Journal on
Artificial Intelligence

Tech Science Press

https://doi.org/10.32604/jai.2026.078359

ARTICLE

Comparative Performance Analysis of Machine Learning Algorithms for Early Detection of Heart Disease

Kadriye Simsek Alan* **and Busra Senel Kahyaoglu**

Department of Mathematical Engineering, Faculty of Chemical and Metallurgical Engineering, Yildiz Technical University, Istanbul, Turkey

*Corresponding Author: Kadriye Simsek Alan. Email: ksimsek@yildiz.edu.tr

Received: 30 December 2025; Accepted: 06 March 2026; Published: 15 April 2026

ABSTRACT: Cardiovascular diseases remain one of the leading causes of mortality worldwide, making early and reliable diagnosis a critical challenge for modern healthcare systems. In this study, a systematic comparative performance analysis of widely used machine learning algorithms is conducted for the early detection of heart disease using tabular clinical data. Rather than proposing a novel model architecture, the primary objective is to provide a fair, reproducible, and clinically meaningful evaluation of commonly adopted classifiers under consistent experimental conditions. The Kaggle Heart Failure dataset is employed, and multiple machine learning models—including tuned Random Forest, tuned XGBoost, and a soft voting ensemble—are evaluated using a unified preprocessing pipeline, hyperparameter optimization strategy, and validation protocol. Model performance is assessed using multiple evaluation metrics, including accuracy, sensitivity, specificity, F1-score, and ROC–AUC, to capture both overall predictive performance and clinically relevant error trade-offs. The experimental results demonstrate that while moderate accuracy values are obtained, the proposed models achieve strong ROC–AUC performance and balanced sensitivity–specificity characteristics, indicating robust discriminative capability across different decision thresholds. These findings highlight the limitations of relying solely on accuracy, particularly in class-imbalanced clinical datasets, and emphasize the importance of multi-metric evaluation for reliable clinical decision support. Overall, this study contributes a transparent and methodologically rigorous comparative framework that facilitates objective assessment of machine learning models for heart disease prediction and supports informed model selection in healthcare applications.

KEYWORDS: Heart disease detection; machine learning; algorithms; logistic regression; support vector machines (SVM); random forest

1 Introduction

Cardiovascular diseases constitute one of the most critical fields of modern medicine and remain the leading cause of death worldwide. According to the most recent reports published by the World Health Organization (WHO), approximately 17.9 million people die each year due to cardiovascular diseases, accounting for nearly one third of all global deaths [1]. Most of these conditions arise from preventable risk factors, and they can be significantly controlled through early diagnosis, accurate risk stratification, and timely medical intervention. Therefore, the early detection of heart diseases has become a strategic necessity not only for individual health but also for reducing the financial burden on healthcare systems.

Traditional diagnostic methods often rely heavily on expert judgment and clinical experience. However, with the advancement of data acquisition technologies, large volumes of biomedical data are now

being generated through electronic health records (EHRs), wearable devices, and IoT-based monitoring systems [2]. The increasing availability of such heterogeneous and high-dimensional clinical data has exposed the limitations of purely rule-based or experience-driven diagnostic approaches, particularly in capturing complex nonlinear relationships among risk factors. To effectively analyze such data, artificial intelligence, machine learning (ML), and deep learning (DL) methods have become increasingly important in the healthcare domain [3,4].

Recent studies have demonstrated that well-trained data-driven models can outperform conventional diagnostic approaches in the early detection of heart diseases by leveraging heterogeneous clinical and physiological data sources [4]. In particular, hybrid and ensemble-based machine learning frameworks have been successfully applied to heart disease prediction [5], while deep learning architectures have also shown promising performance in improving diagnostic accuracy and robustness across different datasets [6]. These approaches benefit from their ability to integrate complementary decision boundaries, reduce model variance, and enhance generalization performance, which is especially valuable in clinical prediction tasks involving noisy or incomplete data [7]. These findings have contributed to the growing recognition of machine learning techniques as essential complementary tools in modern cardiology and clinical decision support systems.

Despite these advances, existing studies on heart disease prediction exhibit considerable methodological diversity, particularly with respect to dataset selection, preprocessing strategies, feature engineering techniques, and evaluation protocols. As a result, reported performance levels vary widely across studies, even when similar datasets and algorithms are employed. This heterogeneity complicates direct comparison of reported results and limits the ability to draw reliable conclusions regarding the relative effectiveness of different machine learning approaches [8].

However, a closer examination of the literature reveals several important methodological gaps. Many studies primarily rely on accuracy-based performance reporting, which may provide an incomplete or misleading assessment of clinical usefulness, particularly in the presence of class imbalance [9]. Moreover, validation strategies are often inconsistently applied, and limited attention is paid to the statistical reliability and generalizability of reported results, which may lead to overly optimistic performance estimates [10]. In addition, comprehensive and statistically grounded comparative analyses across multiple classifiers remain relatively scarce in heart disease prediction research, despite their importance for reliable model comparison [8].

Motivated by these observations, the primary aim of this study is not to introduce a new predictive model, but rather to conduct a systematic and reproducible comparative performance analysis of widely used machine learning algorithms for early heart disease detection. The novelty of the proposed work lies in its focus on fair benchmarking under consistent experimental conditions, including standardized preprocessing, robust validation strategies, multi-metric evaluation, and statistically sound comparison of classifiers. By addressing key methodological limitations observed in recent studies, this work aims to provide clearer insights into the strengths and limitations of commonly adopted machine learning approaches for heart disease prediction using tabular clinical data.

2 Literature Review

The application of machine learning and deep learning techniques for the early detection of heart diseases has shown rapid growth in the academic literature in recent years. An examination of existing studies indicates that the literature can generally be categorized into four main themes: wearable technologies and data acquisition systems, classical machine learning–based classification methods, deep learning and

advanced artificial intelligence models, and hybrid models together with explainable artificial intelligence (XAI) approaches. These studies are thematically presented below.

2.1 Wearable Technologies and Data Acquisition Systems

Wearable technologies have become an important research area in recent years for the monitoring and early detection of heart diseases. The large-scale and continuous physiological data obtained through these technologies not only support clinical monitoring processes but also provide valuable data sources for machine learning–based diagnostic and predictive models. The systematic review conducted by Naseri Jahfari and Tax [11] revealed the growing importance of wearable devices in the detection of cardiovascular diseases and emphasized that these technologies should be evaluated within the framework of technology readiness levels (TRL). The study highlighted key challenges prior to the integration of wearable sensors into clinical practice, including data reliability, calibration, validation, and real-world performance.

Focusing on IoT-based wearable systems, Benali [12] examined the design and development processes of smart wearable devices developed for continuous cardiovascular monitoring. The study emphasized that the real-time and continuous monitoring of key cardiovascular parameters such as heart rate, blood pressure, and electrocardiography (ECG) enables the early detection of cardiac abnormalities and timely medical intervention. Furthermore, the potential of IoT-enabled wearable devices to improve heart health management was evaluated in the context of technological advancements and design criteria, while current limitations and future research directions were also discussed.

In a comprehensive review addressing technological advancements in cardiovascular wearable devices, Iqbal et al. [13] investigated both traditional implantable devices and modern wearable solutions. The study reported that wearable devices offer significant advantages due to their low cost and ability to provide continuous real-time monitoring. Wearable technologies were classified into galvanic contact–based, photoplethysmography (PPG)–based, and radio frequency (RF)–based systems. Additionally, the authors highlighted the increasingly important role of artificial intelligence–based methods in the diagnostic processes of cardiovascular diseases and stated that wearable cardiovascular devices are expected to be more effectively utilized in clinical applications in the future.

Similarly, Lin et al. [14] provided a detailed review of recent developments in wearable and flexible sensor technologies used for real-time cardiovascular disease monitoring. The study emphasized the importance of body-conformal and flexible sensors due to the need for long-term monitoring in cardiovascular diseases. It was reported that advanced materials, integrated electronic systems, the Internet of Things (IoT), and edge computing technologies enable the real-time measurement of various physiological signals, including pulse, ECG, phonocardiogram (PCG), seismocardiogram/ballistocardiogram (SCG/BCG), and apexcardiogram (ACG).

Overall, the literature indicates that wearable technologies hold significant potential in cardiovascular data acquisition and monitoring processes. These systems, which can provide continuous and real-time data, offer a strong infrastructure for both clinical decision support mechanisms and data-driven artificial intelligence applications.

2.2 Heart Disease Prediction Using Classical Machine Learning Methods

Classical machine learning algorithms have long been among the most widely used techniques for heart disease classification. Many studies in this field have been conducted using well-known datasets such as UCI Cleveland, Heart Disease, and Framingham.

The Machine Learning–Based Cardiovascular Disease Diagnosis Framework (MaLCaDD) proposed by Rahim et al. [15] incorporates important data preprocessing steps, including the imputation of missing values using mean substitution and the mitigation of data imbalance through the Synthetic Minority Over-sampling Technique (SMOTE). The framework demonstrated strong performance, achieving accuracy rates of 99.1%, 98.0%, and 95.5% on the Framingham, Heart Disease, and Cleveland datasets, respectively.

Similarly, Amzad Hossen et al. [16] compared the performance of logistic regression and other machine learning algorithms using the UCI Cleveland dataset and reported that logistic regression achieved the best result with an accuracy of 92.10%. The effectiveness of logistic regression was further supported by the study conducted by Ambrish et al. [17], which achieved an accuracy of 87.10% on the UCI dataset. Azmi et al. [18] emphasized that integrating big data analytics with machine learning techniques enhances heart disease prediction performance, particularly highlighting the high success of classification algorithms.

Khan et al. [19] compared various machine learning techniques and demonstrated that support vector machines (SVM) and artificial neural networks (ANN) achieved high accuracy rates across multiple scenarios. Berdinanth et al. [20] also identified logistic regression and decision tree algorithms as effective methods for heart attack prediction. These studies indicate that classical machine learning methods provide a strong foundation, especially when applied to tabular clinical data.

Finally, Kumar et al. [21] compared SVM and ANN models, reporting that SVM was more sensitive in detecting positive cases, whereas ANN achieved higher overall accuracy. These findings suggest that different algorithms may offer distinct advantages depending on the application scenario.

