International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

HEART DISEASE PREDICTION USING MACHINE LEARNING ON A REAL-TIME CLINICAL DATASET WITH MULTI-DIAGNOSTIC FEATURES

Sunanda Budhal[1*], Sheetalrani Kawale[2] , Nitin Agarwal[3]

[1] Department of Computer Science, Government First Grade College, Bagalkot-587103, Karnataka, India

[2] Department of Computer Science, Karnatak State Women's University, Vijayapura-596101, Karnataka, India

[3] Consulting Physician and Cardiologist, Ayush Multi Speciality Hospital and Research Centre(AMSHRC) Pvt, Ltd, Vijayapur,Karnataka, India

sunanda.kiran2003@gmail.com,

sheetalrani@kswu.ac.in

Abstract

Cardiovascular disease remains a significant health concern, being the primary cause of death, which explains why its early and precise prediction is necessary. Machine learning (ML) can be used to discover concealed clinical data patterns and aid in early diagnoses. This paper presents a baseline ML model constructed using an actual dataset of 1100 anonymized patient records obtained from **Ayush Multi Speciality Hospital and Research Centre**(**AMSHRC**) **Pvt, Ltd, Vijayapur,Karnatak, India.** A coordinated pre-processing process was employed, which included missing value treatment, outlier correction, categorical encoding, and scale normalization of numerical values. Exploratory analysis, correlation analysis, and model-based feature importance were used to examine clinically relevant predictors. The six supervised ML algorithms used were Logistic Regression, Support Vector Machine , Decision Tree, K-Nearest Neighbors, Random Forest, and XGBoost, which were trained using an 80:20 stratified split and assessed using Accuracy, Precision, Recall, F1-score, ROC-AUC, MCC, and confusion matrix. Stratified 10-fold cross-validation was employed to provide a good estimation of performance. All the models were extremely predictive, Random Forest and XGBoost (0.93%) and ROC-AUC (0.98) achieved the best accuracy. These aspects, including CAG, ECG,ECHO, TMT, and the nature of chest pain, turned out to be the strongest predictors. This research provides a solid foundation for the estimation of heart disease using clinically rich data through the implementation of ML. In further research, the size of the dataset will be expanded, features will be extracted either automatically or manually, hyperparameter optimization, nested cross-validation, and explainability algorithms like SHAP will be employed in order to build a more robust and reliable model.

Keywords—Heart Disease Prediction, Machine Learning, Clinical Dataset, Cardiovascular Diagnostics, Supervised Learning, Medical Data Analytics

Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

1.INTRODUCTION:

Cardiovascular diseases (CVDs) are the number one cause of death in the world, with the organization reporting just under 18 million cases per year [1]. The increase in heart disease is due to changes in lifestyle, dietary habits, metabolic disorders and people are aging at an increasing rate [2]. Timely intervention is very important in avoiding complications, as it can be achieved through early prediction. Given the recent breakthroughs in the domain of computational technologies, machine learning (ML) has become a potent means of revealing hidden trends in clinical data and helping clinicians to make risk decisions and stratify risks.

Several studies have investigated the predictions of cardiac based on a large amount of publicly available date, including the UCI Cleveland dataset proposed by Janosi et al. [6] , the statlog(heart) dataset from UCI machine Learning repository [21]and different Kaggle repositories [4]. Support Vector Machine (SVM), Logistic Regression, the Random Forest (RF), and the Decision Trees (DT) are classical models of ML that have been extensively tested in the previous research[28][29] . Ram et al. [7], Cenitta et al. [8] and Gnanavelu et al. [10] studies reported competitive to high predictive performance on these datasets. Further progress has been made on hybrid feature selection, ensemble learning, and optimized neural architectures by Chandrasekhar and Peddakrishna [12], Biswas et al. [13], Bouqentar, M. A et.al. [17],and along with comprehensive survey analysis presented by Shinde,P et.al[23].

In addition to traditional datasets, a number of studies have studied more inclusive clinical sets. A study by A. Mahajan *et al*. [5] on UCI and other patients of a hospital using hybrid feature selection and ensembled stack learning model yielded an accuracy of between 97%-99%, whereas M. Alshraideh *et al*. [30] worked on Jordan University Hospital 486 clinical samples and model gave accuracy of 94.3% accuracy. Such works are evidence of movement towards practical use, but indicate a continuing issue: most ML-based studies still use small datasets that do not have the most important modern diagnostic features, such as ECG, ECHO, TMT, and CAG, and which hardly ever use real-time hospital data.

Although it has provided promising developments, there are a few limitations in the current research. Most of them use benchmark data that have limited diagnostic features and fixed sample sizes, such as 303 records [6]. Consequently, the models that are trained using curated datasets might not be generalizable to a real-world clinical setting. Moreover, not many studies utilize extensive cardiological variables such as ECG, ECHO, TMT, and coronary angiography results, which are required to make clinically significant predictions [8], [24], [26].

In order to fill these gaps, the current paper presents real-time clinical data of 1,100 patient records at (AMSHRC)Hospital, Vijayapur. The data consists of the demographic, lifestyle, physiological, and diagnostic values, including pulse rate, SpO2, blood sugar(diabetics), cholesterol, hypertension, habits, ECG, ECHO, TMT, and CAG results. The dataset will go through a strict preprocessing phase as described in the methodology, which will include imputation of missing values, categorical encoding, and re-scaling of the numbers. The six classical machine learning algorithms, namely Logistic Regression, SVM, Decision Tree,

Received: August 09 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

KNN, Random Forest, and XGBoost, are tested on accuracy, precision, recall,, F1-score, ROC-AUC, and confusion matrix metrics to create a solid baseline performance.

With the help of a real-world clinical dataset, this research provides a filler to benchmark-based studies and enhances applicability to the work by presenting an interpretable and evidence-based machine learning model to predict heart diseases. The results can be added to the expanding field of AI-assisted healthcare analytics, as they enhance predictive precision, reliability, and clinical impact.

The principal findings of this paper are as follows:

1. Presentation of a new real-time clinical dataset with 1,100 patient records from **AMSHRC Pvt, Ltd, Vijayapur,Karnatak, India.**

2. Comparison of six traditional machine learning algorithms based on a single common preprocessing and modeling pipeline.

3. Exploratory analysis and model-based investigation of clinically important cardiac markers.

Establishment of baseline machine learning benchmark outcomes to facilitate future lab work with 2,000 or more patient records and sophisticated optimization methods. Subsequently, the rest of the paper is structured as follows: Section II reviews the related literature, Section III describes the dataset, Section IV presents the methodology (data collection, preprocessing, and modeling), Section V reports the experimental results and performance assessment, and finally, Section VI provides a conclusion with clinical implications and future research directions.

2.RELATED WORK

The cardiovascular diseases (CVDs) have been declared as the most extensive cause of death across the globe, according to the World Health Organization (WHO) [1]. Early and accurate diagnosis therefore is the best in order to attain mortality and improved patient outcomes. The past years have seen a wide range of studies concerning the use of machine learning (ML) techniques as an automated approach to heart disease prediction; however, the problems of the scale of the problem, the depth of diagnosis, and the clinical application still exist. Most of the available literature makes use of the data of 303 patients of the traditional clinical dataset offered by Janosi et al. [6], such as age, sex, chest pain type, blood pressure, cholesterol levels, electrocardiographic results, and exercise-induced angina. Although it has been used extensively in benchmarking, the small sample size and inadequately advanced diagnostic data intrude into the extrapolation of the models which are constructed based on this data. The UCI data has been intensively experimented using classical machine learning algorithms of the Logistic Regression (LR), the Support Vector Machine (SVM), the Random Forest (RF), the Decision Tree (DT), the K-Nearest Neighbors (KNN) and the Artificial Neural Networks (ANN). The accuracy according to Ram et al. [7] was 83.6% on CNN-based model. The highest accuracy of 91 % by Saha et al. [11] using Random Forest, and Chandrasekhar and Peddakrishna [12] at 93.44 % and 95 % respectively using ensemble methods on 303 UCI samples and IEEE Dataport 1190 samples respectively. The best performance was given by

Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Biswas et al. [13] in feature selection and highest accuracy of 94.5 % using SVM and LR. According to other researchers, 97% with DT-RF models Jawalkar *et al.* [14], 88.7% with hybrid ML models Kavitha et.al[15], 92 %with feature-engineered SVM Bouqentar, M. A., *et al.* [17], and 71.48% with SVM were also reported Sun *et al.* [19]. There are also hybrid and ensemble approaches to boost robustness whereby Mahajan et al. [5] have achieved between 97-99% accuracy. XGBoost gave Gnanavelu et al. [10] a 93% accuracy in a Kaggle cardiovascular dataset. The deep learning (DL) method has been studied in the recent past in prediction and prognosis of heart disease. Cenitta et al. [8] proposed Hybrid Residual Attention-based LSTM (HRAE-LSTM) model and recorded 97.71% accuracy which demonstrated the ability of time model. However, DL models tend to require massive data and can be hardly interpreted, which restricts their application to clinical practice. Although these improvements have been made, there are three key limitations prevailing in the available literature:

