Programming and Algorithm Fundamentals

E-ISSN: 3123-979X

(Research/Review) Article

Comparison of Random Forest and Support Vector Machine Algorithms for Heart Disease Prediction Using SMOTE and Hyperparameter Tuning Techniques

Randi Trinanda

Universitas Diponegoro; e-mail : randitrinanda2004@gmail.com
***** Corresponding Author: Randi Trinanda

Abstract: Heart disease remains one of the leading causes of death worldwide, making early and accurate detection a critical priority in healthcare. However, building an effective predictive model is challenged by the presence of class imbalance in medical datasets, where the number of healthy patients often exceeds disease cases, potentially biasing model performance toward the majority class. This study aims to compare the performance of Random Forest (RF) and Support Vector Machine (SVM) algorithms for heart disease prediction, incorporating Synthetic Minority Over-sampling Technique (SMOTE) to handle class imbalance and Hyperparameter Tuning to optimize each model's configuration. The dataset used consists of 2,026 patient records with 13 clinical features including age, blood pressure, cholesterol, EKG results, and thallium test results, with heart disease (Absence/Presence) as the target variable. Experiments were conducted in two scenarios: without SMOTE and Tuning as baseline, and with SMOTE and Hyperparameter Tuning as the optimized scenario. Results show that both models improved after optimization, with Random Forest achieving the best overall performance with an accuracy of 87.44%, recall of 0.8634, F1-Score of 0.8610, and AUC-ROC of 0.9449, outperforming SVM which reached an accuracy of 86.45% and AUC-ROC of 0.9396. These findings conclude that Random Forest combined with SMOTE and Hyperparameter Tuning is the most suitable model for heart disease prediction, particularly in minimizing false negatives which are clinically critical in medical diagnosis.

Keywords: Heart Disease Prediction; Hyperparameter Tuning; Random Forest; SMOTE; Support Vector Machine.

Received: Februari 21, 2026
Revised: March 26, 2026
Accepted: April 27, 2026
Published: April 30, 2026
Curr. Ver.: April 30, 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

Cardiovascular disease (CVD) remains the most serious global health threat. CVD has been the leading cause of death for more than three decades, accounting for approximately one-third of all deaths worldwide, with 20.5 million deaths recorded in 2021. The latest data from the Global Burden of Disease study shows an increasingly alarming trend: the number of deaths from CVD worldwide rose sharply from 13.1 million in 1990 to 19.2 million in 2023, with ischemic heart disease becoming the leading cause of death. Furthermore, the prevalence of CVD globally also increased from 311 million in 1990 to 626 million in 2023, making CVD the disease with the greatest burden in the world [1].

In the last decade, machine learning (ML) approaches have proven their great potential in supporting clinical diagnosis, particularly for heart disease prediction. Supervised learning-based classification methods such as Logistic Regression, SVM, RF, and XGBoost have proven their effectiveness in heart disease prediction, and recent studies increasingly integrate data balancing techniques such as SMOTE. Among the various existing algorithms, Random Forest (RF) and Support Vector Machine (SVM) have emerged as the two most widely used algorithms [2].

Random Forest is a decision tree-based ensemble algorithm that works by building a large number of decision trees in parallel and combining their results. On the other hand,

SVM is a margin-based algorithm that is robust in handling high-dimensional data and capable of producing optimal decision boundaries [3].

However, one of the main challenges often encountered in predictive modeling of heart disease is class imbalance in medical datasets. Class imbalance is a significant challenge in medical diagnostics, particularly in heart disease prediction using machine learning models. This condition causes the model to favor the majority class, resulting in very low predictive performance for the minority class, which is actually the positive case of the disease. Imbalanced datasets can lead to biased model performance because the model tends to favor the majority class, making balancing techniques such as SMOTE essential [4][5].

The Synthetic Minority Oversampling Technique (SMOTE) has proven to be an effective solution to this problem. Findings from various studies show that models trained on datasets balanced with SMOTE yield statistically significant improvements in precision, F1-score, and recall metrics compared to models trained on the original imbalanced data. The use of SMOTE has consistently been reported to increase the model's sensitivity in detecting minority disease cases [5].

In addition to handling class imbalance, model optimization through hyperparameter tuning also plays a crucial role. Hyperparameter tuning has been shown to significantly improve model performance and predictive accuracy in research on heart disease prediction using machine learning. This process ensures that each algorithm operates at its optimal configuration, making comparisons between algorithms fairer and more meaningful [6].

Several previous studies have compared ML algorithms for heart disease prediction. One study comparing Random Forest and SVM machine learning models with the Knowledge Discovery in Database (KDD) approach without applying SMOTE on a dataset of heart disease medical records found that both algorithms performed less well [7]. Furthermore, other studies have compared Decision Trees, Random Forest, Gradient Boosting, and SVM models with the SMOTE technique and evaluated them using accuracy, sensitivity, specificity, and AUC metrics, with the aim of finding the best model for coronary heart disease prediction. However, a direct comparison between RF and SVM that specifically integrates SMOTE and hyperparameter tuning simultaneously with a dataset containing complete clinical features still requires further study [8].