2.3 Deep Learning–Based Methods

Deep learning models are increasingly employed in heart disease prediction due to their ability to capture complex patterns within data. Arooj et al. [22] applied a Deep Convolutional Neural Network (DCNN) model to the UCI dataset and achieved an accuracy of 91.7%, demonstrating that deep learning methods are particularly effective in modeling complex relationships among features.

Allheeib et al. [23] developed a hybrid framework combining machine learning and deep learning techniques and achieved an exceptionally high accuracy of 99.99% on real-world datasets. Similarly, Jamila and Roy [24] utilized a Vision Transformer (ViT)–based model to detect valvular heart diseases from phonocardiography (PCG) signals, achieving an accuracy of 99.90% and an F1 score of 99.95%, thereby outperforming existing deep learning models.

Mansoor et al. [25] analyzed coronary artery disease using deep learning techniques and obtained an accuracy of 92%. This study demonstrates that deep learning methods serve as effective tools for high-dimensional clinical datasets.

These findings indicate that deep learning models exhibit strong performance, particularly when applied to datasets involving signals, images, or large feature spaces.

2.4 Hybrid Models and Feature Engineering

Hybrid models aim to achieve higher accuracy by integrating the strengths of different machine learning techniques. Qadri et al. [26] enhanced the feature engineering process through the selection of significant features and achieved a classification accuracy of 100%, highlighting the critical impact of proper feature selection on model performance.

Cuevas-Chávez et al. [27] examined machine learning techniques in conjunction with IoT and IoMT technologies and reported that algorithms such as neural networks and XGBoost achieved accuracy rates

exceeding 90%. However, the authors emphasized that the lack of large-scale datasets remains a limiting factor in advancing this field.

Baghdadi et al. [28] employed the CatBoost algorithm and achieved high accuracy rates in the early detection of heart diseases, suggesting that the integration of this approach into clinical decision support systems could have positive effects on healthcare services. In another related study by the same research group, emphasis was placed on feature selection and engineering for risk prediction.

2.5 Explainable Artificial Intelligence (XAI)

The importance of explainable artificial intelligence (XAI) approaches in the healthcare domain has been steadily increasing. El-Sofany et al. [29] combined the XGBoost algorithm with SHAP-based explainable AI techniques and achieved an accuracy of 97.57%, providing a framework capable of explaining how model decisions are formed to healthcare professionals. In another study published in the same year, El-Sofany [30] developed a low-cost and computationally efficient model, demonstrating the significant role of XAI methods in early diagnosis processes.

Evolutionary and statistical techniques such as genetic algorithms (GA) and the Recursive Feature Elimination Method (RFEM) have also been applied in the literature. Al-Alshaikh et al. [31] achieved an accuracy of 95.5% using a GA + RFEM combination, showing that feature selection substantially enhances model performance.

Finally, Elhadjamor and Harbaoui [32] compared various machine learning methods and reported that the HistGradientBoosting algorithm achieved the highest accuracy. This result indicates that boosting-based methods represent a strong alternative for medical classification problems.

2.6 Recent Comprehensive Comparative Studies

Recent comparative studies on machine learning methods for the early detection of heart diseases systematically present the classification performance of different algorithms. Mandal et al. [33] compared twelve machine learning algorithms and reported that the SVM algorithm achieved the best performance with an accuracy of 89% and a ROC–AUC value of 92%. Similarly, Qian [34] evaluated Logistic Regression, Decision Trees, Random Forest, and XGBoost algorithms using feature engineering and hyperparameter optimization, showing that the Random Forest model provided more balanced and robust performance due to its ensemble structure.

In another study, Temirbayeva and Altybay [35] compared Logistic Regression, SVM, Decision Tree, Random Forest, and Gradient Boosting algorithms using clinical data from 270 patients and found that the Random Forest algorithm outperformed others in terms of ROC–AUC. A broader analysis was presented by Hussain and Aslam [36], who evaluated nine machine learning algorithms using metrics such as Accuracy, Recall, and ROC, comprehensively revealing the strengths and weaknesses of different approaches. These studies clearly demonstrate that algorithm selection in heart disease prediction depends on data characteristics and evaluation metrics.

To synthesize these comparative findings and provide a structured overview of recent benchmark-oriented studies on heart disease prediction, representative works published between 2022 and 2025 are summarized in Table 1. The table focuses on studies employing tabular clinical datasets and reporting commonly used evaluation metrics under heterogeneous experimental settings, thereby enabling a consistent and fair contextualization of the proposed approach within the existing literature.

Table 1: Comparative analysis of recent studies on heart disease prediction (2022–2025).

Study (Year)	Dataset	Sample Size	Main Method	Key Metrics	Best Reported Performance
Zamani et al. [2].	Multi-disease IoT datasets	1190 (Dataset 3)	Ensemble (NN, KNN, Fuzzy)	Acc, ROC	Relative improvement up to 40.1%
Almutairi & Dardouri [3].	Kaggle Heart Failure	918	XGBoost + SVM	Acc, F1	Acc = 89.3%, F1 = 0.905
Ogunpola et al. [4].	Cleveland/CHD	303/1000	XGBoost	Acc	Acc = 98.5%
Bhatt et al. [5].	Kaggle CVD	70,000	MLP, RF, XGBoost	Acc, AUC	Acc = 87.28%
This study	**Kaggle Heart Failure**	**918**	**Tuned XGBoost, Tuned RF, Soft Voting Ensemble**	**Acc, AUC**	**Acc = 84.78%, AUC = 0.9316**

As shown in Table 1, recent studies report a wide range of performance levels even when similar datasets and algorithms are employed. This variability highlights the influence of experimental design choices, including preprocessing strategies, feature engineering, validation protocols, and evaluation metrics. In this context, the present study contributes by offering a controlled and reproducible comparative evaluation under consistent experimental conditions, enabling more reliable assessment of commonly used machine learning classifiers for heart disease prediction.

2.7 Limitations and Research Gaps in the Literature

Despite the growing body of research on machine learning–based heart disease prediction, several gaps remain in how existing studies are designed, evaluated, and interpreted. A large portion of the literature relies on a single publicly available dataset and reports performance primarily in terms of accuracy, often under heterogeneous or weakly specified evaluation protocols. This practice limits the comparability and reproducibility of reported results and may obscure clinically relevant trade-offs between different types of classification errors.

More importantly, many comparative studies do not provide a sufficiently detailed analysis of model behavior beyond aggregate performance scores. In clinical decision-support settings, understanding how models balance sensitivity and specificity, how errors are distributed, and how performance changes across decision thresholds is often as critical as achieving high predictive accuracy. However, these aspects are frequently underexplored or inconsistently reported in prior work.

Motivated by these observations, the primary contribution of this study is not the introduction of a new dataset or a novel model architecture, but rather a systematic and reproducible comparative evaluation of widely used machine learning algorithms under a strictly controlled experimental setting. By employing consistent preprocessing, feature selection, hyperparameter optimization, and a comprehensive set of evaluation metrics—including sensitivity, specificity, F1-score, ROC-AUC, and error distribution analysis—this study provides a transparent performance baseline for heart disease classification.

Although model-level explainability techniques are an important direction for future research, this study focuses on interpretability at the decision and performance level. The presented analysis enables clinically meaningful model comparison and supports informed model selection depending on different clinical

priorities, such as screening-oriented or confirmatory use cases. In this sense, the proposed framework aims to strengthen methodological rigor and interpretability in comparative machine learning studies for cardiovascular risk assessment.

3 Materials and Methods

In this study, the performance of various machine learning (ML) algorithms used for predicting heart diseases was compared. The applied classification methods include Naive Bayes, Support Vector Machines (SVM), Voting, XGBoost, AdaBoost, Bagging, Decision Trees (DT), K-Nearest Neighbors (KNN), Random Forest (RF), and Logistic Regression (LR). These algorithms provide effective tools for the accurate diagnosis of heart diseases and offer support to physicians and data analysts.

In this study, up-to-date datasets and relevant literature related to heart diseases were utilized. The proposed approach consists of data collection, feature extraction, and data exploration stages. During the data preprocessing phase, procedures such as handling missing values, data cleaning, and normalization were performed. After preprocessing, the obtained data were classified using the selected machine learning algorithms. Following the implementation of the models, performance and accuracy were evaluated using various evaluation metrics.

3.1 Dataset Description

The dataset used in this study was obtained from the Kaggle repository and corresponds to the Heart Disease dataset made publicly available by Karimsony [37].

The dataset contains information from a total of 920 patients, with 14 attributes for each patient. These attributes encompass demographic information, clinical measurements, and factors related to heart disease. A detailed description of these attributes and their representative values is provided in Table 2.

Table 2: Description of the attributes used in the dataset.

Sr. No.	Attributes	Representation	Description	Type
1	Age	age	Age in years	Integer
2	Gender	sex	Male and female	Binary (1 for male, 0 for female)
3	Chest pain	cp	Four types of chest pain	Categorical
4	Cholesterol level	chol	Measure of cholesterol in mg/dl	Integer
5	Resting blood pressure	trestbps	Blood pressure when the body is at rest	Integer
6	Fasting blood sugar	fbs	Blood sugar level while fasting	Binary (1 for true, 0 for false)
7	Max HR	thalach	Maximal heart rate	Integer
8	Rest ECG	restecg	Resting electrocardiograph	Categorical
9	Exercise-induced angina	exang	Exercise-induced angina	Binary (1 for yes, 0 for no)
10	Oldpeak	oldpeak	ST depression brought by exercise compared to rest	Continuous
11	Slope	slope	Slope of exercise peak	Discrete

(Continued)

Table 2 (continued)

Sr. No.	Attributes	Representation	Description	Type
12	Vessels	ca	Number of major vessels	Continuous
13	Thalassemia	thal	Normal, fixed, and reversible defects	Discrete
14	Heart disease	target	Predicted attribute	Binary

The dataset underwent a multi-step preprocessing procedure before being used for modeling. In the first step, the target column (heart disease) was separated, and the input variables X and target variable y were defined. Given the significant proportion of missing values in the dataset, a systematic imputation strategy was developed to prevent it from negatively impacting model performance. This strategy involved applying different model-based imputation methods for categorical and numerical variables.

3.1.1 Imputation of Categorical Variables

Categorical variables were predicted using classification-based modeling techniques. In this regard, the following imputations were performed:

- fbs (9.78% missing) was imputed with Logistic Regression with 83.37% accuracy.
- restecg (0.22% missing) was imputed with 66.56% accuracy.
- exang (5.98% missing) was imputed with 78.96% accuracy.
- slope (33.59% missing) was imputed with 67.43% accuracy.
- ca (66.41% missing) was imputed with 64.40% accuracy.
- thal (52.83% missing) was imputed with 69.35% accuracy.