1. Most existing studies rely on the UCI Cleveland Heart disease dataset, which has a small limited sample size of 303 records [6];

2. Limited functions without the use of more complex cardiological tests, including ECG analysis, ECHO results, TMT results, and CAG types [8], [24];

3. Inadequate validation based on large, real-world clinical datasets, restricting generalizability.

Table 1 summarizes Selected representative ML-based heart disease prediction studies, highlighting datasets used, algorithms applied, sample sizes, and reported accuracies

Ref.	**Author(s)**	**Year**	**Dataset**	**Sample Size**	**Algorithms Used**	**Reported Accuracy (%)**	**Best Model**
[7]	Ram *et al.*	2024	UCI Cleveland	303	KNN, SVM, ANN, CNN	**83.61**	CNN
[10]	Gnanavelu *et al.*	2025	Kaggle CVD	-	DT, RF, KNN, NB, XGBoost	**93.00**	XGBoost
[11]	Saha *et al.*	2024	CDC	-	LR, DT, RF, KNN,XGB	**91.00**	**RF**
[12]	Chandrasekhar & Peddakrishna	2023	UCI Cleveland	303	RF,LR,KNN,NB, GB, AB, SVE	**93.44**	Ensemble
			IEEE DataPort	1190		**95.00**	
[14]	Jawalkar *et al.*	2023	UCI Cleveland	303	DT, RF Hybrid DTRF	**97.0**	Hybrid DTRF
[15]	Kavitha *et al.*	2021	UCI Cleveland	303	Hybrid ML	88.7	Hybrid

Received: August 09 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[17]	Bouqentar, M. A., *et al.*	2024	UCI Cleveland	303	RF,LR,KNN,NB, SVM, AB	**92.0**	SVM
[19]	Sun *et al.*	2021	UCI Cleveland	303	LR RF, SVM	**71.48**	SVM

Table 1 includes selected representative studies to highlight common datasets, performance trends, and limitations; additional related works are discussed in the text.

All these studies show that classical machine learning (ML) models may be correct enough; however, they cannot be applied in the clinic due to small sample sizes, the lack of sophisticated diagnostic functions, and the inability to validate them in practice. The enormous largest part of the research is founded on mere properties (age, blood pressure, cholesterol, and chest pain) without a possibility to obtain deeper diagnostic information to come to a clinical conclusion.

To fill such gaps, the present research work utilizes a real-life example of clinical data from 1,100 patients of **AMSHRC Pvt, Ltd, Vijayapur,Karnataka, India.** , who underwent advanced cardiological tests, such as ECG, ECHO, TMT, and CAG tests. Six standard ML models, i.e., Logistic Regression, Decision Tree, Random Forest, KNN, SVM, and XGBoost, are evaluated with the help of rigorous preprocessing, statistical verification, and comprehensive performance measures. The methodology presents a more valid and generalizable model of the clinical aspect of the past studies that used UCI data.

3.DATASET DESCRIPTION

The data includes 1100 anonymized records of patients that had been collected on inpatient and outpatient cardiovascular assessments from **AMSHRC Pvt, Ltd, Vijayapur,Karnatak, India**. All the reports, including demographic and vital signs, lifestyle, laboratory, and diagnostic test results, ECG, ECHO, TMT, and CAG, are documented. The case is reflective of current clinical practice, and the dataset represents instances of some diagnostic tests not being conducted according to medical contraindications based on the patient's condition. The table 1 below summarises the input attributes description to be used in predicting Heart Disease using machine learning.

Table 2 Dataset Feature Description

Feature	**Description**
Age	Age in years (20–95).
Sex	1 = Male, 0 = Female.
BP_systolic	Systolic blood pressure (mmHg); normal range: 90–140.
BP_diastolic	Diastolic blood pressure (mmHg); normal range: 60–90.
PR	Pulse rate (bpm); normal range: 60–100.
SpO_2	Oxygen saturation (%); normal: 95–100.
Weight	Weight in kilograms.

Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

CP	Chest pain type: 0 = No Pain 1 = Mild Pain 2 = Moderate Pain, 3 = Severe pain
Chol	Serum Cholesterol (mg/dL); 0=No, 1= Yes
Diabetics	1=Diabetic, 0= Non Diabetic
Hypertension	1 = Hypertensive, 0 = Non Hypertensive
Habits	0 = Healthy lifestyle, 1 = Smoker/Tobacco/Alcoholic/Unhealthy habits.
ECG	Resting ECG: 0 = Not performed, 1 = Normal, 2 = Abnormal.
ECHO	Echocardiography: 0 = Not performed, 1 = Normal, 2 = Abnormal.
TMT	Treadmill test: 0 = Not performed, 1 = Negative, 2 = Positive.
CAG	Coronary angiography: 0 = Not performed, 1 = Normal, 2–4 = Vessel disease (SVD/DVD/TVD).
Target	1 = Heart disease present, 0 = No disease.

Clinically relevant demographic, physiological, and diagnostic variables are encoded for computational modeling in Table 1. The feature set is presented above. The presence of Age and Sex are known to have a role in predicting cardiovascular risk using baseline demographic measures. The vital cardiovascular function is captured by hemisphere parameters like BP_systolic, BP_diastolic ,Pulse Rate(PR) and SpO2 represented as continuous physiological measurements. The measurement of lipid and glucose profiles is done by metabolic indicators such as serum cholesterol (Chol) and blood sugar (Diabetics), while the use of Hypertension and Habits provides binary markers for chronic conditions and lifestyle risk factors. In cardiology, a four-level classification system known as chest pain type (CP) is used to encode the presentation of symptoms. Advanced cardiological diagnostics make up a significant portion of the dataset. Hence, the prediction framework incorporates structural (external) electrical and functional cardiac assessments as well as temporal or other undefined variables such as resting ECG, ECHO, TMT with categorical constraints on whether these states are normal, abnormal or not-performed. The diagnostic information provided by these features is of high resolution, which cannot be replicated in benchmark datasets such as UCI. The presence of heart disease is confirmed through both the target attribute and the outcome. The feature design enables the multilevel characterization of cardiovascular status by merging routine clinical measurements with advanced cardiac investigations, enabling robust machine learning-based classification.

4. METHODOLOGY

The present section discusses the methodological approach used to develop baseline machine learning (ML) models for heart disease prediction using a real-time dataset of 1100 patients from **AMSHRC Pvt, Ltd, Vijayapur,Karnatak, India.** Additionally, It includes describing the dataset,defining its features and preprocessing of data for statistical analysis using models

Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

and methods; and providing instructions for conducting evaluation strategies. Figure 1 (conceptual flow) illustrates the proposed methodological workflow.

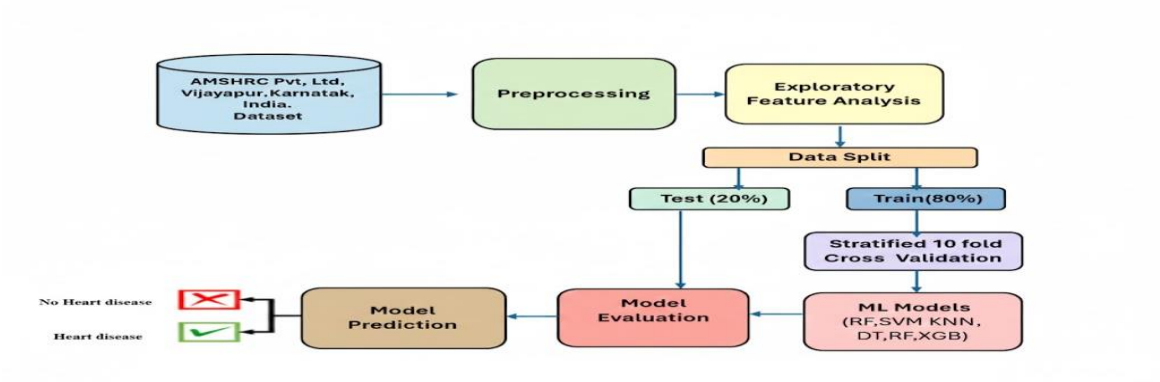


Figure 1 Proposed machine learning workflow on predicting heart disease

Figure 1. Shows proposed machine learning workflow on predicting heart disease. It starts with the acquisition of data based on actual clinical records and then there are preprocessing steps, such as encoding, scaling, and missing-value treatment. To gain insights into the structure and meaning of clinical features, Exploratory Data Analysis (EDA) and statistical analysis are done. The cleaned data is subsequently split into 80/20 stratified train-test. There are several machine learning models that are trained and then 10-fold cross-validation is stratified to guarantee robustness. Accuracy, precision, recall, F1-score, ROC-AUC ,MCR and MCC are performance measures that are calculated. The analysis of the feature importance is done to determine the influential predictors and final results are the predicted presence or absence of heart disease.