The dataset used in this study contains 2,026 patient records with 13 relevant clinical features, including age, gender, type of chest pain, blood pressure, cholesterol levels, ECG results, maximum heart rate, exercise-induced angina, ST depression, ST slope, number of blood vessels detected by fluoroscopy, thallium results, and the target variable of the presence or absence of heart disease. Of the total data, there were 1,113 cases of absence and 913 cases of presence, indicating a class imbalance that needs to be addressed with appropriate techniques.

A broader comparative study evaluating multiple machine learning models including Logistic Regression, Random Forest, SVM, and KNN on the UCI Heart Disease Dataset demonstrated that Random Forest consistently achieved superior performance [9]. This finding reinforces the importance of systematic algorithm comparison under standardized experimental settings. In a related study employing GridSearchCV-based hyperparameter tuning across several classifiers including RF, SVM, Logistic Regression, and Naive Bayes, Random Forest again emerged as the top-performing model for heart disease prediction, highlighting the benefit of combining hyperparameter optimization with ensemble learning [10].

Recent studies have also explored advanced ensemble approaches such as XGBoost combined with SMOTE for heart disease prediction, reporting that this integration effectively addresses class imbalance and improves minority-class detection [11]. Research further confirms that machine learning-based clinical decision support systems hold significant promise for reducing misdiagnosis rates, cutting diagnostic costs, and ultimately improving patient outcomes in cardiovascular care [12].

Based on this background, this study aims to conduct a comprehensive comparison between the Random Forest and Support Vector Machine algorithms in predicting heart disease, integrating the SMOTE technique to address class imbalance and hyperparameter tuning to optimize the performance of each model. The results of this study are expected to contribute to the development of accurate clinical decision support systems for the early detection of heart disease.

2. Literature Review

Heart Disease and the Use of Machine Learning in Diagnosis

Heart disease is a subset of cardiovascular disease (CVD) characterized by structural and functional disorders of the heart, including coronary heart disease, heart failure, arrhythmias, and congenital heart disease. In a clinical context, heart disease is diagnosed through a series of tests such as an electrocardiogram (ECG), echocardiography, stress testing, and laboratory blood tests that include cholesterol levels, blood pressure, and fasting blood glucose. While these conventional diagnostic procedures are quite reliable, they require time, expertise, and significant costs. This situation has driven the development of machine learning-based approaches as clinical decision support systems.

A recent systematic review noted that SVM, Random Forest, Gradient Boosting Machine, and Deep Neural Networks have been successfully used to stratify patients based on cardiovascular risk, and these ML models have demonstrated superior performance compared to traditional statistical prediction tools in predicting adverse cardiovascular outcomes. This superiority is possible because machine learning algorithms are able to capture complex non-linear relationships between clinical variables, something that conventional statistical methods such as logistic regression struggle to do [2].

A recent study evaluated and compared the performance of several machine learning models, including Random Forest and SVM, on four different datasets encompassing clinical laboratory biomarkers, lifestyle risk factors, symptom and ECG data, and syndromic surveillance data for infectious diseases, using AUC as the primary evaluation metric. The study's findings highlight that the context and characteristics of the data significantly influence which algorithms perform best, so careful comparative studies are always necessary [13].

Another study using Random Forest and SVM, models built using the Knowledge Discovery in Database (KDD) approach, showed that the Support Vector Machine model provided the best performance compared to the Random Forest Classifier with 65% accuracy, 70% precision, 68% recall, and 64% f-measure. Testing with cross-validation and ROC analysis also showed the highest AUC value in the Support Vector Machine model, namely 0.67–0.68, which is categorized as poor classification [7].

The significant differences in results between these studies indicate a research gap that needs to be filled: a study is needed that systematically compares RF and SVM by applying data balancing and hyperparameter optimization techniques simultaneously and measurably, so that the contribution of each component to improving model performance can be clearly identified.

Random Forest Algorithm

Random Forest (RF) is an ensemble learning algorithm introduced by Leo Breiman in 2001. Random Forest is an ensemble method that combines a number of uncorrelated decision trees with each other through bagging (bootstrap aggregating) techniques, with the aim of increasing accuracy while reducing variance and avoiding overfitting. The working mechanism of RF can be explained through two main phases: first, forest construction by creating a number of N decision trees from different subsets of data (bootstrap sampling results), second, final decision making through majority voting of all trees that have been built [14].

The basic idea of the random subspace method was developed by Ho in 1995 and later significantly expanded by Breiman through the introduction of the bagging technique, which involves training a large number of trees and using majority voting as the classification rule. Each decision tree in a Random Forest is built using a randomly selected subset of features at each branching node, rather than all available features. This results in high diversity among trees and collectively reduces the risk of overfitting [14].

The main advantages of Random Forest include: (1) the ability to handle high-dimensional data with thousands of input variables without overfitting; (2) robustness to outliers and noise in the data; (3) the ability to generate feature importances useful for model interpretation; and (4) relatively good computational speed because trees can be built in parallel.

The main hyperparameters that need to be optimized in Random Forest include: n_estimators (the number of trees in the forest), max_depth (the maximum depth of each tree), min_samples_split (the minimum number of samples to split a node), min_samples_leaf (the minimum number of samples per leaf), and max_features (the maximum number of features considered when finding the best split). The combination of these parameter values directly affects the model's generalization ability and is the focus of the hyperparameter tuning process.