For each of these variables, hyperparameter optimization was performed using GridSearchCV, and the prediction was completed using the model configurations that provided the highest accuracy.

3.1.2 Imputation of Numerical Variables

For numerical variables, regression-based imputation was preferred. The following results were obtained for numerical variables:

- trestbps (6.41% missing): MAE 9.42, RMSE 12.67, R^2 0.56.
- chol (3.26% missing): MAE 17.01, RMSE 25.09, R^2 0.95.
- thalach (5.98% missing): MAE 10.66, RMSE 13.65, R^2 0.72.
- oldpeak (6.74% missing): MAE 0.35, RMSE 0.53, R^2 0.76.

The relevant variables were imputed using the Random Forest Regressor and KNN Imputer models, and the most suitable method was selected based on performance criteria. As a result, the missing values were imputed with appropriate model predictions, minimizing data loss.

3.2 Dataset Preparation

After handling missing data appropriately, categorical variables in the dataset were converted into numerical format using the LabelEncoder method. The dataset was then split into training and testing subsets using the train_test_split function, with 80% of the data used for training and 20% for testing. The random_state parameter was fixed at 42 to ensure reproducibility.

To ensure numerical stability and comparability across features, the input variables were standardized using the StandardScaler method, resulting in zero mean and unit variance for each feature. The training set consisted of 736 patient records, while the test set contained 184 records. The training data were used to train the evaluated models, and the test data were reserved exclusively for assessing generalization performance.

The class distribution of the dataset was examined prior to model training. Out of 920 instances, 509 cases (55.33%) correspond to patients with heart disease, while 411 cases (44.67%) represent non-disease instances, indicating a mildly imbalanced class distribution.

To account for this imbalance, evaluation metrics beyond accuracy—such as precision, recall (sensitivity), specificity, F1-score, and ROC-AUC—were employed. In addition, class imbalance was explicitly addressed during model training by using balanced class weights in the Random Forest classifier and probability-based decision thresholds across all evaluated models. The overall workflow of the proposed machine learning framework is illustrated in Fig. 1.

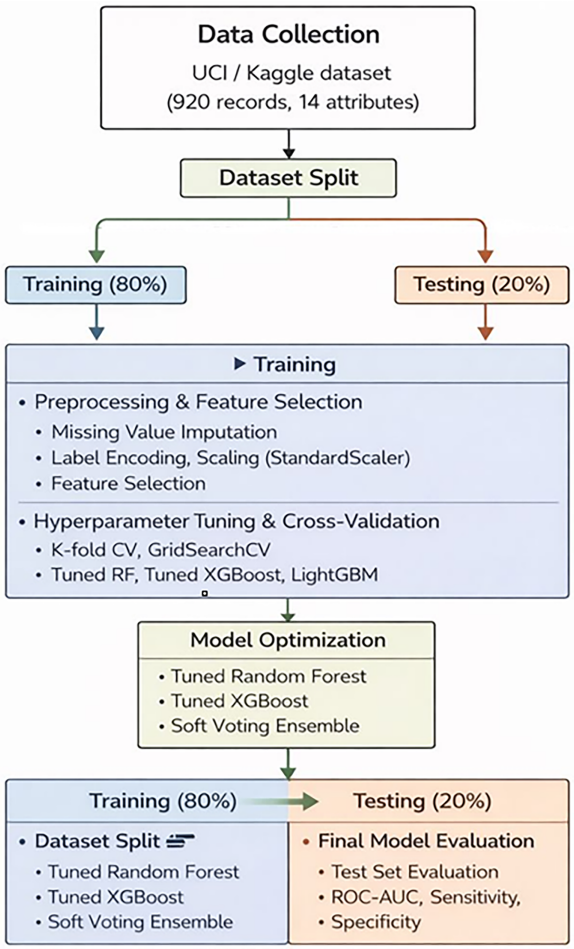


Figure 1: Overview of the optimized machine learning workflow used for heart disease prediction.

3.3 Model Optimization and Ensemble Learning

To improve predictive performance and model robustness, hyperparameter optimization and ensemble learning strategies were integrated into the proposed framework. These techniques were employed to systematically enhance model generalization while preserving interpretability and methodological transparency.

Hyperparameter optimization was performed for the Random Forest and XGBoost classifiers using RandomizedSearchCV combined with stratified cross-validation on the training set. The search was conducted with a fixed number of iterations (N_ITER = 20) under a leakage-free experimental protocol to ensure reproducibility and fair comparison across models.

Building upon the optimized individual models, an ensemble-based approach was subsequently adopted. Specifically, a soft-voting ensemble was constructed by combining multiple optimized tree-based classifiers. By aggregating the probabilistic outputs of the constituent models, the ensemble was designed to leverage complementary decision patterns and improve overall predictive stability.

All experiments were conducted using a fixed train–test split under a strict, leakage-free evaluation protocol to ensure reproducibility and prevent optimistic bias. While some studies report higher absolute accuracy values on this dataset, such results often rely on more complex model architectures or less restrictive evaluation procedures. In contrast, the proposed approach prioritizes robustness, interpretability, and methodological soundness as key design principles.

3.4 Experimental Settings

All machine learning experiments in this study were conducted using Google Colab, a cloud-based platform that provides a Jupyter Notebook environment with a pre-configured Python ecosystem suitable for machine learning research. The experimental environment included widely used libraries such as Scikit-learn, XGBoost, LightGBM, and other standard scientific computing tools.

The virtual machine used during the experiments was configured with approximately 13.2 GB of available RAM and around 108 GB of disk space. The operating system was Ubuntu 22.04.4 LTS (Jammy), running on an x86_64 architecture, which ensured a stable and efficient environment for model development and evaluation.

To ensure full reproducibility and transparency, all experiments were conducted using a fixed stratified train–test split, with 80% of the data allocated for training and 20% reserved for testing. All preprocessing, feature selection, and hyperparameter optimization steps were applied exclusively on the training set to prevent data leakage. The test set was used solely for final model evaluation.

The experimental configuration and the final hyperparameter settings of all evaluated machine learning models are explicitly summarized in Table 3. This setup enables fair model comparison, reproducibility of the reported results, and a clear separation between model optimization and performance assessment.

Table 3: Final hyperparameter configurations of the evaluated machine learning models.

Model	Hyperparameter	Final Value
Random Forest	n_estimators	800
	max_depth	10
	min_samples_split	2
	min_samples_leaf	1
	class_weight	balanced
XGBoost	n_estimators	800
	max_depth	5
	learning_rate	0.05
	subsample	0.85
	colsample_bytree	0.85

(Continued)

Table 3 (continued)

Model	Hyperparameter	Final Value
LightGBM	n_estimators	800
	learning_rate	0.03
	num_leaves	31
Ensemble Model	Base models	RF, XGBoost, LightGBM
	Voting strategy	Soft voting

The final hyperparameter values reported in Table 3 correspond to the configurations that achieved the best performance on the training data during the randomized search procedure with stratified cross-validation. All tuning, feature selection, and preprocessing steps were applied exclusively on the training set, while the test set was reserved solely for final performance evaluation. This experimental protocol ensures reproducibility, fair model comparison, and robustness of the reported results.

3.5 Evaluation Metrics

The performance of the proposed machine learning models was evaluated using multiple complementary metrics to provide a comprehensive and reliable assessment, particularly for a medical diagnosis task. Since relying on a single metric such as accuracy may be misleading in the presence of class imbalance, several widely accepted evaluation measures were employed.

Accuracy represents the proportion of correctly classified instances among all samples and provides an overall indication of classification performance.

Precision measures the proportion of correctly predicted positive cases among all instances predicted as positive, reflecting the reliability of positive predictions.

Recall (Sensitivity) quantifies the ability of a model to correctly identify positive cases and is particularly important in medical screening scenarios, where missing a true positive may have serious clinical consequences.

Specificity measures the ability of the model to correctly identify negative cases, indicating how well false-positive predictions are avoided.

F1-score represents the harmonic mean of precision and recall and provides a balanced evaluation when there is a trade-off between these two metrics.

ROC-AUC (Area Under the Receiver Operating Characteristic Curve) evaluates the discriminative ability of the model across different decision thresholds by measuring its capacity to distinguish between positive and negative classes.

Together, these evaluation metrics enable a balanced and clinically meaningful comparison of model performance and ensure a robust interpretation of the experimental results.

3.5.1 Random Forest

Random Forest (RF) is an ensemble learning algorithm that constructs multiple decision trees using bootstrap sampling and aggregates their predictions to produce a final decision. By introducing randomness in both data sampling and feature selection, Random Forest reduces overfitting and improves generalization performance.

Given an input feature vector x, the final prediction of a Random Forest classifier is obtained through majority voting among T individual decision trees:

$$\hat{y} = \text{mode}\left\{h_1(x), h_2(x), \ldots, h_T(x)\right\}$$

where $h_t(x)$ denotes the prediction of the t-th decision tree.

In this study, the Random Forest classifier was applied to the selected feature set obtained after the feature selection stage. Class imbalance was addressed using balanced class weights, and key hyperparameters were optimized using randomized search with stratified cross-validation. The optimized Random Forest model was subsequently evaluated on the test set for final performance assessment.

3.5.2 XGBoost

Extreme Gradient Boosting (XGBoost) is a gradient boosting framework that builds an ensemble of decision trees sequentially, where each new tree is trained to correct the errors of the previous ensemble. XGBoost incorporates regularization terms and advanced optimization techniques to enhance model robustness and prevent overfitting.

The objective function minimized by XGBoost can be expressed as:

$$\mathcal{L} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right)$$

where $l(\cdot)$ is a differentiable loss function, f_k represents the k-th decision tree, and $\Omega(\cdot)$ denotes a regularization term controlling model complexity.

In this work, XGBoost was trained on the same feature-selected dataset as Random Forest. Hyperparameters such as the number of trees, maximum depth, learning rate, and subsampling ratios were optimized using randomized search. The final tuned XGBoost model demonstrated strong discriminative performance and was included both as an individual classifier and as a base learner in the ensemble model.