4.1.**Data Collection**

The information was gathered retrospectively and anonymously. The inclusion was based on adult patients (over 20 years of age) presenting with a cardiac examination The balance in representation of both cases of heart disease and non-disease is represented in the dataset to ensure the intensity of the ML models.

4.2. **Data Pre-processing.**

Most clinical datasets in practical use exhibit a range of outlier and feature-based classification schemes, including missing values, assorted feature types, outliest attributes, and diagnostic codes that require meticulous analysis preparation[5],[12],[13][30]. The use of a preprocessing pipeline was made in accordance with the established protocols in medical ML research and but also to guarantee reproducibility.

4.2.1. **Removal of Highly Missing Features.**

The removal of features with high missingness (>99%) was done to prevent unreliable imputations and noise. Clinical ML workflows that prioritize particular diagnostic tests require this approach and to be performed.

4.2.2. **Missing-Indicator Feature Construction.**

Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The presence or absence of data was explicitly indicated by binary indicator features in variables that were partially missing. By capturing meaningful clinical patterns, this approach has been proven to enhance model robustness in hospital datasets, particularly for ECG, ECHO, TMT, and CAG

The characteristics were categorized as numeric or categorical, and their nature differed. Numerical variables with low cardinality (Sex, Hypertension, Habits, and Diabetics) were transformed into categorical variables in order to enable the right encoding. The feature typing is made properly to avoid biased transformation and following the best practices in healthcare ML preprocessing

4.2.4 **Numerical Imputation and Scaling of Numerical Features.**

The continuous clinical variables (e.g., blood pressure, cholesterol, SpO2, pulse rate) were imputed by the median to ensure that these variables are resistant to skewed distributions . It was then normalized using standardization in terms of z-score in order to support algorithms that were sensitive to the sizes of features, including SVM and KNN.

4.2.5 Encoding of Categorical Variables.

The most common category was entered in categorical variables (CP, ECG, ECHO, TMT, CAG, and lifestyle-related variables) and coded using one-hot encoding. If the unknown categories exist at test time ,the handle unknown= ignore option was used to safely process such values without disrupting model inference, as recommended by clinical ML systems

4.2.6 Outlier Handling.

The method of interquartile range (IQR) to detect outliers was adopted, and the outliers were not removed but were clipped to physiologically reasonable values. This preserved distinctive but clinically relevant values, which are informed by previous cardiovascular ML studies cautioning against the elimination of potentially large extremes [12],[17].

4.2.7. Final Dataset Construction and Splitting.The dataset was divided into training (80%) and testing (20%) using stratified sampling to ensure class balance. All preprocessing steps were merged into one pipeline for the scikit-learn to ensure identical transformations during training and inference. Further, the total samples considered, heart disease patients, training, and testing set samples are given in Table 3.

Table 3 Total, Training and Testing Samples

Case	Training (80%)	Testing (20%)
Normal	**556**	**139**
Heart Disease	**324**	**81**
Total	**880**	**220**
Total Samples	**1100**	

4.3 STATISTICAL ANALYSIS :

Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Prior to model development, exploratory data analysis was conducted to summarize the characteristics of the dataset and to support subsequent machine learning modeling. Continuous variables were described using means, medians, standard deviations, and ranges, while categorical variables were summarized using frequencies and percentages, following established practices in clinical data analytics . Correlation analysis was performed to examine relationships among predictors and to assess potential multicollinearity. Pearson correlation coefficients were used for continuous variables, while Spearman correlations were applied for ordinal or non-normally distributed features, consistent with standard medical informatics methodology.

4.4 Machine Learning Models

To establish reliable baseline models for heart disease prediction, six widely used machine learning (ML) algorithms were selected. Their inclusion is supported by their frequent use in cardiovascular research and by strong evidence demonstrating high performance on structured clinical datasets [3], [5], [7], [9], [12], [17], [18], [20], [24], [27], [28]. Together, these algorithms represent a broad spectrum of modeling strategies—from simple linear relationships to complex non-linear interactions—allowing for a balanced and comprehensive comparison of predictive capability.

4.4.1 Logistic Regression (LR)

Logistic Regression (LR) is a widely used statistical classification method applied in heart disease detection due to its interpretability, computational efficiency, and ability to model clinical risk factors. LR predicts the likelihood of heart disease by modelling the relationship between features such as age, cholesterol, blood pressure, and ECG abnormalities using a logistic function. The LR model estimates the probability of disease according to the linear combination of input features as shown in Eq. (1).

Eq. (1):

$$P(Y_i = 1|x_i) = \frac{1}{1+e^{-(wx_i+b)}} \quad (1)$$

In Eq. (1), **w** represents the weight vector associated with clinical features, **b** is the bias term, $\mathbf{x_i}$ denotes the feature vector of the i-th patient, and $\mathbf{P(Y_i = 1 \mid x_i)}$ is the predicted probability of heart disease. LR continues to be used extensively used CVD prediction because of its transparency and strong baseline performance in clinical datasets [9],[18], [19]

4.4.2 k-Nearest Neighbors (KNN)

The k-Nearest Neighbors (KNN) algorithm is a simple yet effective supervised learning method widely applied in heart disease detection due to its ability to classify patients based on similarity to previously diagnosed cases. kNN assesses clinical features such as heart rate, cholesterol, blood pressure, and chest pain type by measuring the distance between the test patient and existing labeled patients. The Euclidean distance, commonly used in clinical datasets, is defined in Eq. (2).

$$d(x_i, x_j) = \| x_i - x_j \|_2 \quad (2)$$

Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$\mathcal{N}_k(x_i) = \{x_j \mid x_j \text{ is among the } k \text{ nearest samples to } x_i\} \quad (3)$$

Based on this distance, the prediction for patient x_i is determined using majority voting among its k nearest neighbors as expressed in Eq. (4).

$$\hat{y}_i = \text{mode}\,\{y_j : j \in \mathcal{N}_k(x_i)\} \quad (4)$$

In Eqs. (2)-(4), $\mathbf{x_i}$ is the feature vector of the i-th patient, $\mathbf{N_k(x_i)}$ refers to the set of its k nearest neighbors, and $\mathbf{y_j}$ denotes their class labels. KNN is particularly useful in CVD prediction because patient clusters with similar physiological features. KNN is particularly useful in CVD prediction because patient clusters with similar physiological features often correlate strongly with disease presence [13],[17],[**22],**

4.4.3 **The SVM** is a powerful supervised learning algorithm widely applied in heart disease detection due to its robustness in handling high-dimensional and non-linear medical data. The primary objective of SVM is to construct a hyperplane that maximally separates the classes, i.e., in this work, patients with heart disease and those without, based on their clinical features such as age, cholesterol, resting blood pressure, and ECG outcomes. Mathematically, the SVM attempts to solve the following optimization problem using Eq. (5).

$$\min_{w,b} \quad \frac{1}{2} \| w \|^2 \text{ subject to} y_i(w \cdot x_i + b) \geq 1,\ i = 1,2,\dots,n \quad (5)$$

In Eq. (5), $\mathbf{w}$ is the weight vector perpendicular to the hyperplane, $\mathbf{b}$ is the bias term, $\mathbf{x_i}$ represents the feature vector of the i-th patient, and $\mathbf{y_i} \in \{-1, +1\}$ denotes the class label (diseased or non-diseased). By maximizing the margin between classes, SVM enhances the model's generalization ability and reduces misclassification. For non-linear data relationships, kernel functions such as linear, polynomial, or Radial Basis Function (RBF) are applied using Eq. (6).

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \quad (6)$$

In Eq. (6), $\boldsymbol{\phi}$ maps input features into a higher-dimensional space to allow linear separation, enabling the classifier to capture complex clinical patterns. SVM has consistently demonstrated competitive performance in predicting CVD due to its margin maximization and kernel-based learning capability [7], [19], [28].

4.4.4. Decision Tree (DT)

The Decision Tree (DT) is a rule-based supervised learning algorithm frequently applied in heart disease detection due to its transparency and resemblance to clinical decision-making. DT recursively splits the dataset into subgroups based on features such as cholesterol level, blood pressure thresholds, chest pain type, and ECG characteristics. The quality of each split is evaluated using an impurity measure such as the Gini index shown in Eq. (7).

$$I(t) = 1 - p_{0,t}^2 - p_{1,t}^2 \quad (7)$$

In Eq. (7), $\mathbf{p_{0},t}$ and $\mathbf{p_{1},t}$ denote the proportions of healthy and diseased patients in node t. A split is selected to maximize impurity reduction as defined in Eq. (8).

Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$\Delta I = I(t) - \left(\frac{N_L}{N} I(t_L) + \frac{N_R}{N} I(t_R)\right) \quad (8)$$

Here, $\mathbf{N^L}$ and $\mathbf{N^R}$ represent the number of patients in the left and right child nodes, respectively. Where N is the total numbers of samples at the parent node. DT models have been shown to perform effectively in CVD classification due to their interpretability and ability to capture nonlinear decision boundaries [10],[11], [14].

