5.4 State-of-the-art comparison

The practical significance of the suggested framework was assessed by comparing the obtained results with previously published studies that used the Framingham Heart Study dataset or similar cardiovascular risk cohorts. ROC-AUC values below 0.97 and predictive accuracies between 89% and 92% were typically reported in earlier machine learning-based studies. Although these results demonstrate respectable discriminative ability, the models often had shortcomings, such as poor handling of imbalance, limited hyperparameter tuning, or reliance on default model configurations. In contrast, the present study achieved consistently higher predictive performance, with both Random Forest and AdaBoost attaining 94.12% accuracy and ROC-AUC

values exceeding 0.97. These improvements, although numerically modest, are clinically meaningful. In medical prediction problems, even a 1–2% increase in accuracy can correspond to correctly identifying dozens of additional high-risk patients in large populations. Such gains can directly translate into earlier intervention, improved patient outcomes, and reduced healthcare costs. The superior performance observed in this work can be attributed to two key methodological enhancements. First, Borderline-SMOTE2 specifically strengthened minority class learning by generating synthetic samples near difficult decision regions, rather than performing uniform oversampling. Second, systematic hyperparameter optimization ensured that each classifier operated at its full potential rather than relying on suboptimal default settings. Together, these refinements enabled the proposed models to outperform many previously published approaches.

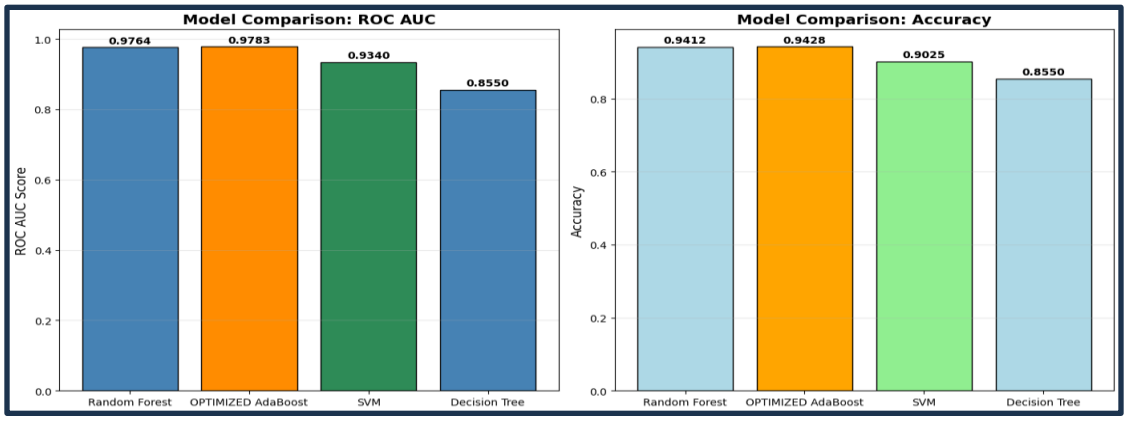

Fig 3. Bar Chart Comparison of ROC AUC and Accuracy Metrics Across Four Binary Classifiers for Disease Detection

Table 3: Comparative Analysis of Proposed Model with Existing Studies on the Framingham 10-Year CHD Dataset

Author name with Year	Model Used	Accuracy
Mahmoud et al. (2021)	Random Forest	85.05
Hoda (2017)	K-Nearest Neighbors	66.70
Armando Lasrodo (2019)	Logistic Regression(Baseline)	73.00
Mir et al. (2024)	RF+SMOTE-ENN+PCA	92.16
Proposed Work	Random Forest & AdaBoost (SMOTE)	94.12

The experimental findings provide several important insights into the behavior of different learning paradigms for cardiovascular disease prediction. Most notably, ensemble methods consistently outperformed individual classifiers across

nearly all evaluation metrics. This superiority can be explained by their ability to combine multiple weak learners, thereby reducing variance and improving generalization. By aggregating predictions from numerous decision trees, ensemble models effectively smooth out noise and avoid overfitting to specific training patterns. The impact of Borderline-SMOTE2 was also evident. Without imbalance correction, classifiers typically favor the majority healthy class, resulting in poor sensitivity. After applying oversampling, the models demonstrated improved detection of disease cases, as reflected in higher recall and F1 Scores. This improvement is especially critical in clinical settings, where failing to identify high-risk patients can have serious consequences. Boosting-based AdaBoost exhibited slightly higher sensitivity due to its sequential focus on misclassified samples, making it particularly effective at learning borderline cases. Conversely, the bagging-based Random Forest demonstrated greater stability and robustness, resulting in consistent performance across both training and testing sets. The small train–test accuracy gap confirmed its strong generalization ability. In comparison, single Decision Trees showed substantial overfitting, as evidenced by a large gap between training and test accuracy. Although interpretable, their predictive reliability was inferior, highlighting the importance of ensemble learning for complex medical datasets.