3.5.3 LightGBM

Light Gradient Boosting Machine (LightGBM) is a gradient boosting algorithm designed for efficiency and scalability. Unlike traditional level-wise tree growth, LightGBM employs a leaf-wise growth strategy, which can achieve faster convergence and improved accuracy under appropriate regularization.

LightGBM follows the same general boosting objective as gradient boosting methods, iteratively minimizing a loss function while adding decision trees to the ensemble. Its efficiency stems from histogram-based feature discretization and optimized tree growth strategies.

In this study, LightGBM was evaluated as a comparative baseline model. It was trained on the selected features using fixed hyperparameter settings to assess its performance relative to the optimized Random Forest and XGBoost models. Although computationally efficient, LightGBM exhibited lower overall performance on the test set compared to the other evaluated approaches.

3.5.4 Soft Voting Ensemble

To further investigate the potential benefits of model combination, a soft voting ensemble was constructed using the optimized Random Forest and XGBoost models along with LightGBM. In soft voting, the predicted class probabilities of individual base classifiers are averaged, and the final prediction is determined by the highest aggregated probability.

Formally, the ensemble prediction is given by:

$$\hat{y} = \arg\max_{c} \left(\frac{1}{M} \sum_{m=1}^{M} P_m \left(y = c \mid x \right) \right)$$

where $P_m(y = c \mid x)$ denotes the predicted probability of class c produced by the m-th model, and M is the number of base learners.

The soft voting ensemble was evaluated under the same experimental protocol as the individual models. While the ensemble achieved competitive performance, it did not surpass the best-performing individual classifiers, highlighting that ensemble effectiveness depends on the diversity and complementarity of its constituent models.

3.6 Exploratory Data Analysis

In this section, the overall structure of the dataset was examined within the scope of exploratory data analysis (EDA). The distributions of key demographic and clinical variables and their relationships with heart disease were systematically analyzed. EDA enables a comprehensive understanding of the data structure prior to the modeling stage, facilitates the identification of potential patterns, and reveals the possible contributions of variables to model performance. Accordingly, fundamental features such as age, gender, chest pain type, and geographic region were analyzed in a structured manner.

Significant differences were observed in age distributions across datasets obtained from four different regions. The VA Long Beach dataset exhibited the highest mean age, approximately 59 years, whereas the Hungary dataset showed a considerably lower mean age of around 47 years. The Cleveland and Switzerland datasets were positioned between these two extremes. This demographic heterogeneity indicates that the combined use of these datasets may introduce regional biases, which could potentially influence model performance.

An examination of chest pain types by gender revealed notable differences in symptom presentation between male and female individuals. In particular, the proportions of typical angina and non-anginal pain differed significantly across genders. These clinical observations highlight the importance of considering gender in symptom evaluation and suggest that gender serves as a critical explanatory variable in the modeling process.

Regional analyses demonstrated that the distribution of chest pain types varied according to geographic origin. Asymptomatic chest pain was more prevalent in the Cleveland and VA Long Beach datasets, whereas atypical angina was more dominant in the Hungary dataset. In contrast, the Switzerland dataset exhibited a more balanced distribution of chest pain types. Furthermore, age-group-based analyses indicated that the asymptomatic symptom type was more frequently observed in older age groups. This finding suggests that symptom types may reflect age-related physiological changes.

When distributions by gender and region were evaluated, female individuals were most prominently represented in the Cleveland dataset (n = 97), followed by Hungary (n = 81), Switzerland (n = 10), and VA Long Beach (n = 6). Male individuals were predominantly represented in Hungary (n = 212), Cleveland (n = 207), and VA Long Beach (n = 194), with Switzerland including 113 male participants. These findings indicate the presence of pronounced regional and gender-based imbalances within the dataset.

Evaluations based on heart disease severity revealed that the number of male individuals without heart disease exceeded 250, whereas this number was approximately 150 among female individuals. As disease severity increased, the number of male patients became markedly higher than that of female patients. These results suggest that male individuals may carry a higher risk of heart disease.

An analysis of the thal variable by gender showed that the most common category among males was reversible defect (52.3%), followed by normal thal (33.8%) and fixed defect (13.8%). Among females, the dominant category was normal thal (65.8%), while reversible defect accounted for 31.6% and fixed defect for only 2.6%. Regional distribution analysis further indicated that the Cleveland dataset exhibited higher frequencies of both fixed defect (n = 18) and reversible defect (n = 117) compared to other regions.

Overall, both the independent distributions of key demographic and clinical variables—such as age, gender, chest pain type, and geographic region—and their relationships with heart disease were thoroughly examined. The findings provide a robust foundation for variable selection and interpretation of feature importance in the subsequent modeling stage. Based on these summarized results, the analysis proceeds to the visual exploration of selected demographic and clinical variables. Among demographic factors, age is widely recognized as one of the most critical determinants of heart disease risk. Accordingly, examining the age distribution is essential for assessing age-related trends in disease occurrence. Fig. 2 presents the overall distribution of the age variable within the dataset.

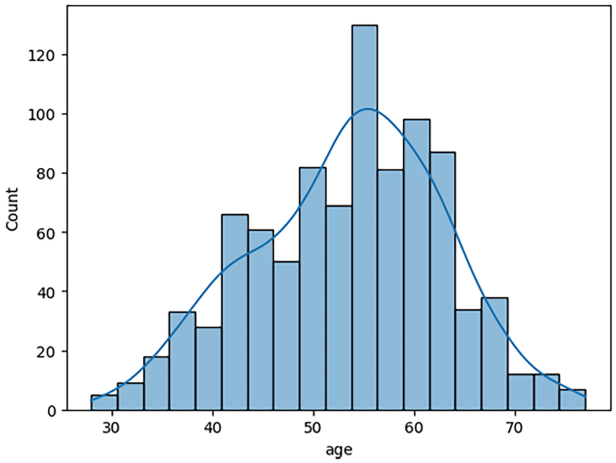


Figure 2: Distribution of the age variable.

The age histogram shown in Fig. 2 indicates that individuals are concentrated within the age range of 28–77 years, with a pronounced density observed particularly in the 50–60 age group. The mean age is approximately 53.5 years, which is consistent with the existing literature reporting that heart disease is more prevalent at older ages. This distribution supports the discriminative importance of the age variable in the modeling process. To further examine the effect of age on heart disease, the distribution of disease severity levels across age groups is analyzed in Fig. 3.

Fig. 3 demonstrates that as age increases, heart disease tends to appear at higher severity levels. Individuals without heart disease are predominantly concentrated at younger ages, whereas higher *num* values are more frequently observed among older individuals. This finding confirms that age is a strong risk factor for heart disease and highlights the necessity of explicitly considering the effect of age during the modeling process.

Resting blood pressure, which is one of the clinical indicators influencing heart disease risk, may exhibit different distributions depending on the presence of the disease. Accordingly, the relationship between resting blood pressure and heart disease status is examined in Fig. 4.

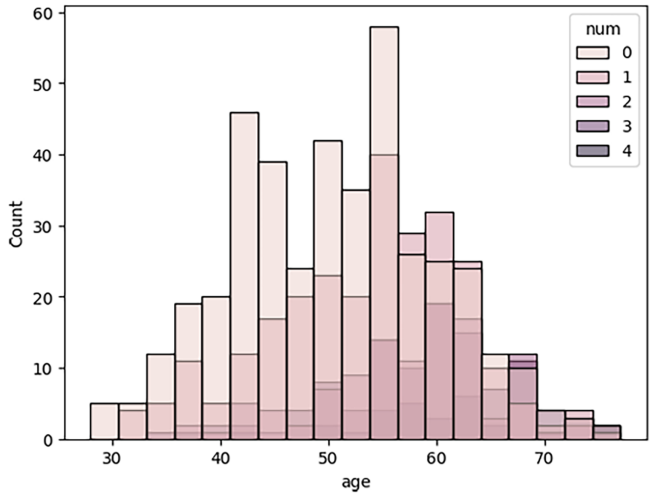


Figure 3: Distribution of heart disease severity levels (num) by age.

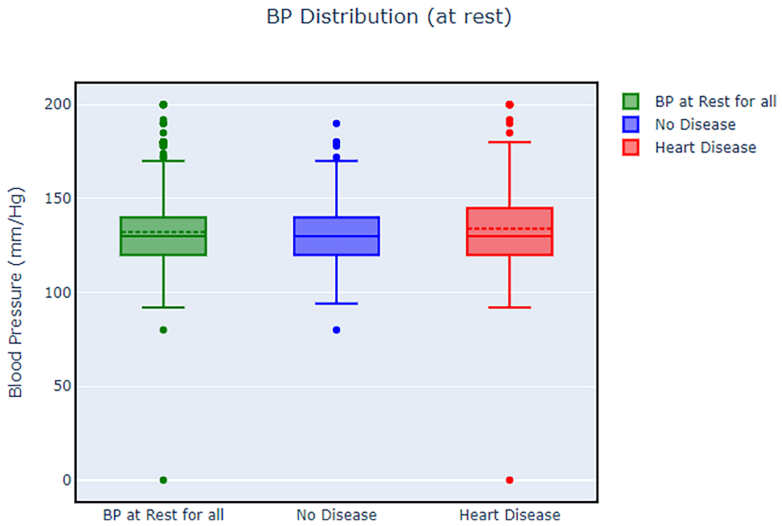


Figure 4: Distribution of resting blood pressure by heart disease status.

Fig. 4 presents a box plot illustrating the distribution of resting blood pressure according to heart disease status. The overall population is represented in green, individuals without heart disease in blue, and individuals with heart disease in red. The box plot clearly displays the minimum and maximum values, quartiles, and median. An examination of the figure reveals that individuals with heart disease exhibit a wider variance in resting blood pressure and higher median values compared to those without the disease. This observation suggests that elevated resting blood pressure may be significantly associated with the presence of heart disease.

Following the analysis of distributional differences in resting blood pressure, examining the basic statistical properties of this variable is important for a more comprehensive understanding of the data structure. In this context, the fundamental descriptive statistics of the resting blood pressure (*trestbps*) variable are presented in Fig. 5.