4.4.5. Random Forest (RF)

Random Forest (RF) is a powerful ensemble learning method applied extensively in heart disease prediction due to its stability, resistance to overfitting, and ability to learn complex clinical patterns. RF generates multiple decision trees using bootstrap sampling and random feature selection. The final classification is obtained through majority voting as shown in Eq. (9).

$$\hat{y}_i = \text{mode}\{h_m(x_i): m = 1,2,\dots,M\} \quad (9)$$

In Eqs. (9) , $\mathbf{h_m}$ denotes the prediction from the m-th tree, and **M** is the total number of trees in the forest. RF has consistently demonstrated strong predictive power and robustness in heart disease detection across diverse clinical datasets [13], [18],[29].

4.4.6. XGBoost (Extreme Gradient Boosting)

XGBoost is an advanced boosting algorithm that has gained significant importance in heart disease prediction due to its superior accuracy, scalability, and capability to model complex nonlinear feature interactions. XGBoost constructs an ensemble of regression trees trained sequentially to correct the errors of previous trees. The final prediction is expressed in Eq. (10).

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \quad (10)$$

In Eq. (10), l denotes loss function, y denotes actual label and $\hat{y}_i$ denotes predicted label and $\Omega(f_k)$ denotes regularization term controlling model complexity. Recent studies have highlighted XGBoost's superior performance in predicting CVD risk compared to traditional models [10], [12],[27].

The whole process of preprocessing, encoding, scaling, model training, evaluation, and visualization was programmed in Jupyter/Colab Notebook, which ensured the same partitions of data and preprocessing conditions, ensuring fair comparison of models.

4.5 Model Evaluation

The performance of the trained models i.e. Support Vector machine (SVM), random forest (RF), k-Nearest neighbors (KNN), decision tree (DT) and XGBoost with respect to forecasting heart diseases have been evaluated based on popular performance indicators i. e. accuracy, precision, recall and F1-score on the test dataset. Accuracy merely provides a rough notion of the data that have been properly classified but in clinical practice, more quantifications are required to quantify the predictive performance in terms of classes. Class 0 in the current study is the absence of heart disease and class 1 is its presence. In addition to accuracy, the Misclassification rate (MCR) was also used to measure prediction error. In assessing the

Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

discriminative performance between the thresholds, ROC-AUC was applied where Matthews Correlation Coefficient (MCC) provided a trade-off test of the performance by putting all the elements of the confusion matrix into consideration and put into consideration the gap between the classes. The time cost of computation of each algorithm was measured with the help of Python time module.

The accuracy is evaluated using Eq. (11), where TP denotes true-positive, FP denotes false-positive, FN denotes false positive and TN denotes true-negative.[12],[16]

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (11)$$

Precision is evaluated using Eq. (12), which is the ratio of the number of correct positive predictions among all those the model identified as positive, or the extent to which the positive predictions can be trusted[14].

$$Precision = \frac{TP}{TP + FP} \qquad (12)$$

Recall is evaluated using Eq. (13), which refers to the proportion of correct positive responses that the model has made, or what the model can identify[25].

$$Recall = \frac{TP}{TP + FN} \qquad (13)$$

The F1-score is evaluated using Eq. (14), which is a harmonic combination of two measures, precision and recall, that weigh both. It comes in especially when you are training models with skewed data. These indicators provide the complete picture of the categorical performance of any model, [24], [26]

$$F - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \qquad (14)$$

Specificity refers to the degree to which the model is able to classify negative cases correctly (patients without the disease). Its importance in the medical use is critical because it can be used to ascertain the effectiveness of the model in avoiding falses since healthy patients would not be falsely diagnosed[13]. Specificity is defined as in the equation below (15).

$$Specificity = \frac{TN}{TN + FP} \qquad (15)$$

All the measures of evaluation were executed with standardized Scikit-learn functions to promote reproducibility and methodological consistency [5], [7], [10]–[14], [18].

4.5.1 K-Fold Cross-Validation

Stratified 10-fold cross-validation was used to achieve robustness even in data sets that were not too large (i.e., moderate size). This is a popular cross-validation method with moderately sized clinical data [12], [17]. The data was divided into ten folds of equal sizes, and one of the folds was used as the validation subset, and the rest were utilized as the training folds. Stratification maintained the distribution of classes and did not favour one side of the disease and non-disease. Each of the models was computed to produce mean and standard-deviation values of Accuracy, Precision, Recall, F1-score, and ROC-AUC values across all folds, which is a more reliable and generalizable estimate of its performance compared to a single split .

Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

All preprocessing steps, including imputation, encoding, and scaling, were performed within each training fold to prevent data leakage.

5.RESULTS AND DISCUSSION:

5.1 Results

In this section, the results of machine learning framework tested on a real-time clinical cardiac dataset are in brief. Feature relevance was assessed using model-based importance scores from ensemble learning methods, providing insight into clinically influential predictors.

Exploratory data visualizations, feature importance analysis, and performance evaluation of six (LR,SVM,KNN,DT, DF and XGBoost)supervised machine learning models were conducted The statistical test determines features that have significant differences between the groups, which justifies their clinical importance. Distribution, correlations, and the proportion of classes visualization should also serve as an extra proof of preprocessing procedures, as they are the best practices in medical data analytics . The evaluation of model performance entails the Accuracy, Precision, Recall, F1-score, ROC-AUC, MCC, and confusion-matrix measures, as the commonly used measures in recent CVD ML literature [17], [22], [30].

5.1.1 Cross-Validation Results

.Table 4 is an extension of the analysis conducted on the test-sets previously but they provide the mean performance of the entire set of classifiers recovered by running k-fold(10) cross-validation in mean +/- standard deviation. Although the central result of the model behavior analysis was provided in form of confusion matrices and ROC curves on one test split analyzed in the earlier tables and figures, the cross-validation results ensure validity and consistency of results on an immense number of data partitions. Once again the ensemble models and specifically the Random Forest and XGBoost have the highest accuracy (0.945), recall (0.933, 0.931), F1-score(0.925, 0.926) as well as ROC-AUC (0.983) indicating that that the accuracy generalizes better. Similarly to the above where they are conservative in their classification behaviour, both the Regression Logistic and SVM are of high precision but relatively low recall. The trade-off between the predictive and computational cost can also be identified since the time taken by the cross-validation can also be, involved. As a rule, cross-validation analysis confirms the outcomes of the prior test-set results and makes the proposed classification framework more real.

Table 4 Cross-validated performance of machine learning models (mean ± standard deviation).

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC	CV Time (s)
Logistic Regression	0.922 ± 0.016	0.941 ± 0.037	0.842 ± 0.036	0.888 ± 0.023	0.979 ± 0.011	0.67
SVM	0.914 ± 0.022	0.952 ± 0.036	0.807 ± 0.047	0.873 ± 0.035	0.980 ± 0.012	2.92

Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Decision Tree	0.937 ± 0.021	0.925 ± 0.031	0.904 ± 0.044	0.914 ± 0.030	0.974 ± 0.018	1.28
KNN	0.915 ± 0.016	0.910 ± 0.032	0.857 ± 0.052	0.881 ± 0.026	0.950 ± 0.017	0.60
Random Forest	0.945 ± 0.022	0.918 ± 0.026	0.933 ± 0.045	0.925 ± 0.031	0.983 ± 0.011	2.57
XGBoost	0.945 ± 0.023	0.923 ± 0.035	0.931 ± 0.048	0.926 ± 0.032	0.983 ± 0.011	2.43

The feature-importance algorithm (Figure 2) shows that the andragogy of CAG, ECHO, ECG, TMT and CP are the most significant predictors, which is consistent with the clinical evidence of the literature that highlighted the prognostic power of anatomical and functional cardiac measurements . Clinical parameters (Age, BP, PR) were found to be moderately influential, whereas metabolic and lifestyle characteristics features showed lower influence.