Empirical evidence further supports the superiority of Random Forest in clinical prediction tasks. A study using a large multi-source cardiovascular dataset found that ensemble methods including Random Forest consistently produced strong discriminative performance while their ability to quantify feature importance enabled clinicians to interpret model decisions in a clinically meaningful way [15]. Furthermore, research applying SHAP value analysis to Random Forest models for heart disease prediction identified Thallium test results, ST-segment slope, and chest pain type as the most influential predictors, aligning well with established clinical knowledge [16].

Support Vector Machine Algorithm

Support Vector Machine (SVM) is a supervised learning algorithm developed based on Statistical Learning Theory (SLT) by Vapnik and Chervonenkis. SVM works by finding the optimal hyperplane that separates two classes with the largest margin (maximum margin classifier). The data points closest to the separating hyperplane are called support vectors, and the maximized margin is the distance between the support vectors of each class.

SVM have the advantage of good generalization and show particular advantages in solving pattern recognition problems on small, non-linear, and high-dimensional data, especially in handling classification problems. To handle data that is not linearly separable, SVM use kernel functions that map the data to a higher-dimensional feature space where linear separation is possible. Commonly used kernels include the linear kernel, polynomial kernel, Radial Basis Function (RBF) kernel, and sigmoid kernel [17].

In its application to heart disease prediction, SVM demonstrates competitive performance. Research comparing Linear SVM and Polynomial SVM found that Polynomial SVM, especially when applied to transformed data, yielded higher accuracy in heart disease diagnosis than Linear SVM, with test accuracies of 80% (polynomial) and 83% (linear on transformed data), respectively [18].

Critical hyperparameters in SVM that need to be optimized include: C (a regularization parameter that controls the balance between wide margins and misclassification), kernel (the type of kernel function used), and gamma (which determines the range of influence of a single relevant training sample for RBF and polynomial kernels). Selecting the right C and gamma values is crucial for SVM performance, and this process is generally done through a structured search in the hyperparameter space.

A comprehensive comparison of machine learning algorithms for cardiovascular disease detection using multiple optimized classifiers found that machine learning-based predictive models are highly effective tools for supporting clinical decision-making, with SVM demonstrating competitive recall rates especially when paired with appropriate preprocessing techniques [19]. The effectiveness of SVM in cardiac prediction tasks has also been confirmed in studies using various kernel configurations, where the RBF kernel consistently outperformed linear kernels in capturing complex non-linear feature interactions inherent in clinical datasets [20].

SMOTE Technique for Handling Class Imbalance

Class imbalance is a condition in which the number of samples in one class is significantly greater than in another. In the medical field, this condition is very common because the number of sick patients is usually lower than the number of healthy patients in a population. In the healthcare world, where timely and precise diagnosis is crucial, addressing the challenge of imbalanced classification becomes even more crucial, as misclassifying sick individuals as healthy can have serious and potentially life-threatening health implications [21].

The Synthetic Minority Oversampling Technique (SMOTE) was introduced by Chawla in 2002 as a solution to this problem. SMOTE works by generating synthetic samples of the minority class through an interpolation process between existing minority samples and their nearest neighbors of the same minority class. This mechanism differs from random oversampling, which simply duplicates existing samples; SMOTE creates new, unique samples, thus enriching the diversity of the minority class distribution and reducing the risk of overfitting [21].

SMOTE produces a more diverse set of examples than simple duplication of minority class samples, because the synthetics are constructed as linear combinations of the original samples and their neighbors. The SMOTE technique is crucial for addressing class imbalance, which can significantly impact the performance of machine learning models, and experimental results are presented both before and after applying SMOTE to the training data to compare

the model's performance on untouched test data. Variants of SMOTE have also been developed, including SMOTE-ENN, SMOTE-Tomek, Borderline-SMOTE, and ADASYN, each with different characteristics in handling different imbalance conditions [22].

A systematic review on modified SMOTE strategies confirmed that SMOTE and its variants remain the most widely adopted oversampling approaches in healthcare classification, with SMOTE-ENN and SMOTE-Tomek proving particularly effective in reducing class overlap while preserving data diversity [23]. More recent work applying over-sampling techniques across multi-class imbalanced medical datasets further concluded that SMOTE-based methods are especially valuable when combined with robust classifiers such as Random Forest, yielding more generalized and clinically meaningful predictions [24].

Hyperparameter Tuning

Hyperparameters are configurations determined before model training and are not learned directly from data, unlike model parameters (such as weights in a neural network), which are optimized during training. Selecting the right hyperparameters significantly determines the final performance of a machine learning model. Hyperparameter tuning has been shown to significantly improve model performance and predictive accuracy, and this process is a crucial component in developing machine learning-based heart disease prediction models.

The two most commonly used hyperparameter tuning methods are Grid Search Cross-Validation (GridSearchCV) and Randomized Search Cross-Validation (RandomizedSearchCV). GridSearchCV performs an exhaustive search on all pre-defined hyperparameter combinations, guaranteeing the best parameters but requiring a very large computational time; while RandomizedSearchCV simply samples randomly from the parameter space, is much more efficient although it does not guarantee absolute best parameters [21].

The recommended best practice is to start with RandomizedSearchCV to narrow the search space, then use GridSearchCV for further refinement. Both methods use k-fold cross-validation to evaluate each hyperparameter combination, ensuring that model performance estimates are not biased toward any particular data split. In this study, hyperparameter tuning was applied to both the Random Forest and SVM algorithms to ensure comparisons were conducted on the optimal configuration of each model.