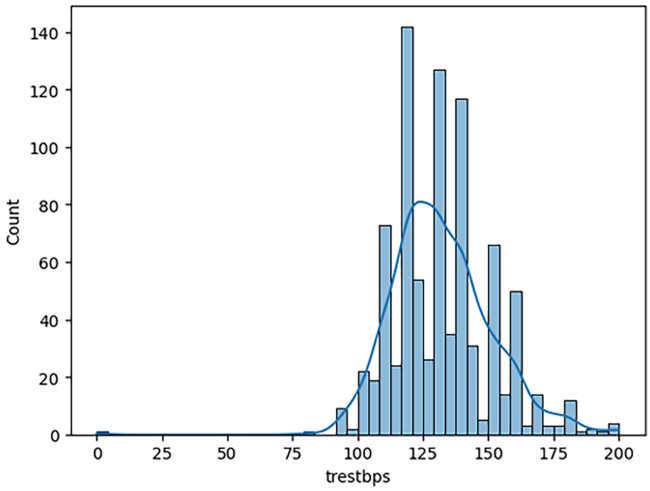


Figure 5: Descriptive statistics of the resting blood pressure (*trestbps*) variable.

Fig. 5 presents the basic descriptive statistics of the resting blood pressure (*trestbps*) variable. The data are largely concentrated within the range of 120–140 mmHg, with a mean value of 132.13 mmHg. The standard deviation of 19.07 mmHg indicates substantial variability among individuals. The minimum recorded value of 0 mmHg suggests the presence of a potential missing or erroneous measurement. Overall, the distribution is primarily concentrated around the upper-normal and mildly hypertensive clinical ranges.

Following the analysis of resting blood pressure, another important clinical indicator influencing heart disease—cholesterol levels—was examined. Cholesterol is a critical biochemical parameter in assessing cardiovascular risk, as it may increase the likelihood of heart disease through mechanisms such as atherosclerosis and coronary artery narrowing. The relationship between cholesterol levels and heart disease status is illustrated in Fig. 6.

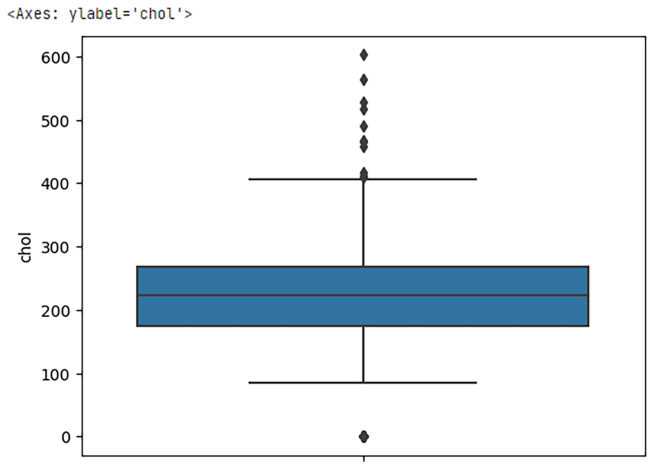


Figure 6: Relationship between cholesterol levels (*chol*) and heart disease status (*num*).

Fig. 6 illustrates the relationship between cholesterol levels (*chol*) and heart disease status (*num*). The cholesterol variable exhibits a wide distribution, with an average value of approximately 199.13 mg/dL. The relatively high standard deviation of 110.78 indicates substantial variability in cholesterol levels among individuals. The presence of a minimum value of 0 mg/dL suggests potential data quality issues, while

the maximum value of 603 mg/dL reveals that some individuals exhibit severe hypercholesterolemia. Although Fig. 6 does not indicate a clear linear relationship between cholesterol levels and heart disease status, the observed high variance and extreme values provide noteworthy insights for both clinical evaluation and the modeling process.

After examining the overall relationship between cholesterol levels and heart disease, a more detailed analysis of the distribution across different patient groups becomes essential. To this end, Fig. 7 presents a comparative visualization of cholesterol distributions among the overall population, individuals without heart disease, and individuals with heart disease.

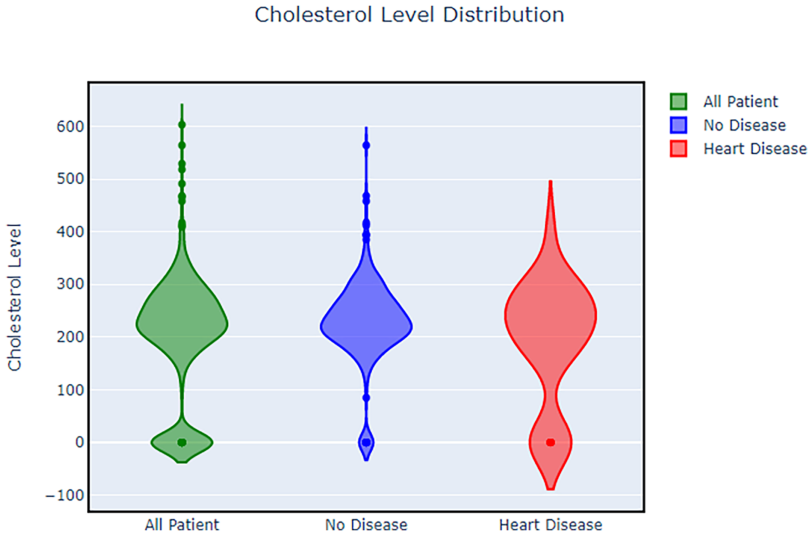


Figure 7: Distribution of patients' cholesterol levels.

The violin plot in Fig. 7 illustrates the density distribution of cholesterol values among all individuals, as well as separately for those without and with heart disease. The central markers in the plots represent the measures of central tendency for each group. The analysis shows that cholesterol values are distributed across a wide range in all groups (approximately 0–600 mg/dL), with density particularly concentrated in the 200–300 mg/dL range. In individuals with heart disease, the distribution shifts toward higher values, and the measures of central tendency are higher compared to other groups. These findings indicate that cholesterol is associated with heart disease but cannot be used as a definitive marker on its own.

Chest pain type (cp), gender, and thalassemia (thal) test results are critical variables in the clinical assessment of heart disease. When analyzed together, these variables allow for a multidimensional evaluation of cardiovascular risk. To this end, Figs. 8–10 present the relevant visual analyses. The chest pain types in the dataset are divided into four categories: asymptomatic, non-anginal, atypical angina, and typical angina. Fig. 8 shows the distribution of these types according to gender.

In Fig. 8, the distribution of chest pain types by gender is examined. It can be observed that asymptomatic and non-anginal types are more frequent in males, whereas the atypical angina type is more evenly distributed among females.

Following these gender-based differences, the relationship between chest pain types and thalassemia (thal) test results is visually presented in Fig. 9.

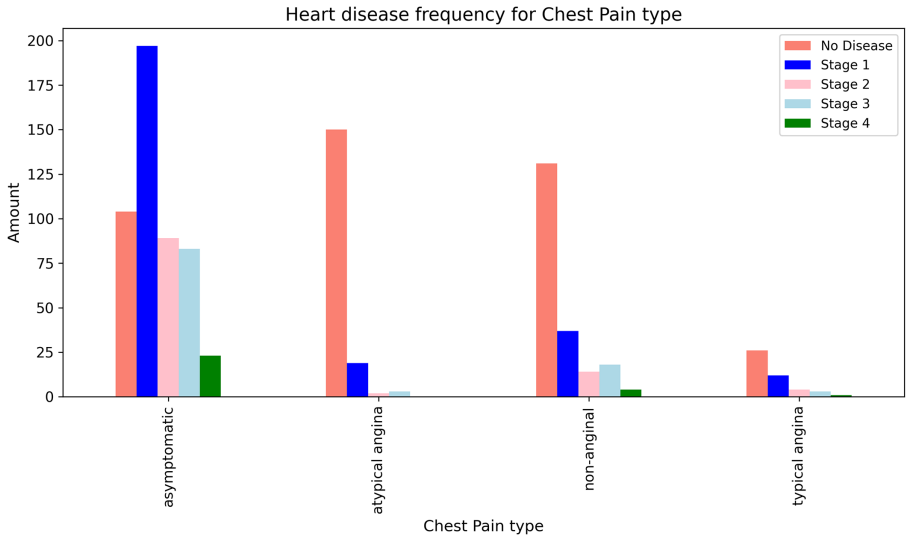


Figure 8: Distribution of chest pain types by gender.

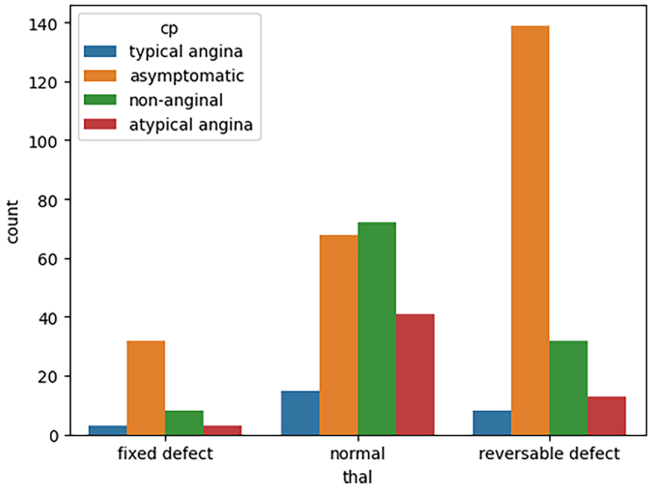


Figure 9: Distribution of chest pain types by thalassemia test results.

Fig. 9 illustrates the relationship between chest pain types and thalassemia (thal) test results. A significant portion of asymptomatic individuals exhibit abnormal thal results, such as reversible or fixed defects, whereas the non-anginal group predominantly shows normal test results.

Additionally, the distribution of heart disease severity according to thalassemia test results is also presented in Fig. 9, clearly highlighting the relationship between thal findings and disease severity.

Fig. 10 shows the distribution of heart disease severity according to thalassemia (thal) test results. The findings reveal strong and consistent relationships among chest pain types, gender, thal results, and heart disease severity. Notably, the presence of advanced heart disease in asymptomatic individuals indicates that clinical assessments based solely on symptoms may be insufficient.