5.1.2 Feature Importance & Visualizations

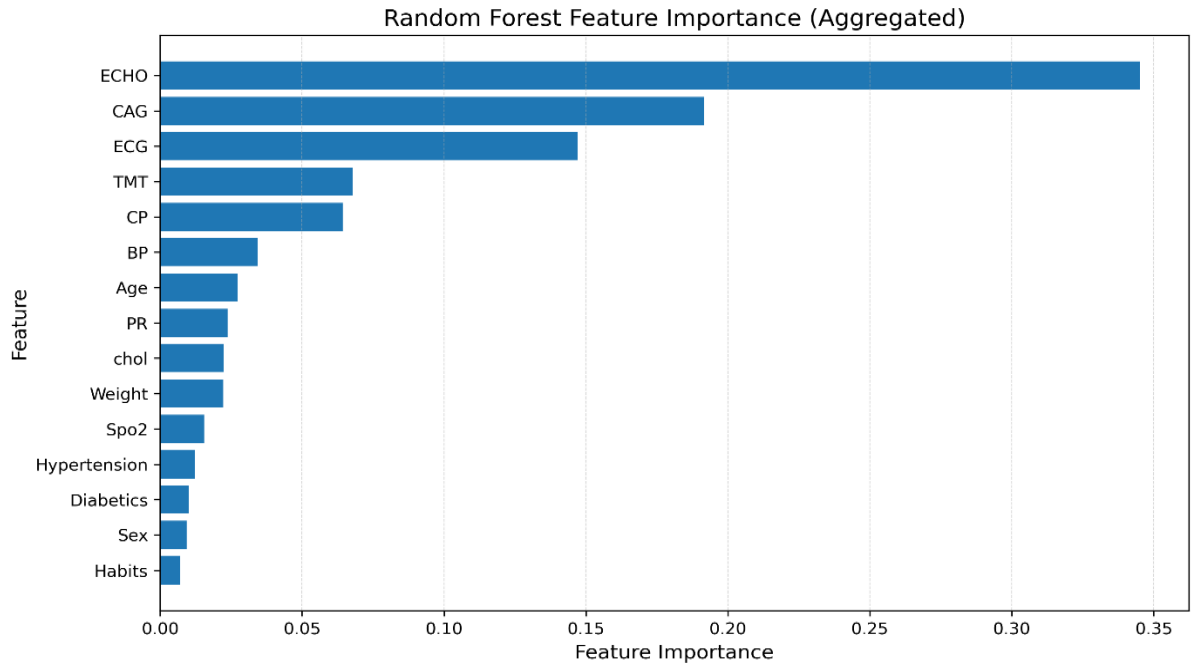


Figure 2 . Random Forest Feature Importance for Heart Disease Prediction.

To depict the structural properties of the dataset and the predictive models performance, articulate representation of the dataset in the form of visualizations (Figure 3–Figure 9) was utilized. The figure distributions and demographic graphs (Figure 3-Figure 7) indicate variation in the clinical variables, imbalance of classes, and population characteristics, which provides a hint about the background trends that can affect the model usage. The heatmap of correlation (Figure 3) shows that there are strong correlations between the features and that the multicollinearity may also occur. In addition, the confusion matrices and ROC curves of all the

Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

classification models (Figure 9) make visual assessment of the discriminative capability, classification errors, and generally the power of the individual model. All of these visualizations altogether contribute to the acquisition of the general image of the nature of the data as well as the comparative outcomes of the machine learning models.

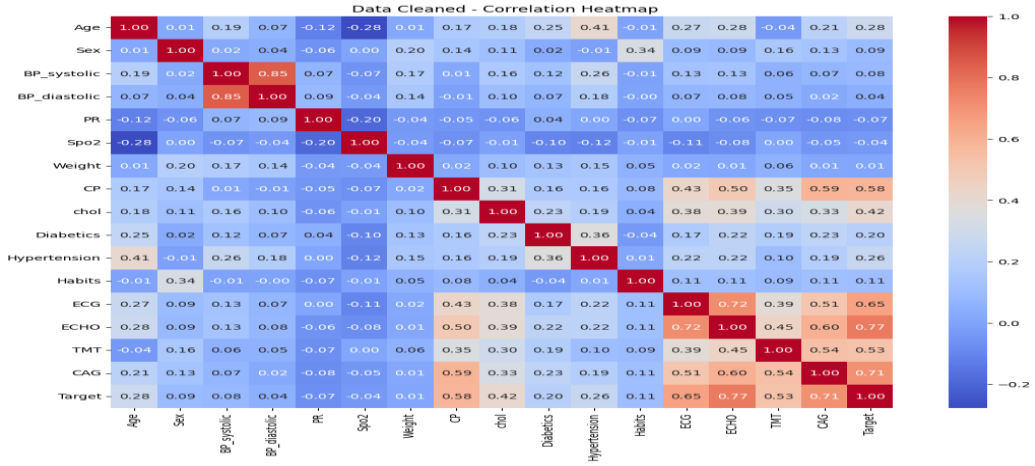


Figure 3 Correlation Heatmap of Dataset.

The correlation heatmap of the cleansed clinical data presented in Figure 3 indicates that ECG, ECHO, TMT and CAG are predictive because they show positive correlation with the heart disease target. Clinical risk factors (chest pain, cholesterol level, hypertension, diabetes, and age) have moderate correlations because it is expected by the existing cardiovascular risk profiles. On the other hand, such demographic and physiological factors as sex, weight, pulse rate, and SpO 2 demonstrate low correlation. The multicollinearity observed by the high inter-feature correlations, in particular, systolic and diastolic blood pressure, and the diagnostic modalities is successfully addressed using the selected machine learning models.

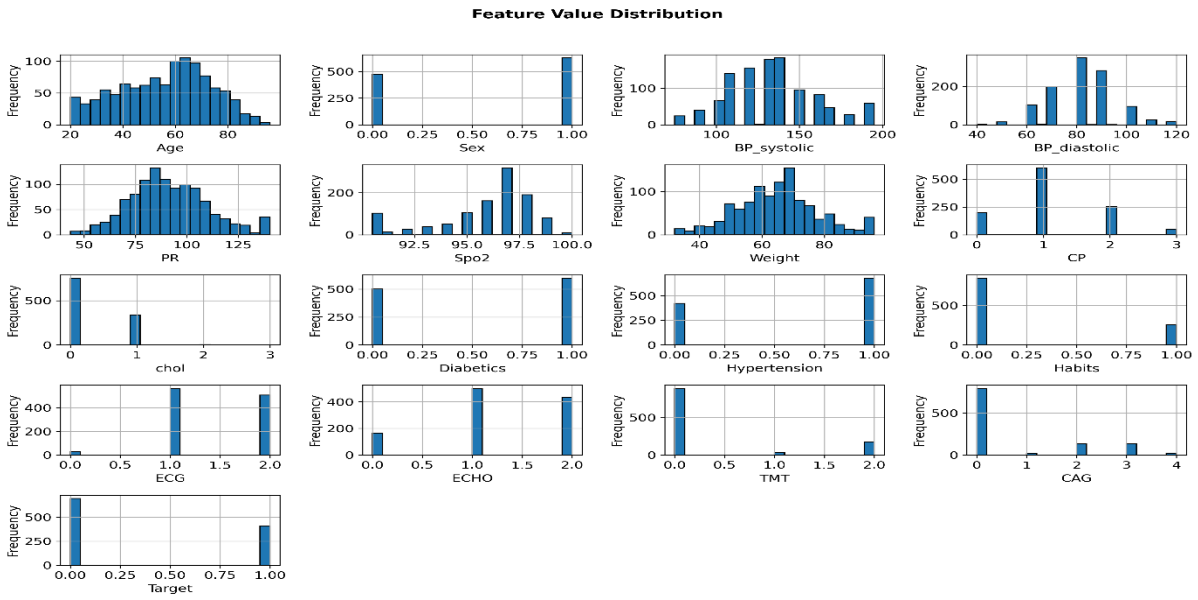


Figure 4 Distribution of Clinical, Demographic, and Diagnostic Feature Values in the Cleaned Dataset

Figure 4. shows Distribution of Clinical, Demographic, and Diagnostic Feature Values in the Cleaned Dataset presents how the values of the features to be used in the current research

Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

should be distributed. Continuous variables namely age, systolic blood pressure, diastolic blood pressure, pulse rate, and weight have about unimodal distributions with age data ranging 40-70 years, systolic blood pressure of 110-160 mmHg and diastolic blood pressure of 70-90 mmHg, pulse rate 70-100 beats per minute, and weight between 60-80kg. There is no large variability of SpO 2 with a difference of 95-100 report. The sex, chest pain, cholesterol, diabetes, hypertension, habits, ECG, ECHO and desired variable are categorical and binary with the apparent skew of classes existing in some of them. Overall, the distributions of the observed features assumed that the data features were stable, and the skewness is not extreme , supporting the suitability of the dataset for subsequent classification experiments.

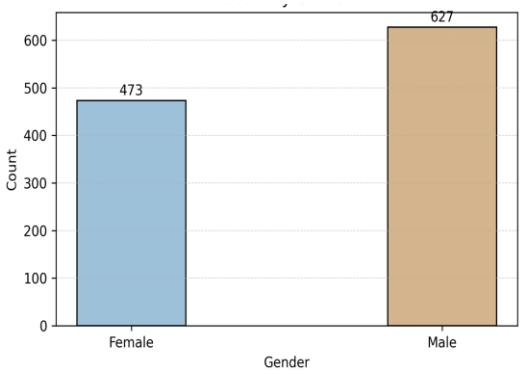


Figure 5 Gender Distribution of dataset

Figure 5 show the gender balance that the proportion of males to the dataset is larger with 627 males and 473 females. The skew in this shows that the data is moderately skewed towards male subjects, and this may have an impact on behavior in the model were there are any gender specific clinical patterns in different groups. This difference should be thus considered in the downstream analyses and predictive modelling so that no unfair performance is observed in the demographic subpopulations.