A study applying GridSearchCV and RandomizedSearchCV for cardiovascular disease prediction with multiple classifiers confirmed that hyperparameter tuning consistently yields statistically significant accuracy improvements, and that the benefit is most pronounced for ensemble-based methods such as Random Forest [25].

3. Proposed Method

The research procedure is illustrated in a flowchart, which shows a series of systematic and logical steps, from the beginning to the end of the research.

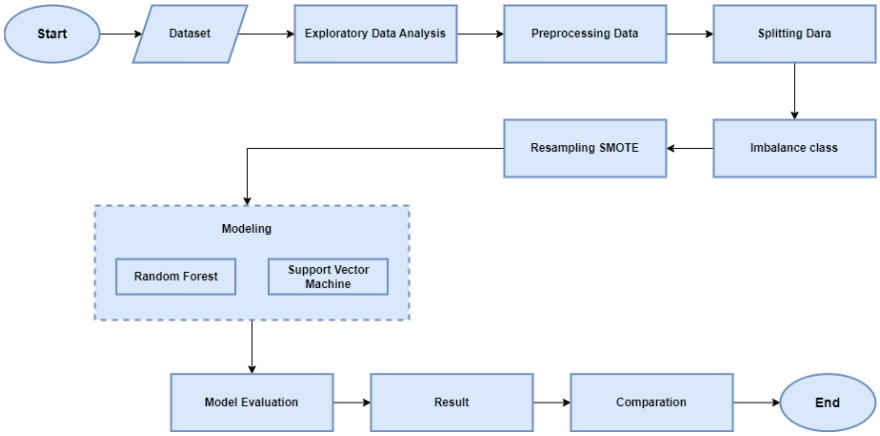


Figure 1. Research flowchart.

Data Collection

The dataset used is sourced from Kaggle https://www.kaggle.com/datasets/algozee/heart-decices from a public repository, containing 2,026 patient medical record data with 13 clinical features such as age, blood pressure, cholesterol, and ECG results, as well as 1 target variable, namely heart disease (Absence/Presence).

Exploratory Data Analysis

This stage aims to understand the characteristics of the dataset before modeling. Descriptive statistical analysis (min, max, mean, standard deviation), visualization of the distribution of each feature, correlation analysis between variables using heatmaps, and observation of the distribution of the target variable, which shows 1,113 cases of absence (54.9%) and 913 cases of presence (45.1%) were performed.

Preprocessing Data

The data was cleaned and prepared for modeling. The id column was removed due to its irrelevance, the target variable was converted from a string to a binary numeric value (Absence=0, Presence=1) using Label Encoding, and a missing values check was performed, resulting in no missing values being found in the dataset. Next, StandardScaler was applied to normalize the numeric features, which is particularly important for the SVM algorithm, which is sensitive to data scale.

Splitting Data

The dataset was divided into two subsets using the holdout method, with 80% training data and 20% testing data. The splitting process was performed using the train_test_split function from the Scikit-Learn library with the stratify=y parameter to ensure the class proportions in the training and testing data remained consistent with their original distribution. A fixed random_state parameter was also applied to ensure the reproducibility of the experimental results. This technique ensures there is no overlap between the data used to train the model and the data used for evaluation.

Imbalance Class

A class imbalance check was performed on the training data. The distribution showed an Absence:Presence ratio of 1.22:1, which, while mild, still has the potential to bias the model toward the majority class. In a medical context, failure to detect Presence cases (False Negatives) has serious clinical consequences, so handling imbalance is still necessary.

Resampling SMOTE

SMOTE is applied only to the training data to generate synthetic samples from the minority class (Presence) through linear interpolation between nearest neighbors (k-nearest neighbors), until the numbers of both classes are balanced. The test data is left untouched to ensure the evaluation reflects the real-world data.

Modeling

Two algorithms were trained and compared: Random Forest and SVM. Each was run in two scenarios without SMOTE (baseline) and with SMOTE + Hyperparameter Tuning. Hyperparameter optimization was performed using GridSearchCV with 5-fold cross-validation to objectively find the best configuration for each model.

Evaluation

The performance of each model was evaluated using Confusion Matrix, Accuracy, Precision, Recall, F1-Score, AUC-ROC, as well as sensitivity and specificity on the test data. Comparisons were conducted between four scenarios (RF without SMOTE, RF with SMOTE, SVM without SMOTE, SVM with SMOTE) to determine the best algorithm and configuration for predicting heart disease.

4. Results and Discussion

Data Collection

The dataset contains 2,026 patient records with 13 feature attributes and 1 target variable, for a total of 14 columns. All features in this dataset represent clinical parameters commonly used in the diagnosis of heart disease in healthcare facilities. Details of the variables in the dataset are presented in Table 1 below.

Table 1. Dataset Variable Description.