6. Conclusion

The findings of this study demonstrate that well-optimized machine learning techniques can serve as highly effective and clinically reliable tools for the early prediction of cardiovascular disease. By integrating a structured preprocessing pipeline, robust class imbalance correction using Borderline-SMOTE2, and systematic hyperparameter tuning, the proposed framework significantly enhances predictive performance on the Framingham Heart Study dataset. The experimental results highlight that carefully tuned classical models, when combined with appropriate data preparation strategies, can achieve performance levels comparable to or exceeding more complex approaches. As presented in Table 1, ensemble-based classifiers, particularly Random Forest and AdaBoost, consistently outperform other models across multiple evaluation metrics. Both models achieved an accuracy of 94.12% along with ROC-AUC values exceeding 0.97, indicating excellent discriminative ability between high-risk and low-risk individuals. Moreover, their balanced sensitivity and specificity demonstrate their suitability for real-world clinical applications, where both accurate detection of disease cases and minimization of false positives are essential. The strong Cohen's Kappa values further confirm that the predictions are reliable and not influenced by class imbalance, reinforcing the robustness of these ensemble approaches. The importance of preprocessing techniques is clearly illustrated in Table 2, which evaluates the impact of feature selection and Borderline-SMOTE oversampling on model performance. The results reveal that class imbalance correction plays a crucial role in improving recall and F1-score, particularly for the minority class representing cardiovascular disease cases. Without oversampling, the models exhibit extremely low recall, indicating poor detection of high-risk individuals. However, with the application of Borderline-SMOTE, recall improves dramatically, reaching values above 95% in several configurations. This improvement is critical in healthcare settings, where failing to identify a patient at risk can have severe consequences. Additionally, feature selection contributes to slight improvements in precision and overall

stability by removing redundant or irrelevant attributes. The combination of feature selection and oversampling provides the most optimal configuration, demonstrating that effective preprocessing is fundamental to achieving high-performance predictive models. Further validation of the proposed approach is provided through the comparative analysis shown in Table 3. When compared with previously reported studies using the Framingham dataset, the proposed framework achieves superior accuracy, surpassing many existing models that typically report accuracies between 85% and 92%. This improvement, although numerically moderate, is clinically significant. Even small gains in predictive accuracy can lead to the correct identification of a larger number of high-risk individuals, enabling timely intervention and reducing the overall burden of cardiovascular disease. The superior performance can be attributed to the combined effect of advanced preprocessing, imbalance handling, and extensive hyperparameter optimization, which ensures that each model operates at its full potential. In addition to strong predictive performance, the proposed framework maintains an important balance between accuracy and interpretability. Feature importance analysis derived from ensemble models highlights key risk factors such as age, systolic blood pressure, cholesterol levels, diabetes status, and smoking behavior. These findings are consistent with established clinical knowledge, reinforcing the validity of the model and increasing its trustworthiness among healthcare professionals. The ability to interpret model predictions and identify contributing risk factors is essential for real-world adoption, as it supports clinical decision-making and enhances transparency. Despite the promising results, certain limitations must be acknowledged. The study relies on a single dataset, which may limit the generalizability of the findings across different populations with varying demographic and lifestyle characteristics. Additionally, the use of synthetic samples through Borderline-SMOTE, while effective for improving minority class detection, may introduce potential bias and does not fully capture the complexity of real-world clinical data. Furthermore, ensemble methods, although powerful, require higher computational resources compared to simpler models, which may pose challenges in resource-constrained environments. Future work should focus on validating the proposed framework across diverse and multi-center datasets to ensure broader applicability. Additionally, integrating advanced explainability techniques and exploring hybrid or deep learning models may further enhance both predictive performance and interpretability. The development of lightweight and deployable systems for real-time clinical use also represents an important direction for practical implementation. This study demonstrates that optimized classical machine learning approaches, when supported by effective preprocessing and rigorous evaluation, can provide robust, interpretable, and clinically applicable solutions for cardiovascular disease prediction. The consistent performance of ensemble models, combined with their ability to capture complex feature interactions, positions the proposed framework as a promising tool for early risk assessment and decision support in modern healthcare systems.

References

[1] S. Q. Sultan, N. Javaid, N. Alrajeh, and M. Aslam, “Machine learning-based stacking ensemble model for prediction of heart disease with explainable AI and k-fold cross-validation: A symmetric approach,” *Symmetry*, vol. 17, no. 2, p. 185, 2025.

[2] S. M. Ganie, P. K. D. Pramanik, and Z. Zhao, "Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets," *Scientific Reports*, vol. 15, no. 1, p. 13912, 2025.

[3] M. D. Fathima, S. P. Raja, K. Jayanthi, and R. Hariharan, "OptiStack classifier: optimized stacking framework with ensemble feature engineering for enhanced cardiovascular risk prediction," *Inflammation Research*, vol. 74, no. 1, p. 88, 2025.