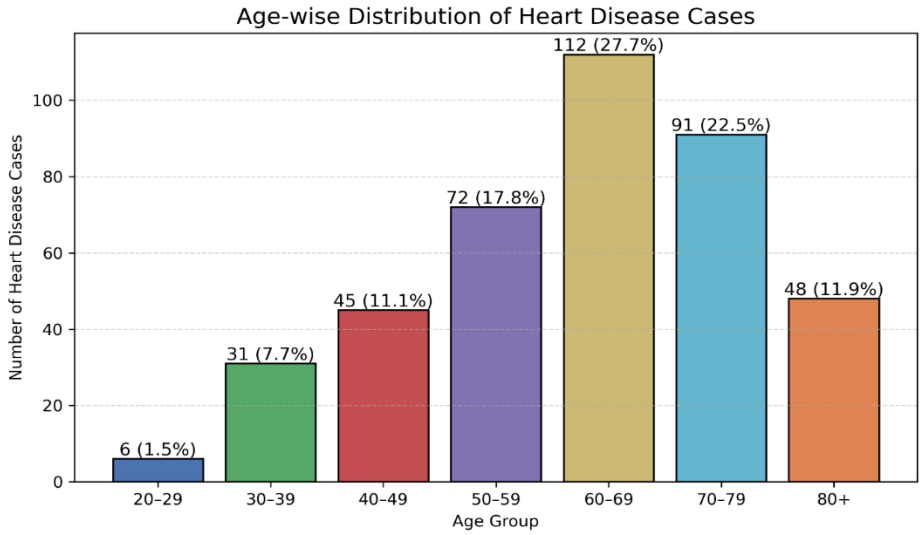


Figure 6 Age-wise Distribution of Heart Disease Cases.

Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Fig. 6. Age-wise Distribution of heart disease cases reveals the percentage relationship of the case incidences of heart disease in various ages. The highest rates are between 60-69 (28.2%), 70-79 (22.0%), and 50-59 (17.6%), which implies that heart diseases are increasing at a high rate among the elderly people. Middle-aged ages 40 to 49 (11.1%) and 30 to 39 (7.7%) contribute relatively little, while very young people, aged 20 to 29 (1.5%), also make a minimal contribution. The 80 plus and above segment accounts for 11.4 percent of the cases, showing that the prevalence of old age is still present but slightly less in the 80 plus and above category. Overall, the figure shows that age is a severe risk factor, and the cases of heart diseases are centered on individuals aged 50 years and above.

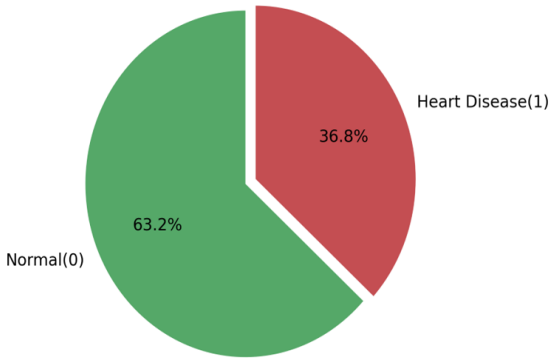


Figure 7 Class Distribution of the Heart Disease Target Variable

Fig. 7. Heart Disease Target Class Distribution illustrates the sample size rate of heart disease in the sample and the rate of sample size without heart disease in the sample. The majority of the normal cases (0) represents 63.2% of the samples, while 36.8% represent the heart disease class (1). This implies that there is an intermediate disparity in class towards non-heart disease cases. Given the observed disproportion, which is not severely skewed, the application of corresponding evaluation measures, such as ROC-AUC and confusion matrix-based measures, is necessary to offer consistent measures of classification performance.

5.1.3 Model Performance

Table 5 and Figure 8 is an abridged performance comparison of six classification models including: Random Forest, XGBoost, KNN, Decision Tree, Logistic Regression and SVM with RBF kernel were evaluated using the values of accuracy(Acc), Precision(Prec), recall, F1-score, ROC-AUC and Matthews Correlation Coefficient (MCC), Misclassification Rate (MCR), and confusion matrix. Random Forest and XGBoost were the most accurate at 93 % and high recall of 0.88 and 0.89 respectively and identical F1-score of 0.90 or good coherent classification. XGBoost showed the most true positive values (TP = 72), and the least negative values (FN = 9), whereas the same values were recorded in the case of the Random Forest (TP = 71, FN = 10); the highest value of MCC was 0.84 and the lowest value of MCR was 0.07 in XGBoost and the Random Forest, respectively, which shows that these two models are good predictors. KNN and Decision tree performed at the same level with the 91 % accuracy with a mild increase in the misclassification rates with slight increment in the MCC value of 0.80. The best precision belonged to Logistic Regression and SVM(0.95), and very low false positives (FP = 3), but lower precision due to greater false negatives, resulting in lower values of MCC of 0.75 and 0.71, respectively. Computational efficiency part comprised of Decision Tree and KNN that

Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

took the shortest execution time as compared to the more time consuming random Forest and SVM. Overall, among the Random Forest and XGBoost, it could be concluded that the ensemble-based models were the most successful models in comparison with the traditional classifiers because of the best proportion of accuracy and robustness and minimal computational expenses that would be more suitable in the suggested classification problem.

Table 5 Comparative Performance Evaluation of Traditional and Ensemble Machine Learning Classifiers

Model	Acc	Preci	Recall	F1 Score	ROC-AUC	MCC	MCR	TP	FP	TN	FN	Time (s)
RF	0.93	0.92	0.88	0.90	0.98	0.84	0.07	71	6	133	10	0.29
XGBoost	0.93	0.91	0.89	0.90	0.98	0.84	0.07	72	7	132	9	0.11
KNN	0.91	0.94	0.80	0.87	0.94	0.80	0.09	65	4	135	16	0.06
DT	0.91	0.92	0.83	0.87	0.96	0.80	0.09	67	6	133	14	0.05
LR	0.88	0.95	0.72	0.82	0.98	0.75	0.12	58	3	136	23	0.06
SVM	0.86	0.95	0.67	0.78	0.98	0.71	0.14	54	3	136	27	0.27

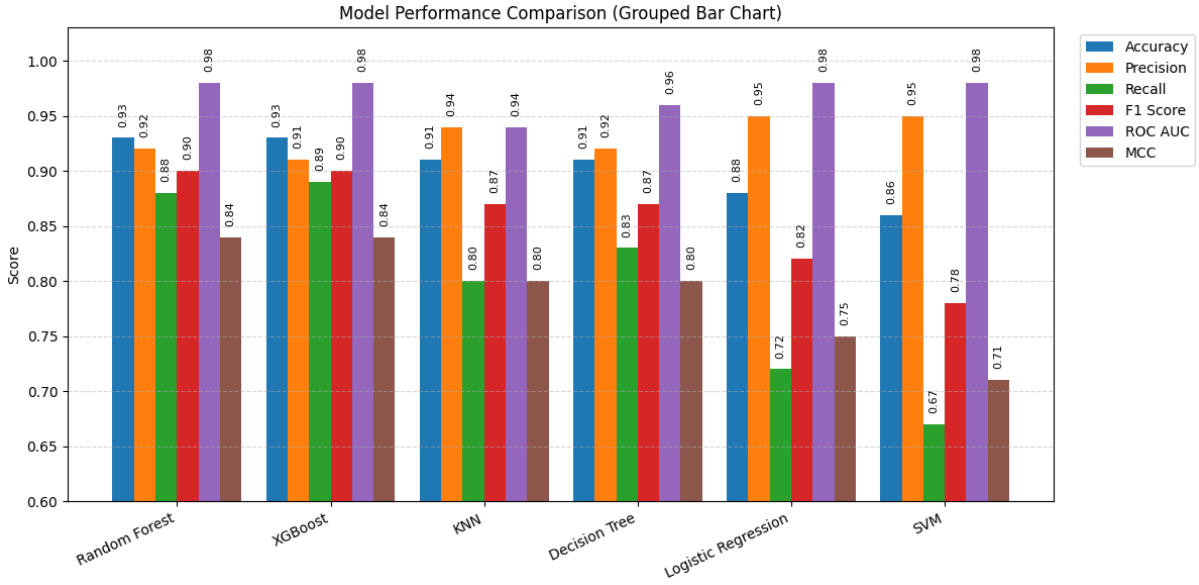
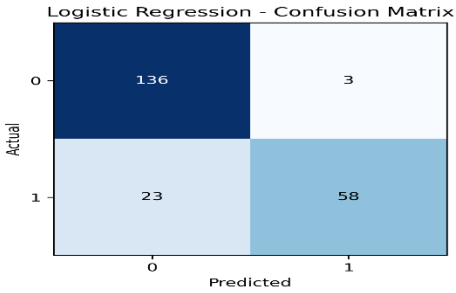