Feature Name	Data Type	Explanation
Age	Integer	Patient age (29–77 years)
Sex	Integer	Gender (0 = Female, 1 = Male)
Chest Pain Type	Integer	Tipe nyeri dada (1–4)
BP	Integer	Systolic blood pressure (mmHg)
Cholesterol	Integer	Serum cholesterol level (mg/dl)
FBS over 120	Integer	Fasting blood sugar > 120 mg/dl (0 = No, 1 = Yes)
EKG Results	Integer	Resting electrocardiogram results (0, 1, 2)
Max HR	Integer	Maximum heart rate achieved
Exercise Angina	Integer	Exercise-induced angina (0 = No, 1 = Yes)
ST Depression	Float	Exercise-induced ST depression
Slope of ST	Integer	Peak ST segment slope (1, 2, 3)
Number of Vessels Fluro	Integer	Number of major vessels (0–3)
Thallium	Integer	Thallium test results (3 = Normal, 6 = Fixed Defect, 7 = Reversible Defect)
Heart Disease	String	Target Variable (Absence / Presence)

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the stage of gaining a deeper understanding of the characteristics of a dataset before modeling it.

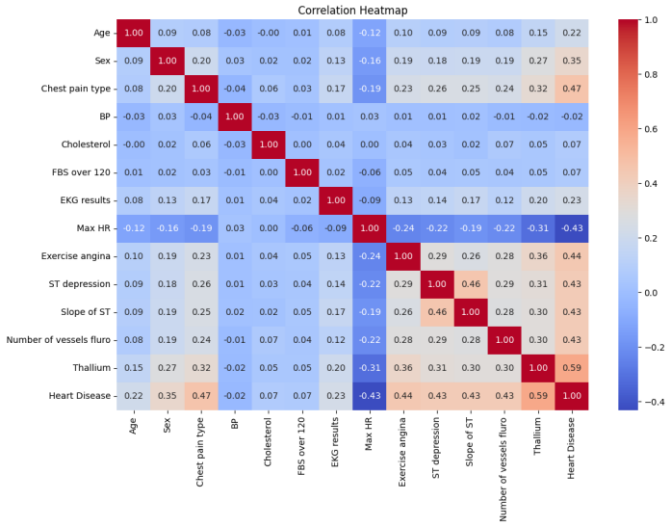


Figure 2. Correlation Heatmap.

Based on the heatmap correlation results, several variables have a fairly strong relationship with heart disease. The variable with the highest correlation is Thallium at 0.59, which indicates that thallium test results have a significant influence on a person's likelihood of developing heart disease. In addition, the variables Chest pain type at 0.47, Exercise angina at 0.44, ST depression at 0.43, Slope of ST at 0.43, and Number of vessels fluro at 0.43 also show a significant positive relationship with heart disease. Meanwhile, the variable Max HR has a negative correlation of -0.43, which indicates that the higher the maximum heart rate, the likelihood of developing heart disease tends to decrease.

Preprocessing Data

The preprocessing stage aims to prepare the data for use in the machine learning modeling process. The id column was removed from the dataset because it is a unique identifier that has no predictive value for the target variable. The results of checking the dataset showed that there were no missing values in all 14 features or the target variable. Thus, no data imputation technique was required at this stage. The target variable heart disease, which has string values (Absence and Presence), was converted into binary numeric values using Label Encoding: Absence = 0 and Presence = 1. This conversion is necessary so that the target variable can be processed by the machine learning algorithm. Feature normalization was performed using the StandardScaler technique (z-score normalization), which transforms the values of each feature so that they have a mean of 0 and a standard deviation of 1. This process is especially crucial for the SVM algorithm, which is sensitive to differences in scale between features, because SVM is based on calculating distances in the feature space.

Splitting Data

The dataset was divided into two subsets using the holdout method, with 80% training data and 20% testing data. Of the 2,026 datasets, 1,620 were used to train the model, and 406 were used to test the model's performance on previously unseen data.

Imbalance Class

Before applying the balancing technique, it is necessary to first determine whether the dataset experiences significant class imbalance. The class distribution in the dataset shows: Absence: 1,113 samples (54.9%) and Presence: 913 samples (45.1%), resulting in an imbalance ratio of 1.22:1.

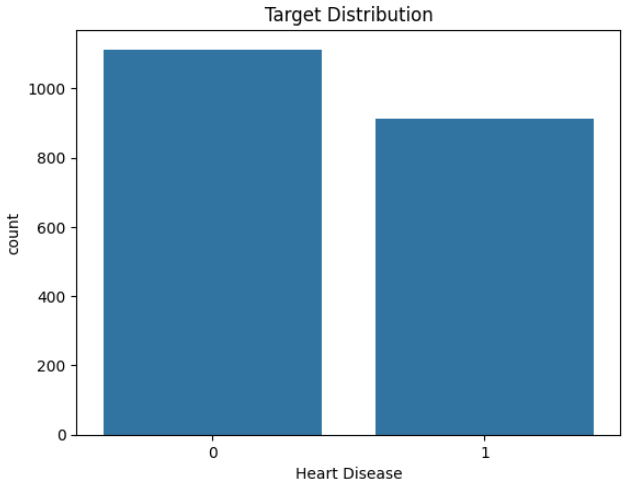


Figure 3. Target Distribution.

Although this imbalance rate is relatively low compared to other medical cases, in the context of heart disease prediction, any misclassification in the minority class (Presence) has serious clinical consequences. Failing to identify patients with heart disease (False Negatives) is far more dangerous than diagnosing healthy patients as having the disease (False Positives). Therefore, class balancing techniques are still needed to optimize the model's sensitivity in detecting positive cases.