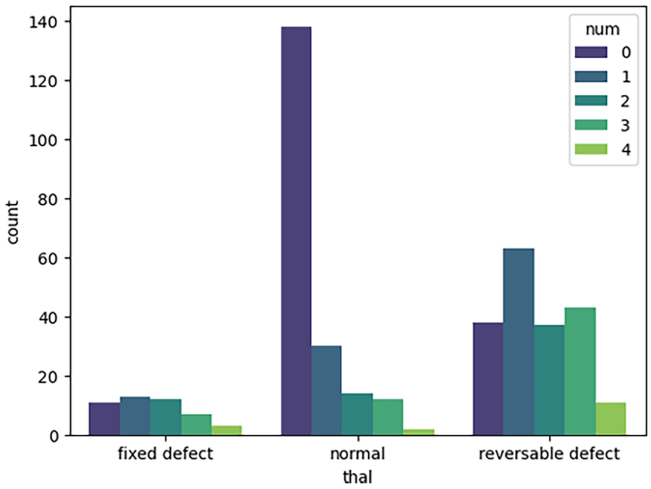


Figure 10: Distribution of heart disease severity by thalassemia test results.

The descriptive insights obtained from the exploratory data analysis provide the foundation for the subsequent modeling and experimental evaluation.

4 Experimental Results

This section presents the experimental evaluation of the proposed approach using the selected heart disease dataset. The experimental analysis is structured in two stages. First, an exploratory data analysis (EDA) is conducted to examine the underlying characteristics of the dataset, identify clinically meaningful patterns, and provide preliminary insights into the relationships between key variables and the target outcome. These observations help to motivate the subsequent modeling choices. In the second stage, the classification performance of the selected machine learning models is evaluated and compared using standard performance metrics.

4.1 Exploratory Data Analysis Results

Although the dataset is provided as a unified Kaggle heart disease dataset, it aggregates patient records originally collected from multiple clinical sources, including Cleveland, Hungary, Switzerland, and the VA Long Beach database. In this study, these records are analyzed within a single integrated dataset, and the exploratory data analysis is conducted to examine distributional patterns across the original subgroups without treating them as independent datasets.

This subsection summarizes the key patterns observed from the exploratory data analysis (EDA) conducted on the heart disease dataset. The EDA provides descriptive insights into the distributional characteristics of the variables and their relationships with the target outcome, thereby motivating the subsequent modeling stage.

Figs. 8–10 jointly highlight clinically meaningful evidence regarding chest pain type (cp), gender, and thalassemia test results (thal). As shown in Fig. 8, the distribution of chest pain categories differs by gender, where asymptomatic and non-anginal types appear more frequently among males, while atypical angina is more evenly represented among females. Extending this analysis, Fig. 9 illustrates the association between chest pain categories and thal results, suggesting that a substantial portion of asymptomatic individuals exhibit abnormal thal findings (e.g., reversible or fixed defects), whereas the non-anginal group is more frequently associated with normal thal outcomes. Finally, Fig. 10 presents the distribution of heart disease

severity across thal categories and indicates a consistent relationship between abnormal thal results and more severe disease levels. Notably, the presence of advanced disease among asymptomatic individuals suggests that symptom-based clinical assessment alone may be insufficient, underlining the need for data-driven decision support.

Overall, the descriptive evidence from the EDA forms the empirical basis for the experimental evaluation presented in the next subsection and supports the rationale for employing machine learning models for reliable risk classification.

4.2 Model Performance Results

This subsection presents the classification performance of the final models evaluated on the test set. The models were assessed using accuracy, precision, recall (sensitivity), specificity, F1-score, and ROC-AUC metrics. The comparative performance results of the evaluated models are summarized in Table 4, enabling a quantitative comparison of the discriminative ability and overall classification performance of the final model configurations.

Table 4: Comparison of classification performance of machine learning models.

Model	Accuracy	Precision	Recall (Sensitivity)	Specificity	F1	ROC-AUC
Tuned XGBoost + FeatureSel	0.8478	0.8700	0.8529	0.8415	0.8614	0.9316
Tuned Random Forest + FeatureSel	0.8478	0.8491	0.8824	0.8049	0.8654	0.9313
Soft Voting Ensemble	0.8424	0.8544	0.8627	0.8171	0.8585	0.9264
LightGBM + FeatureSel	0.8152	0.8269	0.8431	0.7805	0.8350	0.8911

According to the results reported in Table 4, the tuned XGBoost and tuned Random Forest models exhibit the highest overall performance. Both models achieve an accuracy of 0.8478, while their ROC-AUC values are 0.9316 and 0.9313, respectively. The tuned Random Forest model provides the highest sensitivity (0.8824), whereas the tuned XGBoost model yields higher specificity (0.8415) and precision (0.8700). The soft voting ensemble model demonstrates competitive performance with an accuracy of 0.8424 and a ROC-AUC value of 0.9264, but does not surpass the best-performing individual models. In contrast, the LightGBM model remains at a lower performance level compared to the other evaluated approaches.

Although some studies in the literature report higher accuracy values for heart disease prediction, accuracy alone may provide a limited view of clinical performance, particularly in datasets exhibiting class imbalance. In such scenarios, metrics such as sensitivity, specificity, F1-score, and ROC–AUC offer more informative insights into a model's diagnostic reliability and error trade-offs.

In the present study, the proposed models achieved moderate accuracy values while demonstrating strong ROC–AUC performance and balanced sensitivity–specificity characteristics. This indicates that the models are capable of effectively discriminating between positive and negative cases across varying decision thresholds, which is particularly important in clinical screening applications where false negatives carry significant risk.

Compared to recent studies employing similar tabular clinical datasets, the proposed approach emphasizes methodological consistency and reproducibility rather than maximizing a single performance metric. All models were evaluated under identical preprocessing pipelines, validation strategies, and hyperparameter optimization procedures. This controlled experimental setting enables a fair comparison of commonly used classifiers and reduces the risk of optimistic performance estimates arising from selective reporting.

Overall, the results suggest that while higher accuracy values can be achieved under specific experimental configurations, balanced multi-metric performance and methodological transparency are more appropriate criteria for evaluating machine learning models intended for clinical decision support.

Fig. 11 presents a heatmap summarizing the main performance metrics of the final models on the test set, including accuracy, precision, sensitivity, specificity, F1-score, and ROC-AUC. In the heatmap, each row corresponds to a model and each column represents an evaluation metric, while the color intensity reflects the relative magnitude of the metric values, enabling a clear visual comparison of performance differences among the models.

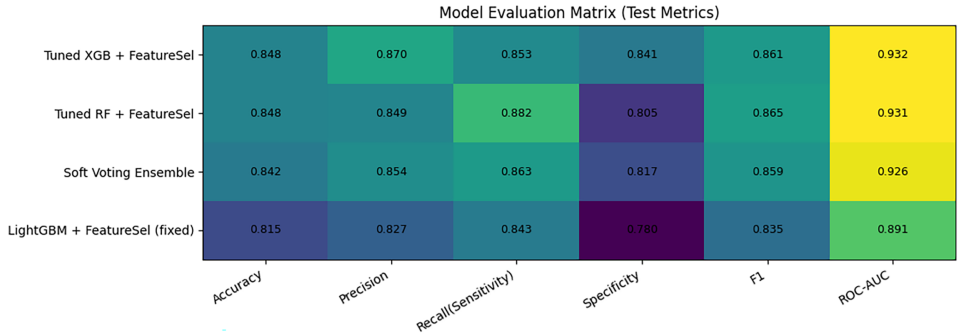


Figure 11: Heatmap of performance metrics for the evaluated models on the test set.

As observed in the heatmap presented in Fig. 11, the tuned XGBoost and tuned Random Forest models exhibit higher performance across all evaluation metrics. While both models achieve similar accuracy levels, the XGBoost model stands out in terms of precision and specificity, whereas the Random Forest model attains higher sensitivity. The soft voting ensemble model delivers performance close to these two models across most metrics; however, it falls behind the individual models, particularly with respect to accuracy and ROC-AUC. In contrast, the LightGBM model demonstrates consistently lower performance across all evaluated metrics.

To assess the discriminative ability of the final models across different decision thresholds, ROC curves obtained on the test set are presented in Fig. 12.

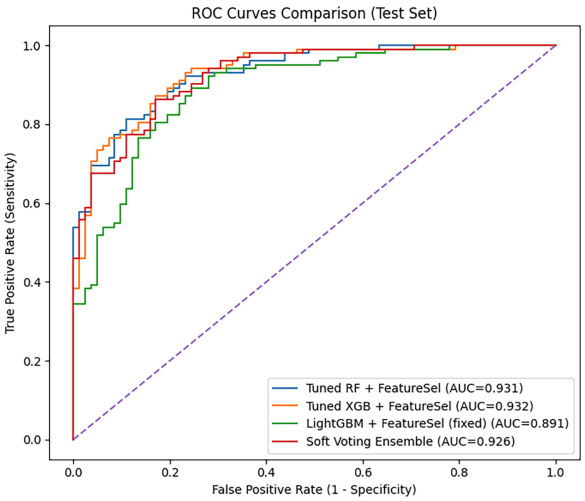


Figure 12: ROC curves of classification models.

As observed from the ROC curves presented in Fig. 12, the tuned XGBoost and tuned Random Forest models demonstrate high discriminative performance on the test set. This behavior is further quantified through ROC-AUC values, as discussed below. Although the soft voting ensemble model follows a similar trend, it achieves a comparatively lower ROC-AUC value. In contrast, the ROC curve of the LightGBM model points to a lower level of discriminative ability compared to the other models.

To further quantify the discriminative performance observed in the ROC curves, the ROC-AUC values of the evaluated models on the test set are summarized in Fig. 13.

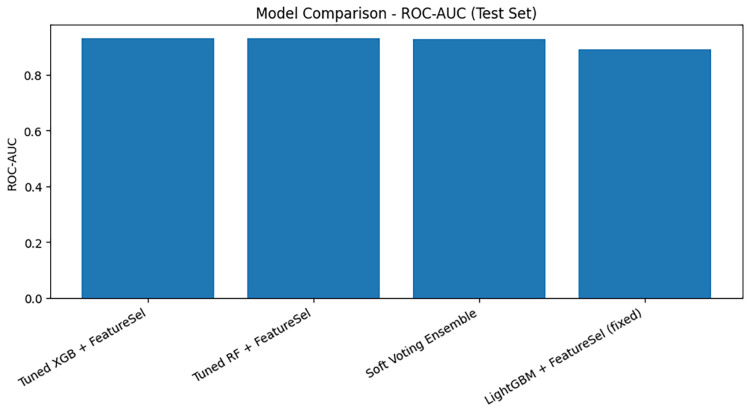


Figure 13: ROC-AUC values of the evaluated models on the test set.