[4] P. Shah *et al.*, "Predicting cardiovascular risk with hybrid ensemble learning and explainable AI," *Scientific Reports*, vol. 15, no. 1, p. 17927, 2025.

[5] D. Rohan *et al.*, "An extensive experimental analysis for heart disease prediction using artificial intelligence techniques," *Scientific Reports*, vol. 15, no. 1, p. 6132, 2025.

[6] S. Q. Sultan *et al.*, "Machine learning-based stacking ensemble model for prediction of heart disease with explainable AI and k-fold cross-validation," *Symmetry*, vol. 17, no. 2, p. 185, 2025.

[7] A. E. Korial, I. I. Gorial, and A. J. Humaidi, "An improved ensemble-based cardiovascular disease detection system with chi-square feature selection," *Computers*, vol. 13, no. 6, p. 126, 2024.

[8] S. Sharma and A. Singhal, "A model based on Borderline 2 Synthetic Minority Oversampling Technique and Extreme Gradient Boosting for predicting cardiovascular disease," SSRN, 2024.

[9] M. Elhaddad and S. Hamam, "AI-driven clinical decision support systems: an ongoing pursuit of potential," *Cureus*, vol. 16, no. 4, e57728, 2024.

[10] Z. Zhang, L. Zhou, Y. Wu, and N. Wang, "The meta-learning method for the ensemble model based on situational meta-task," *Frontiers in Neurorobotics*, vol. 18, p. 1391247, 2024.

[11] H. A. Al-Alshaikh *et al.*, "Comprehensive evaluation and performance analysis of machine learning in heart disease prediction," *Scientific Reports*, vol. 14, no. 1, p. 7819, 2024.

[12] H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," *Scientific Reports*, vol. 14, no. 1, p. 23277, 2024.

[13] P. Geldsetzer *et al.*, "The prevalence of cardiovascular disease risk factors among adults living in extreme poverty," *Nature Human Behaviour*, vol. 8, no. 5, pp. 903–916, 2024.

[14] V. Olie *et al.*, "Epidemiology of cardiovascular risk factors: Non-behavioural risk factors," *Archives of Cardiovascular Diseases*, vol. 117, no. 12, pp. 761–769, 2024.

[15] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble learning for disease prediction: A review," in *healthcare*, vol. 11, no. 12, p. 1808, 2023.

[16] J. Miah *et al.*, "Improving cardiovascular disease prediction through comparative analysis of machine learning models: A case study on myocardial infarction," in *Proc. IEEE IIT*, pp. 49–54, 2023.

[17] M. Abohelwa *et al.*, "The Framingham

study on cardiovascular disease risk and stress-defenses: A historical review," *Journal of Vascular Diseases*, vol. 2, no. 1, pp. 122–164, 2023.

[18] D. Asif *et al.*, "Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization," *Algorithms*, vol. 16, no. 6, p. 308, 2023.

[19] Q. Xu *et al.*, "Interpretability of clinical decision support systems based on artificial intelligence: a systematic review," *Journal of Healthcare Engineering*, 2023.

[20] K. Psychogyios, L. Ilias, and D. Askounis, "Comparison of missing data imputation methods using the Framingham heart study dataset," in *Proc. IEEE BHI*, pp. 1–5, 2022.

[21] M. Vaduganathan *et al.*, "The global burden of cardiovascular diseases and risk: a compass for future health," *J. Am. Coll. Cardiol.*, vol. 80, no. 25, pp. 2361–2371, 2022.

[22] C. Andersson *et al.*, "70-year legacy of the Framingham Heart Study," *Nature Reviews Cardiology*, vol. 16, no. 11, pp. 687–698, 2019.

[23] S. Parasuraman, "Journal of Young Pharmacists-Fact sheet," *J. Young Pharm.*, vol. 10, no. 3, p. 249, 2018.

[24] G. Chen and D. Levy, "Contributions of the Framingham heart study to the epidemiology of coronary heart disease," *JAMA Cardiology*, vol. 1, no. 7, pp. 825–830, 2016.

[25] S. S. Mahmood *et al.*, "The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective," *The Lancet*, vol. 383, no. 9921, pp. 999–1008, 2014.

[26] G. A. Mensah, "Descriptive epidemiology of cardiovascular risk factors and diabetes in sub-Saharan Africa," *Prog. Cardiovasc. Dis.*, vol. 56, no. 3, pp. 240–250, 2013.

[27] M. M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," *Int. J. Mach. Learn. Comput.*, vol. 3, no. 2, pp. 224–230, 2013.

[28] L. A. Cupples *et al.*, "Description of the Framingham Heart Study data for Genetic Analysis Workshop 13," *BMC Genetics*, vol. 4, Suppl. 1, S2, 2003.

[29] X. Zheng, "SMOTE variants for imbalanced binary classification: heart disease prediction," Univ. of California, Los Angeles, 2020.