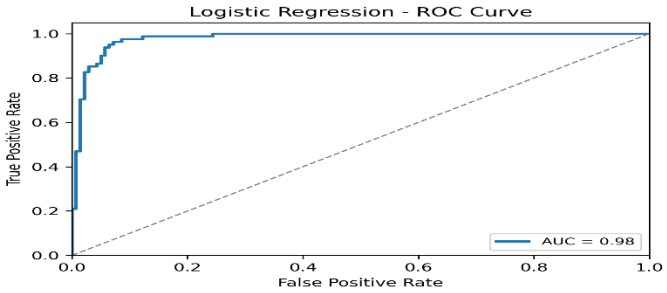
Figure 8 Comparative Analysis of Machine Learning Models Based on Multiple Performance Metrics

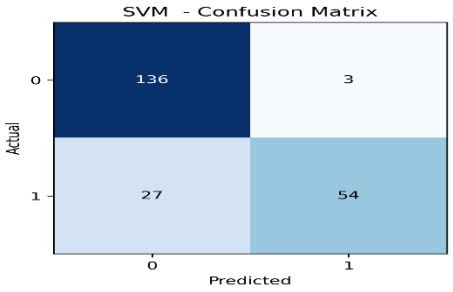
Figure 9 shows confusion matrices and ROC curves, indicating that all the experimented models were found to be highly discriminative, and the ROC-AUC values were always above 0.94. Random Forest and XGBoost showed the most reasonable pattern of classification, with a high true-positive rate and a low false classification rate. The same applies to Logistic Regression and KNN, which demonstrated consistency, but SVM did not, as it demonstrated low sensitivity with a high degree of accuracy. Overall, it may be concluded that tree-based ensemble models provide the most powerful and accurate predictions for identifying heart diseases in this dataset.

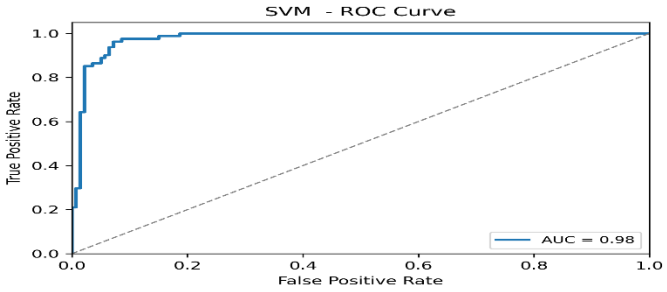
Received: August 09 ,2025

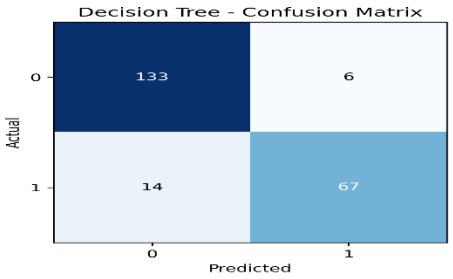
International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

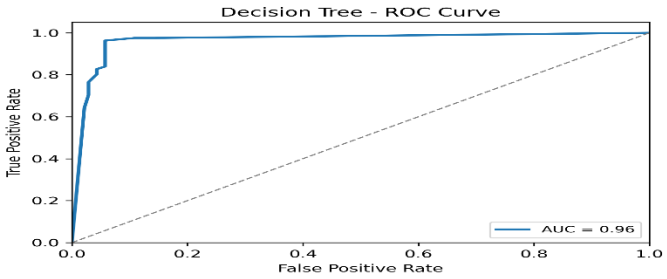


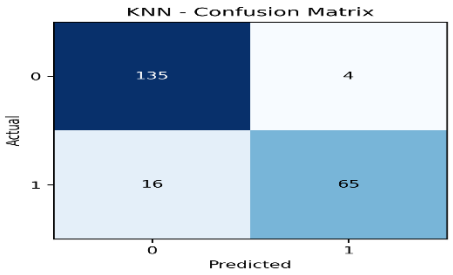


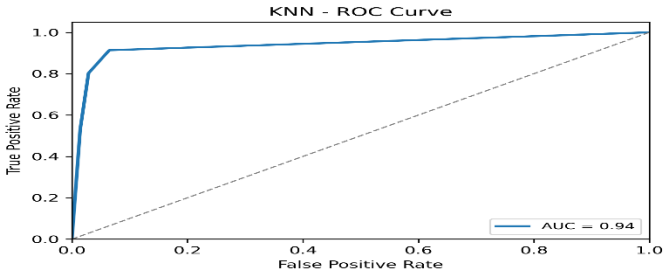












Received: August 09 ,2025

International Journal of Applied Mathematics
Volume 38 No. 12s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

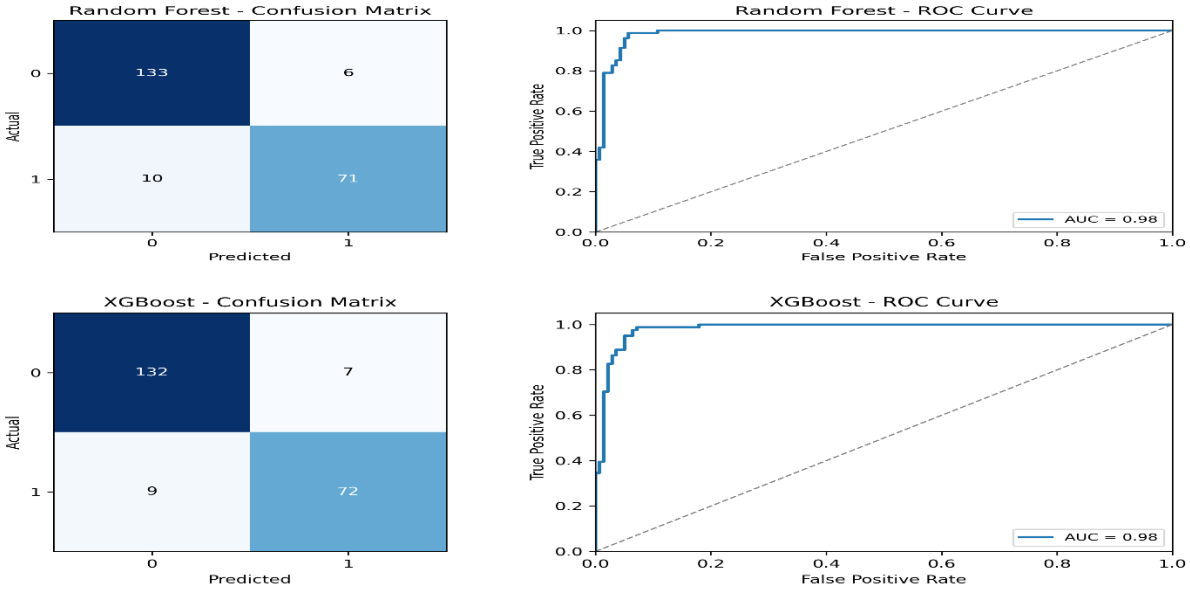


Figure 2 Confusion Matrices and ROC Curves for All Classification Models.

5.1.4 Comparison with Prior Work

The table 7 presents the comparative analysis between the proposed approach and the recent research on the classification of heart diseases. Ram et al. [7] found that CNN was able to achieve a maximum accuracy of 83.6% using the UCI Cleveland data set (303 samples), but Kavitha *et al.* [15] achieved 88.7% when using DT and RF with the same dataset. Bouqentar, M. A.,et al. [17] also utilized a UCI dataset to determine 92% accuracy with the usage of SVM. Alternatively, the proposed research evaluates the conservative, as well as ensemble, classifiers, including LR, DT, RF, KNN, SVM, and XGBoost, on a significantly larger clinical sample (AMSHRC) with 1100 samples. Compared to the already existing research, the proposed method gives a higher accuracy of 93%, and XGBoost and RF are identified as the most successful models. It can be explained by enhanced clinical information and more extensive data.

Table 6 Comparative analysis of proposed model against established methodologies

Ref.	**Author(s)**	**Year**	**Dataset**	**Sample Size**	**Algorithms Used**	Reported Accuracy (%)	**Best Model**
[7]	Ram *et al.*	2024	UCI Cleveland	303	KNN, SVM, ANN, CNN	83.6	CNN
[17]	Bouqentar, M. A., *et al.*	2024	UCI Cleveland	303	RF,LR,KNN,NB, SVM, AB	92.0	SVM

Received: August 09 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[15]	Kavitha *et al.*	2023	UCI Cleveland	303	Hybrid DTRF	88.7	Hybrid DTRF
	Proposed		**AMSHRC Pvt, Ltd, Vijayapur,Karnataka, India**	**1100**	**LRDT, RF, KNN, SVM XGB**	**93**	**XGB and RF**

5.2 DICSUSSION

The findings of this article indicate that machine learning models can be useful in the categorization of heart disease when a real-life clinical data containing both normal physiological parameters and contemporary diagnostic variables is used. Ensemble models, specifically XGBoost and Random Forest exhibited the best performance in the context of accuracy, recall, F1-score, ROC-AUC and confusion matrix analysis; this shows properly robust and high generalization. As previous studies have shown, logistic regression and SVM were highly precise but low recalling, meaning that they tend to behave cautiously in prediction and confirm that the information is not perfectly linearly separable. Considerable prognostic features, including the type of chest pain, treadmill tests, coronary angiography, and echocardiography, were well in line with established clinical data. The improved performance in contrast with the past researches that utilize limited benchmark data such as the UCI Cleveland might reflect the advantage of more diagnostic information and more patients.The high internal validity can be achieved due to the consistent findings of stratified train-test analysis and cross-validation, however, external validation is required due to the fact that the data were collected within one clinic. The overall outcome is that this research offers a decent basis of future growth. The future work will be grounded on some methodological extensions that can make the model more effective and strong. These include: expansion of data set, automatic or manual feature extraction, systematic cross validation or cross validation repetition in order to ensure valid generalization. In addition to that, interpretable methods of artificial intelligence such as SHAP will be introduced to improve the interpretability of the model, and make predictive outcomes driven by clinical meaningful trends and not by accidental correlations.