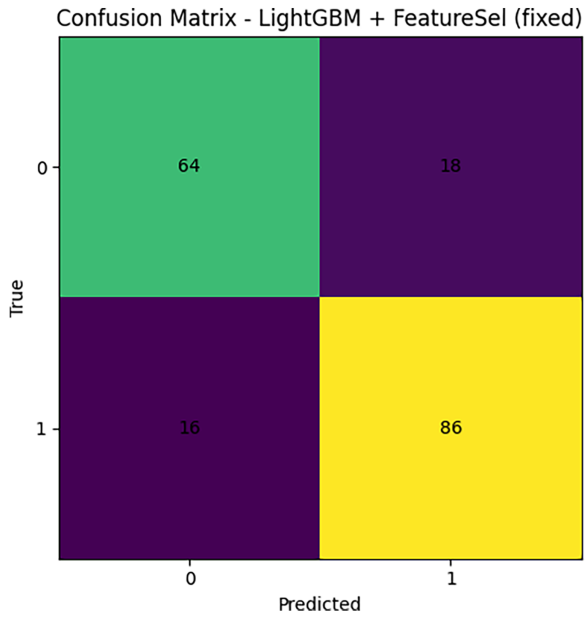
As shown in Fig. 13, the tuned XGBoost and Random Forest models achieved the highest ROC-AUC values (approximately 0.93), outperforming the ensemble model and LightGBM, thereby confirming their superior discriminative capability on the test data.

Beyond overall performance metrics, a more detailed analysis of classification behavior requires examining the distribution of errors; accordingly, the confusion matrices obtained on the test set are presented in Fig. 14.

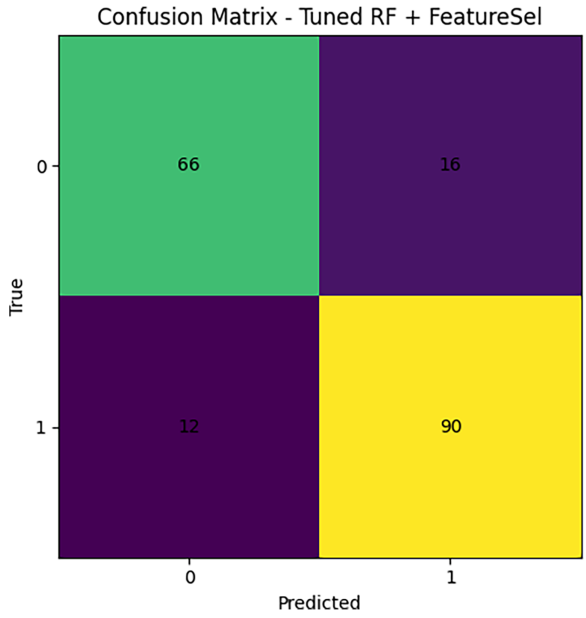
Fig. 14 presents the confusion matrices of the evaluated models on the test set, highlighting differences in false positive and false negative rates, with the tuned XGBoost and Random Forest models exhibiting a more balanced error distribution compared to the ensemble and LightGBM models.

To further interpret the error characteristics identified in the confusion matrices, the sensitivity and specificity values of the evaluated models are compared in Fig. 15.

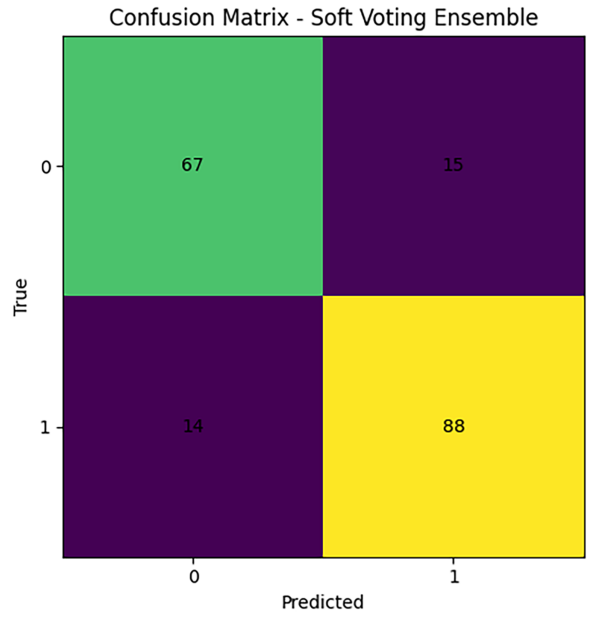
Fig. 15 illustrates the trade-off between sensitivity and specificity across the models, showing that the tuned Random Forest model achieves higher sensitivity, whereas the tuned XGBoost model provides a more balanced specificity, highlighting their suitability for different clinical decision-making scenarios.



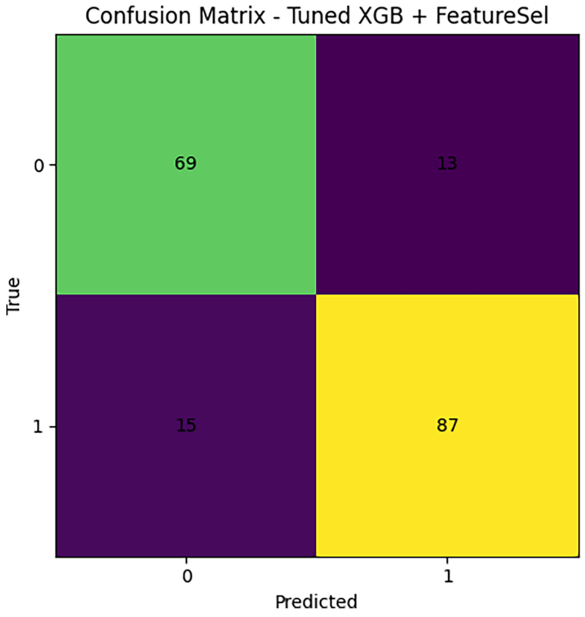
(a) Confusion matrix of the tuned XGBoost model



(b) Confusion matrix of the tuned random forest model



(c) Confusion matrix of the soft voting ensemble model



(d) Confusion matrix of the LightGBM model

Figure 14: Confusion matrices of the evaluated models on the test set.

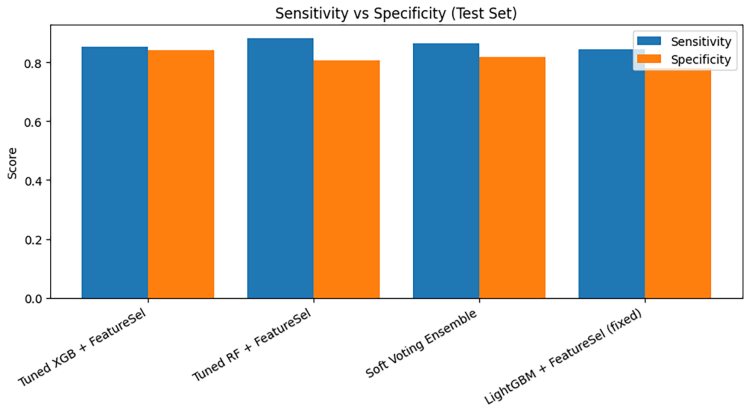


Figure 15: Sensitivity and specificity comparison of the evaluated models on the test set.

5 Discussion

The results obtained in this study demonstrate that tree-based machine learning models can achieve strong and clinically meaningful performance for heart disease classification when evaluated under a strict and controlled experimental setting. Across the tested configurations, the best-performing models consistently achieved ROC-AUC values close to 0.93, indicating a high level of discriminative capability between positive and negative cases. Importantly, this performance was accompanied by balanced accuracy and F1-score values, suggesting that the models do not rely on a single dominant metric but rather maintain stable behavior across multiple evaluation criteria.

A notable finding is the difference in operating characteristics between the top-performing models. The Random Forest model achieved the highest sensitivity, corresponding to a lower number of false-negative predictions. From a clinical perspective, this behavior is particularly desirable in screening-oriented scenarios, where failing to identify a patient with potential heart disease may lead to delayed diagnosis and increased risk. In contrast, the XGBoost model exhibited higher specificity, resulting in fewer false-positive predictions. Such a profile may be preferable in contexts where unnecessary follow-up examinations, patient anxiety, or additional healthcare costs need to be minimized. These complementary behaviors highlight that no single model is universally optimal; instead, model selection should be guided by the intended clinical use case and the relative cost of false-negative vs. false-positive decisions.

Although ensemble learning is often expected to enhance predictive performance by combining multiple models, the soft-voting ensemble evaluated in this study did not outperform the strongest individual classifiers. While the ensemble achieved competitive results, its overall discriminative power and accuracy remained slightly below those of the best-performing single models. This outcome suggests that when individual learners already exhibit strong and correlated decision patterns, simple aggregation strategies such as soft voting may offer limited additional benefit. In such cases, ensemble performance can be constrained by the weakest component model, potentially diluting the discriminative signal of the strongest learner. These findings emphasize that ensemble construction should be approached carefully and that more complex ensemble strategies do not automatically guarantee superior performance.

Beyond raw predictive accuracy, the observed balance between sensitivity and specificity constitutes one of the most important contributions of this study. The results demonstrate that high discrimination can be achieved without excessively favoring one type of error over another. This balance is particularly relevant for clinical decision-support systems, where both missed diagnoses and unnecessary interventions carry significant consequences. The reported confusion matrix patterns indicate that the proposed models manage

this trade-off in a controlled and interpretable manner, supporting their potential applicability in real-world clinical settings.

Despite the encouraging results, several limitations should be acknowledged. First, model evaluation was performed using a fixed train–test split, which supports reproducibility but may not fully capture performance variability under different data partitions. Second, while this study focuses on classification performance, it does not explicitly address probability calibration or decision-threshold optimization, both of which are critical for deployment in clinical workflows. Future research could explore calibrated probability estimates and threshold selection strategies aligned with specific clinical objectives, such as prioritizing sensitivity for early screening or specificity for confirmatory assessment.

In summary, the findings of this study indicate that optimized tree-based models provide strong, stable, and clinically interpretable performance for heart disease classification. The results underscore the importance of evaluating models through multiple complementary metrics and interpreting performance in light of clinical priorities rather than relying on a single summary score. Together, these insights contribute to a more nuanced understanding of model behavior and support the use of machine learning as a reliable tool for heart disease risk assessment.