SMOTE

To handle class imbalance, this study applies the Synthetic Minority Over-sampling Technique (SMOTE) technique. SMOTE works by balancing the dataset through inserting synthetic samples among the existing minority class samples, generating new samples through linear interpolation with the k-nearest neighbor algorithm.

```
Distribution after SMOTE
Heart Disease
0    890
1    890
Name: count, dtype: int64
```

Figure 4. SMOTE Result.

SMOTE is only applied to the training set and not to the testing set. It is crucial to maintain the integrity of the model evaluation the test data must reflect the distribution of real-world data. After applying SMOTE to the training data, the class distribution in the training set becomes balanced with an equal number of samples between the Absence and Presence classes.

Evaluation

A comprehensive model performance evaluation was conducted using several metrics relevant to binary classification problems in medical data. All evaluations were conducted on a test set that was never used during the training or tuning process.

- Random Forest

Table 2. Random Forest Result.

	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Sensitivity	Specificity
without SMOTE + Tuning	0.8596	0.85	0.8360	0.8429	0.9342	0.8361	0.8789
SMOTE + Tuning	0.8744	0.8587	0.8634	0.8610	0.9449	0.8634	0.8834

Random Forest model testing was conducted in two scenarios: without SMOTE and Hyperparameter Tuning, and with SMOTE and Hyperparameter Tuning. The evaluation results showed that SMOTE and Hyperparameter Tuning consistently improved model performance across all measured metrics.

In the first scenario (without SMOTE and Tuning), the Random Forest model produced an accuracy of 85.96%, with a precision value of 0.85, a recall of 0.8360, and an F1-Score of 0.8429. The AUC-ROC value obtained of 0.9342 indicates that the model has good discrimination ability in distinguishing patients with and without heart disease. In terms of detection ability, the model recorded a sensitivity value of 0.8361 and a specificity of 0.8789, which means the model is better at identifying patients who do not have heart disease (Absence class) than detecting patients who actually have heart disease (Presence class).

In the second scenario (with SMOTE and Hyperparameter Tuning), all evaluation metrics improved. Accuracy increased to 87.44%, precision to 0.8587, recall to 0.8634, and F1-Score to 0.8610. The most significant improvements were seen in recall and sensitivity, which rose from 0.8360 to 0.8634, indicating that the application of SMOTE successfully improved the model's ability to detect positive cases of heart disease that were previously likely missed due to class imbalance. The AUC-ROC value also increased from 0.9342 to 0.9449, confirming that the optimized model had better overall classification ability. Specificity also increased from 0.8789 to 0.8834, indicating that the increased sensitivity did not significantly compromise the model's ability to identify negative cases.

Overall, these results show that the combination of SMOTE and Hyperparameter Tuning makes a positive contribution to the performance of the Random Forest model, with the most significant improvement in the detection ability of Presence cases which is the most critical aspect in the context of clinical heart disease prediction.

- Support Vector Machine

Table 3. SVM Result.

	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Sensitivity	Specificity
without SMOTE + Tuning	0.8596	0.8663	0.8142	0.8394	0.937	0.8142	0.8969
SMOTE + Tuning	0.8645	0.8636	0.8306	0.8468	0.9396	0.8306	0.8924

Support Vector Machine (SVM) model testing was also conducted in two similar scenarios: without SMOTE and Hyperparameter Tuning, and with SMOTE and Hyperparameter Tuning. Overall, the evaluation results showed that SMOTE and Hyperparameter Tuning

improved the performance of the SVM model, albeit at a more moderate level than Random Forest.

In the first scenario (without SMOTE and Tuning), the SVM model achieved an accuracy of 85.96%, the same as Random Forest in the same scenario. However, there are quite striking differences in characteristics: SVM recorded a precision value of 0.8663, which is higher than Random Forest, but a lower recall value of 0.8142. This condition indicates that without optimization, SVM tends to be more conservative in predicting the Presence class only predicting positive when the model is absolutely certain, resulting in high precision but many positive cases are missed (low recall). This is also reflected in the sensitivity value of 0.8142 and a fairly high specificity of 0.8969, indicating that the model is more inclined to recognize the Absence class than the Presence class. The AUC-ROC value of 0.9370 indicates good overall model discrimination ability.

In the second scenario (with SMOTE and Hyperparameter Tuning), improvements were seen in almost all evaluation metrics. Accuracy increased to 86.45%, recall increased from 0.8142 to 0.8306, and F1-Score increased from 0.8394 to 0.8468. The increase in recall and sensitivity values from 0.8142 to 0.8306 indicates that the application of SMOTE successfully pushed the model to be more sensitive in detecting Presence cases that were previously likely to be missed. The AUC-ROC value also increased from 0.9370 to 0.9396, confirming an improvement in the model's overall classification ability. One thing worth noting, the specificity value decreased slightly from 0.8969 to 0.8924, which is a natural consequence of increased sensitivity a more aggressive model detecting positive cases potentially slightly increases False Positives in the negative class.

Overall, the application of SMOTE and Hyperparameter Tuning to the SVM model has a positive impact, especially on the ability to detect Presence cases, although the trade-off between sensitivity and specificity still needs to be considered in the context of its clinical use.

- Comparison of RF and SVM Performance

Table 4. RF and SVM Performance Comparison Table.