6. CONCLUSION AND FUTURE WORK

This study explored the application of supervised machine learning models for heart disease prediction using a clinical dataset of approximately 1,100 patient records collected at AMSHRC, Vijayapur, Karnataka, India. A powerful preprocessing pipeline, such as missing values imputation, outlier treatment, feature encoding, and feature normalization, was applied to provide accurate measurement and analysis. It was found that performance evaluation proved that the most accurate and reliable predictions were provided by Random Forest and XGBoost, and that they yielded the highest accuracy (0.93), F1-score (0.90), Matthews Correlation Coefficient (0.84), and ROC-AUC values, and the strongest discriminative

Received: August 09 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

capability was provided by the Random Forest (0.98). The stability of proposed models was also ascertained by stratified 10-fold cross-validation. Although the results indicate that ensemble learning can be used in clinical measurement and decision support, the size of the existing dataset is a limitation. Future studies will focus on the expansion of the dataset to about 2,000 records of patients, automatic or manual feature extraction, optimizing hyperparameter and incorporation of explainable artificial intelligence systems to increase transparency in measurements and clinical dependability.

Acknowledgement

The authors gratefully acknowledge the academic support provided by the Department of Computer Science, Government First Grade College, Bagalkot, and the Department of Computer Science, Karnatak State Women's University, Vijayapura, Karnataka, India. The authors also thank the staff of Ayush Multi Speciality Hospital and Research Centre (AMSHRC), Vijayapura, Karnataka, India, for their clinical support, with special thanks to Dr. Rashmi Biradar (AMSHRC) for her guidance during data collection and processing. The authors further acknowledge Dr. Laxmi Kattimani, College Librarian, Government First Grade College, Bagalkot, for assistance with plagiarism checking.

7. REFERENCES

[1.] World Health Organization, "Cardiovascular diseases (CVDs)," *WHO Fact Sheets*, Jul. 31, 2025.

[2.] M. Di Cesare *et al*., "The heart of the world," *Global Heart*, vol. 19, no. 1, Art. no. 11, 2024, doi: 10.5334/gh.1310.

[3.] A. R. Vijayaraj and S. Pasupathi, "Nature inspired optimization in context-aware-based coronary artery disease prediction: A novel hybrid Harris Hawks approach," *IEEE Access*, vol. 12, pp. 92635–92651, 2024, doi: 10.1109/ACCESS.2024.3414662.

[4.] The Devastator, "Predicting heart disease risk using clinical variables," Kaggle Dataset, 2025. [Online]. Available: https://www.kaggle.com/datasets/thedevastator/predicting-heart-disease-risk-using-clinical-var. Accessed: 2025.

[5.] A. Mahajan *et al*., "A hybrid feature selection and ensemble stacked learning model on multivariant cardiovascular disease datasets," *IEEE Access*, vol. 12, pp. 87023–87038, 2024, doi: 10.1109/ACCESS.2024.3412077.

[6.] A. Janosi *et al*., "Heart disease dataset," *UCI Machine Learning Repository*, Univ. of California, Irvine, 1988. [Online]. Available: https://archive.ics.uci.edu/dataset/45/heart+disease.

[7.] P. Ram *et al*., "Investigations on cardiovascular diseases using machine learning algorithms," *Cogent Engineering*, vol. 11, no. 1, Art. no. 2386381, 2024, doi: 10.1080/23311916.2024.2386381.

[8.] D. Cenitta *et al*., "Ischemic heart disease prognosis: A hybrid residual attention-enhanced LSTM model," *IEEE Access*, vol. 13, pp. 4281–4289, 2024, doi: 10.1109/ACCESS.2024.3524604.

Received: August 09 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[9.] H. Kamal *et al*., "Heart disease prediction using machine learning," in *Proc. Int. Conf. Data Intelligence, Communication and Artificial Intelligence Engineering Innovations (IDICAIEI)*, 2024. 2024 IEEE | DOI: 10.1109/IDICAIEI61867.2024.10842908

[10.] C. Gnanavelu *et al*., "Cardiovascular disease prediction using machine learning performance metrics," *Journal of Young Pharmacists*, vol. 17, no. 1, pp. 226–233, 2025 doi: 10.5530/jyp.20251231

[11.] S. Saha *et al*., "Heart disease prediction using machine learning algorithms: Performance analysis," in *Proc. Int. Conf. Advancement in Electrical and Electronic Engineering (ICAEEE)*, pp. 1–6, 2024.

[12.] N. Chandrasekhar and S. Peddakrishna, "Enhancing heart disease prediction accuracy using machine learning techniques," *Processes*, vol. 11, no. 4, Art. no. 1210, 2023, doi: 10.3390/pr11041210.

[13.] N. Biswas *et al*., "Machine learning-based model to predict heart disease in early stage employing different feature selection techniques *",BioMed Research International*, Art. no. 6864343, pp. 1–14, 2023, doi: 10.1155/2023/6864343.

[14.] .P. Jawalkar *et al*., "Early prediction of heart diseasewith data analysis using supervised learning
with stochastic gradient boosting," *Journal of Engineering and Applied Science*, vol. 70, Art. no. 280, 2023, doi: 10.1186/s44147-023-00280-y.

[15.] M. Kavitha *et al*., "Heart disease prediction using a hybrid machine learning model," in *Proc. IEEE 6th Int. Conf. Inventive Computation Technologies (ICICT)*, Coimbatore, India, Jan. 2021, pp. 1329–1333, doi: 10.1109/ICICT50816.2021.9358597.

[16.] A. Ahdal *et al*., "Integrated machine learning technique for accurate heart disease prediction," in *Proc. MECON*, pp. 1–5, 2022. doi: 10.1109/MECON53876.2022.9752342

[17.] Bouqentar, M. A., et.al "Early heart disease prediction using feature engineering and machine learning algorithms". *Heliyon*,2024 *10*(19).

[18.] Rehman, M. U., et,al "Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment". *Scientific Reports*, 2025*15*(1), 13361..

[19.] W. Sun, P. Zhang, Z. Wang, and D. Li, "Prediction of cardiovascular diseases based on machine learning," *ASP Transactions on Internet of Things*, vol. 1, no. 1, pp. 30–35, 2021.

[20.] E. Dritsas and M. Trigka, "Application of Deep Learning for Heart Attack Prediction with Explainable Artificial Intelligence" *Computers*, vol. 13, no. 10, Art. no. 244, 2024, doi: 10.3390/computers13100244.

[21.] UCI Repository, "Statlog (Heart) Dataset Documentation," Univ. of California, Irvine.Available: https://archive.ics.uci.edu/dataset/54/statlog+heart

[22.] Babu *et al*., "Prediction and diagnosis of cardiovascular disease using cloud and machine learning design," *Journal of Cloud Computing*, vol. 13, no. 1, Art. no. 45, 2024, https://doi.org/10.1186/s13677-024-00720-x

[23.] Shinde, P et.al, "A survey on machine learning techniques for heart disease prediction," *SN Computer Science*, *6*(4), 334., 2025, doi: https://doi.org/10.1007/s42979-025-03860-2

Received: August 09 ,2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[24.] A. Abdellatif *et al*., "An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Method," *IEEE Access*, vol. 10, pp. 42568–42582, 2022, doi: 10.1109/ACCESS.2022.3191669

[25.] L. Fitriyani *et al*., "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," *IEEE Access*, vol. 8, pp. 133034 - 133050, 2020, doi: https://doi.org/10.1109/ACCESS.2020.3010511

[26.] A. T. Ashika *et al*., "Enhancing heart disease prediction with stacked ensemble and MCDM-based ranking: an optimized RST-ML approach," *Frontiers in Digital Health*, Art. no. 1609308, 2025, doi: 10.3389/fdgth.2025.1609308.

[27.] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[28.] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[29.] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[30.] M. Alshraideh *et al.*, "Enhancing heart attack prediction with machine learning: A study at Jordan University Hospital," *Applied Computational Intelligence and Soft Computing*, vol. 2024, Art. no. 5080332, 2024, doi: 10.1155/2024/5080332.

Received: August 09 ,2025