6 Conclusion

In this study, we presented a systematic and reproducible comparative evaluation of machine learning models for heart disease classification using a publicly available tabular clinical dataset. A structured preprocessing pipeline was employed, including model-based imputation for missing values, feature encoding and scaling, and a leakage-free train–test protocol. The final experimental results show that the optimized tree-based models achieve strong and clinically meaningful performance on the test set, with the best configurations reaching an accuracy of 0.85 and ROC-AUC values close to 0.93. In particular, tuned XGBoost provides higher precision and specificity, while tuned Random Forest yields higher sensitivity, indicating that model preference should be guided by clinical priorities (e.g., screening-oriented vs. confirmatory use). Although the soft-voting ensemble delivered competitive results, it did not surpass the strongest individual models, suggesting limited gains when base learners exhibit correlated decision behavior under the same feature space.

Overall, the findings support the use of optimized, interpretable tree-based classifiers as reliable baselines for heart disease risk assessment from tabular clinical data. Future work will focus on evaluating performance variability under repeated resampling or external validation, investigating probability calibration and threshold optimization for deployment-oriented settings, and extending the framework with clinically grounded interpretability analyses.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Kadriye Simsek Alan: Conceptualization, Methodology, Formal analysis, Writing—original draft preparation, Writing—review and editing, Software, Supervision, Project administration. Busra Senel Kahyaoglu: Data curation, Validation, Visualization. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in the Heart Disease Dataset published on the Kaggle platform by Karimsony at the following URL: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Nomenclature

AI	Artificial Intelligence
SVM	Support Vector Machines
KNN	K-Nearest Neighbors
LR	Logistic Regression
RF	Random Forest
XGBoost	Extreme Gradient Boosting algorithm
LightGBM	Light Gradient Boosting Machine algorithm
AdaBoost	Adaptive Boosting algorithm
Bagging	Bootstrap Aggregating algorithm
DT	Decision Trees
MAE	Mean Absolute Error (a regression evaluation metric)
RMSE	Root Mean Squared Error (a regression evaluation metric)
R^2	Coefficient of Determination (explained variance score)
EDA	Exploratory Data Analysis
UCI	University of California, Irvine Machine Learning Repository
Feature Importance	Ranking of input variables by predictive impact

References

1. World Health Organization. Cardiovascular diseases (CVDs): key facts [Internet]. Geneva, Switzerland: WHO Fact Sheet; 2023 [cited 2026 Jan 1]. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
2. Zamani AS, Hashim AHA, Ali Shatat AS, Akhtar MM, Rizwanullah M, Mohamed SSI. Implementation of machine learning techniques with big data and IoT to create effective prediction models for health informatics. Biomed Signal Process Control. 2024;94(3):106247. doi:10.1016/j.bspc.2024.106247.
3. Almutairi M, Dardouri S. Intelligent hybrid modeling for heart disease prediction. Information. 2025;16(10):869. doi:10.3390/info16100869.
4. Ogunpola A, Saeed F, Basurra S, Albarrak AM, Qasem SN. Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics. 2024;14(2):144. doi:10.3390/diagnostics14020144.
5. Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective heart disease prediction using machine learning techniques. Algorithms. 2023;16(2):88. doi:10.3390/a16020088.
6. Banerjee T, Paçal İ. A systematic review of machine learning in heart disease prediction. Turk J Biol. 2025;49(5):600–34. doi:10.55730/1300-0152.2766.
7. Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access. 2019;7:81542–54. doi:10.1109/access.2019.2923707.
8. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1–30.
9. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6. doi:10.1186/s12864-019-6413-7.
10. Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. NeuroImage. 2018;180(11):68–77. doi:10.1016/j.neuroimage.2017.06.061.
11. Naseri Jahfari A, Tax D, Reinders M, van der Bilt I. Machine learning for cardiovascular outcomes from wearable data: systematic review from a technology readiness level point of view. JMIR Med Inform. 2022;10(1):e29434. doi:10.2196/29434.
12. Benali A. Continuous cardiovascular health monitoring with IoT-enabled smart wearable devices: designs IoT-based wearable devices for continuous monitoring of cardiovascular parameters, facilitating early detection of cardiac abnormalities and improving heart health management. J Deep Learn Genom Data Anal. 2024;4(1):84–93.

13. Iqbal SMA, Leavitt MA, Mahgoub I, Asghar W. Advances in cardiovascular wearable devices. Biosensors. 2024;14(11):525. doi:10.3390/bios14110525.
14. Lin J, Fu R, Zhong X, Yu P, Tan G, Li W, et al. Wearable sensors and devices for real-time cardiovascular disease monitoring. Cell Rep Phys Sci. 2021;2(8):100541. doi:10.1016/j.xcrp.2021.100541.
15. Rahim A, Rasheed Y, Azam F, Anwar MW, Rahim MA, Muzaffar AW. An integrated machine learning framework for effective prediction of cardiovascular diseases. IEEE Access. 2021;9:106575–88. doi:10.1109/access.2021.3098688.
16. Amzad Hossen MD, Tazin T, Khan S, Alam E, Sojib HA, Monirujjaman Khan M, et al. Supervised machine learning-based cardiovascular disease analysis and prediction. Math Probl Eng. 2021;2021:1792201. doi:10.1155/2021/1792201.
17. Ambrish G, Ganesh B, Ganesh A, Srinivas C, Dhanraj, Mensinkal K. Logistic regression technique for prediction of cardiovascular disease. Glob Transitions Proc. 2022;3(1):127–30. doi:10.1016/j.gltp.2022.04.008.
18. Azmi J, Arif M, Nafis MT, Alam MA, Tanweer S, Wang G. A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. Med Eng Phys. 2022;105(1):103825. doi:10.1016/j.medengphy.2022.103825.
19. Khan Y, Qamar U, Yousaf N, Khan A. Machine learning techniques for heart disease datasets: a survey. In: Proceedings of the 2019 11th International Conference on Machine Learning and Computing; 2019 Feb 22–24; Zhuhai, China. p. 27–35.
20. Berdinanth M, Syed S, Velusamy S, Suseelan AD, Sivanaiah R. Analysis of traditional machine learning approaches on heart attacks prediction. Rev Română De Informatică Şi Autom. 2024;34(1):23–30. doi:10.33436/v34i1y202403.
21. Kumar A, Khan AA, Singh J. Enhancing the diagnosis of cardiovascular disease: a comparative examination of support vector machine and artificial neural network models. WSEAS Trans Comput. 2024;23(31):318–27. doi:10.37394/23205.2024.23.31.
22. Arooj S, Rehman SU, Imran A, Almuhaimeed A, Alzahrani AK, Alzahrani A. A deep convolutional neural network for the early detection of heart disease. Biomedicines. 2022;10(11):2796. doi:10.3390/biomedicines10112796.
23. Allheeib N, Kanwal S, Alamri S. An intelligent heart disease prediction framework using machine learning and deep learning techniques. Int J Data Warehous Min. 2023;19(1):1–24. doi:10.4018/ijdwm.333862.
24. Jamila S, Roy AM. An efficient PCG-based valvular heart disease detection framework using Vision Transformer. Comput Biol Med. 2023;158(25):106734. doi:10.1016/j.compbiomed.2023.106734.
25. Mansoor CMM, Chettri SK, Naleer HMM. Development of an efficient novel method for coronary artery disease prediction using machine learning and deep learning techniques. Technol Health Care. 2024;32(6):4545–69. doi:10.3233/thc-240740.
26. Qadri AM, Raza A, Munir K, Almutairi MS. Effective feature engineering technique for heart disease prediction with machine learning. IEEE Access. 2023;11:56214–24. doi:10.1109/ACCESS.2023.3281484.
27. Cuevas-Chávez A, Hernández Y, Ortiz-Hernandez J, Sánchez-Jiménez E, Ochoa-Ruiz G, Pérez J, et al. A systematic review of machine learning and IoT applied to the prediction and monitoring of cardiovascular diseases. Healthcare. 2023;11(16):2240. doi:10.3390/healthcare11162240.
28. Baghdadi NA, Farghaly Abdelaliem SM, Malki A, Gad I, Ewis A, Atlam E. Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. J Big Data. 2023;10(1):144. doi:10.1186/s40537-023-00817-1.
29. El-Sofany H, Bouallegue B, El-Latif YMA. A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Sci Rep. 2024;14(1):23277. doi:10.1038/s41598-024-74656-2.
30. El-Sofany HF. Predicting heart diseases using machine learning and different data classification techniques. IEEE Access. 2024;12(4):106146–60. doi:10.1109/ACCESS.2024.3437181.
31. Al-Alshaikh HA, Prabu P, Poonia RC, Saudagar AKJ, Yadav M, AlSagri HS, et al. Comprehensive evaluation and performance analysis of machine learning in heart disease prediction. Sci Rep. 2024;14(1):7819. doi:10.1038/s41598-024-58489-7.
32. Elhadjamor EA, Harbaoui H. A comparison analysis of heart disease prediction using supervised machine learning techniques. In: Proceedings of the 2024 IEEE Symposium on Computers and Communications (ISCC); 2024 Jun 26–29; Paris, France. p. 1–6. doi:10.1109/iscc61673.2024.10733656.

33. Mandal AK, Dehuri S, Sarma PKD. Analysis of machine learning approaches for predictive modeling in heart disease detection systems. Biomed Signal Process Control. 2025;106(4):107723. doi:10.1016/j.bspc.2025.107723.
34. Qian B. A comparative study of machine learning models for cardiovascular disease prediction. In: Proceedings of the 1st International Conference on Modern Logistics and Supply Chain Management; 2024 Sep 27–29; Singapore. Setúbal, Portugal: SCITEPRESS—Science and Technology Publications; 2024. p. 377–81. doi:10.5220/0013332300004558.
35. Temirbayeva AB, Altybay A. Machine learning methods for predicting cardiovascular diseases: a comparative analysis. Vestn Ross Univ Družby Nar Ser Inž Issled. 2025;26(2):168–80. doi:10.22363/2312-8143-2025-26-2-168-180.
36. Hussain A, Aslam A. Cardiovascular disease prediction using risk factors: a comparative performance analysis of machine learning models. J Artif Intell. 2024;6(1):129–52. doi:10.32604/jai.2024.050277.
37. Karimsony R. Heart disease dataset [Internet]. 2022 [cited 2026 Jan 1]. Available from: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data.