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Sensitivity	Specificity
Random Forest	0.8744	0.8587	0.8634	0.8610	0.9449	0.8634	0.8834
SVM	0.8645	0.8636	0.8306	0.8468	0.9396	0.8306	0.8924

Random Forest demonstrated significant improvement over SVM across nearly all metrics. RF achieved 87.44% accuracy compared to SVM's 86.45%, a difference of 0.99%. In the recall and sensitivity metrics, RF again outperformed SVM with a score of 0.8634 compared to SVM's 0.8306 a significant difference in a clinical context, given that higher recall means fewer missed heart disease cases. RF's F1-Score of 0.8610 was also higher than SVM's 0.8468, indicating a better precision-recall balance. RF's AUC-ROC in this scenario reached 0.9449, surpassing SVM's 0.9396. In terms of optimization impact, SMOTE and Hyperparameter Tuning had a greater impact on Random Forest than SVM. RF experienced an increase in accuracy of +1.48%, recall of +2.74%, F1-Score of +1.81%, and AUC-ROC of +1.07% after optimization. Meanwhile, SVM only experienced an increase in accuracy of +0.49%, recall of +1.64%, and AUC-ROC of +0.26%. This indicates that Random Forest responded better to data balancing and optimal hyperparameter search than SVM on this dataset.

Based on a comprehensive comparison of all evaluation metrics, Random Forest with SMOTE and Hyperparameter Tuning was determined to be the best model in this study. This model excelled in accuracy (87.44%), recall (0.8634), F1-Score (0.8610), AUC-ROC (0.9449), and sensitivity (0.8634) the most relevant metrics in the context of heart disease prediction. High recall and sensitivity values are particularly important considerations, as in medical applications, the ability to detect all positive cases of heart disease is far more critical than simply achieving high overall accuracy. Therefore, the optimized Random Forest model is considered most suitable for development as a clinical decision support system for heart disease prediction.

Comparison

The results of this study demonstrate superior performance compared to several previous studies using similar algorithms for heart disease prediction. A comprehensive comparison is presented in the following table.

Table 5. Comparison Table with Previous Research.

Research	Method	Accuracy	Precision	Recall	F1 Score	AUC-ROC
[7]	RF + KDD	0.62	0.61	0.61	0.61	0.66
[7]	SVM + KDD	0.65	0.70	0.68	0.64	0.68
This Research	RF + SMOTE + Tuning	0.8744	0.8587	0.8634	0.8610	0.9449
This Research	SVM + SMOTE + Tuning	0.8645	0.8636	0.8306	0.8468	0.9396

Overall, the model developed in this study performed significantly better than previous studies. The Random Forest model with SMOTE and Hyperparameter Tuning achieved an accuracy of 87.44%, significantly surpassing the RFC results of the previous study, which only achieved 62%. Similarly, the SVM model achieved an accuracy of 86.45%, compared to 65% in the previous study. The most significant difference was seen in the AUC-ROC values, with this study producing values of 0.9449 (RF) and 0.9396 (SVM), while the previous study only achieved values of 0.66–0.68, which are considered poor classification. This indicates that the model in this study has significantly better discrimination ability in distinguishing patients with heart disease from those without.

5. Conclusions

This study compared the performance of the Random Forest (RF) and Support Vector Machine (SVM) algorithms in predicting heart disease using SMOTE and Hyperparameter Tuning techniques on a dataset of 2,026 patient records. Experimental results showed that Random Forest with SMOTE and Hyperparameter Tuning produced the best performance with an accuracy of 87.44%, a recall of 0.8634, an F1-Score of 0.8610, and an AUC-ROC of 0.9449, outperforming SVM, which achieved an accuracy of 86.45% and an AUC-ROC of 0.9396 in the same scenario.

These findings address the research objective: the combination of SMOTE and Hyperparameter Tuning consistently improves the performance of both models. Random Forest proved more optimal than SVM for predicting heart disease on this dataset, particularly in the recall and sensitivity metrics, which are the most critical indicators in the context of medical diagnosis to minimize the risk of false negatives.

Practically, this study provides empirical evidence that the simultaneous integration of data balancing and hyperparameter optimization techniques can produce an excellent classification model (AUC > 0.90), which has the potential to be developed as a clinical decision support system for early detection of heart disease. This study has limitations in terms of model generalization due to the use of only one public dataset. Future research is recommended to validate the model using local medical record data, explore advanced ensemble algorithms such as XGBoost or LightGBM, and apply SHAP value analysis to improve model interpretability for healthcare professionals.

References

Stark, B. A., et al. (2025). Global, regional, and national burden of cardiovascular diseases and risk factors in 204 countries and territories, 1990–2023. JACC, 86(22), 2167–2243. https://doi.org/10.1016/j.jacc.2025.08.015

Banerjee, T., & Paçal, İ. (2025). A systematic review of machine learning in heart disease prediction. Turkish Journal of Biology, 49(SI-5), 600–634. https://doi.org/10.55730/1300-0152.2766

Humaira, A. P., & Ningsih, S. (2026). Perbandingan algoritma SVM dan Random Forest untuk prediksi penyakit jantung menggunakan metode feature selection ANOVA F value. Jurnal Sistem Informasi Bisnis, 7, 7–17. https://doi.org/10.55122/junsibi.v7i1.2047

Alsabhan, W., & Alfadhly, A. (2025). Effectiveness of machine learning models in diagnosis of heart disease: A comparative study. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-09423-y

Baddah, W., Qasem, H. A., Alsabry, A., Al Gawani, R. S., Alzuraiqi, W. M., & Hanash, F. E. (2024). Optimizing heart disease prediction models through SMOTE: Addressing data imbalance. In 4th International Conference on Emerging Smart Technologies and Applications (eSmarTA 2024). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/eSmarTA62850.2024.10638899

Ahamad, G. N., et al. (2023). Influence of optimal hyperparameters on the performance of machine learning algorithms for predicting heart disease. Processes, 11(3). https://doi.org/10.3390/pr11030734

Alfajr, N. H., Garno, G., & Yusup, D. (2025). Studi komparasi algoritma Random Forest Classifier dan Support Vector Machine dalam prediksi penyakit jantung. Jurnal Informatika dan Teknik Elektro Terapan, 13(3). https://doi.org/10.23960/jitet.v13i3.6569

Wanyonyi, M., Morris, Z. N., Musyoka, F. M., & Kitavi, D. M. (2025). Enhanced machine learning and hybrid ensemble approaches for coronary heart disease prediction. PLOS ONE, 20(12), 1–51. https://doi.org/10.1371/journal.pone.0328338

Nasution, N., Hasan, M. A., & Nasution, F. B. (2025). Predicting heart disease using machine learning: An evaluation of logistic regression, Random Forest, SVM, and KNN models on the UCI heart disease dataset. IT Journal Research and Development, 9(2), 140–150. https://doi.org/10.25299/itjrd.2025.17941

Rasheed, S., Kumar, G. K., Rani, D. M., Kantipudi, M. V. V. P., & Anila, M. (2024). Heart disease prediction using GridSearchCV and Random Forest. EAI Endorsed Transactions on Pervasive Health and Technology, 10. https://doi.org/10.4108/eetpht.10.5523

Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics, 14(2). https://doi.org/10.3390/diagnostics14020144

Roopa, T., & Ramanjinappa, G. D. (2025). Heart disease predictive modeling with XGBoost and SMOTE-driven class imbalance mitigation. Engineering, Technology and Applied Science Research, 15(6), 29914–29918. https://doi.org/10.48084/etasr.14301

Mohamed, A., Abdelrehim, M., & Al-Barazie, R. (2025). Context matters in machine learning based disease prediction with insights from diverse clinical and symptom data. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-26855-8

Bulagang, A. F., Weng, N. G., Mountstephens, J., & Teo, J. (2020). A review of recent approaches for emotion classification using electrocardiography and electrodermography signals. Information in Medicine Unlocked, 20, 100363. https://doi.org/10.1016/j.imu.2020.100363

Mienye, I. D., & Jere, N. (2024). Optimized ensemble learning approach with explainable AI for improved heart disease prediction. Information, 15(7). https://doi.org/10.3390/info15070394

Lin, J. (2025). In-depth analysis, model optimization, and interpretability of heart disease risk factors using machine learning and stacking ensembles. In Proceedings of the 2nd International Conference on Image Processing, Machine Learning, and Pattern Recognition (IPMLP 2025) (pp. 177–185). Association for Computing Machinery. https://doi.org/10.1145/3759928.3759957

Wang, J., He, F., & Sun, S. (2023). Construction of a new smooth support vector machine model and its application in heart disease diagnosis. PLOS ONE, 18(2). https://doi.org/10.1371/journal.pone.0280804

Hoque, R., Billah, M., Debnath, A., Hossain, S. M. S., & Sharif, N. B. (2024). Heart disease prediction using SVM. International Journal of Science and Research Archive, 11(2), 412–420. https://doi.org/10.30574/ijsra.2024.11.2.0435

Asif, D., Bibi, M., Arif, M. S., & Mukheimer, A. (2023). Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization. Algorithms, 16(6). https://doi.org/10.3390/a16060308

Olawade, D. B., Soladoye, A. A., Omodunbi, B. A., Aderinto, N., & Adeyanju, I. A. (2025). Comparative analysis of machine learning models for coronary artery disease prediction with optimized feature selection. International Journal of Cardiology, 436. https://doi.org/10.1016/j.ijcard.2025.133443

Nizam-Ozogur, H., & Orman, Z. (2024). A heuristic-based hybrid sampling method using a combination of SMOTE and ENN for imbalanced health data. Expert Systems, 41(8). https://doi.org/10.1111/exsy.13596

Kahraman, A. (2025). Machine learning techniques for improved prediction of cardiovascular diseases using integrated healthcare data. Frontiers in Artificial Intelligence, 8. https://doi.org/10.3389/frai.2025.1694450

Hairani, H., Widiyaningtyas, T., & Prasetya, D. D. (2024). Addressing class imbalance of health data: A systematic literature review on modified Synthetic Minority Oversampling Technique (SMOTE) strategies. International Journal on Informatics Visualization, 8, 1310–1318. www.joiv.org/index.php/joiv

Yang, Y., Khorshidi, H. A., & Aickelin, U. (2024). A review on over-sampling techniques in classification of multi-class imbalanced datasets: Insights for medical problems. Frontiers in Digital Health, 6. https://doi.org/10.3389/fdgth.2024.1430245

Patel, H. B. (2026). An optimization of cardiovascular disease prediction by incorporating hyperparameters tuning in machine learning algorithms. International Journal of Scientific Research in Computer Science and Engineering, 14(1), 1–9. https://doi.org/10.26438/ijsrcse.v14i1.799