ISSN: 2093-5374 / E-ISSN: 2093-5382

Predictive Analytics for Heart Disease Classification Using KNN, SVM, and XGBoost Models

Sandhyarani[1], Dr.B. Nethravathi[2], Syed Abdul Hafeez[3*], Dr.S. Sunil Kumar [4], and Dr. Vikram Patil[5]

[1]Research Scholar, Department of Computer Science and Engineering, JSS Academy of Technical Education, Bengaluru, Karnataka, India. apsandhya21@gmail.com, https://orcid.org/0009-0007-2086-8471

[2]Department of Information Science and Engineering, JSS Academy of Technical Education, Bengaluru, Karnataka, India. nethravathi.sai@gmail.com, https://orcid.org/0000-0003-0340-7888

[3*]Department of Cardiology, JSS Medical College, JSS Academy of Higher Education & Research, Mysuru, Karnataka, India. syedhafeez33@gmail.com, https://orcid.org/0009-0004-4799-2438

[4]Department of Cardiology, JSS Medical College, JSS Academy of Higher Education & Research, Mysuru, Karnataka, India. sunil_cardio@yahoo.co.in, https://orcid.org/0000-0001-9222-9627

[5]Department of Radiology, JSS Medical College, JSS Academy of Higher Education & Research, Mysuru, Karnataka, India. vikrampatil@jssuni.edu.in, https://orcid.org/0000-0002-7684-7860

Received: June 30, 2026; Revised: August 14, 2026; Accepted: September 02, 2026; Published: September 30, 2026

Abstract

Cardiovascular diseases (CVDs) and sudden cardiac events still represent a major challenge world-wide, thus early detection of the risk is very relevant. Early identification of people who may be vulnerable can improve treatment outcomes and reduce mortality significantly. This work presents a comparative machine learning (ML) method for predicting heart disease using three different algorithms namely KNN, SVM, and XGBoost. To ensure reliable data and a consistent work, a well-established Kaggle heart disease dataset is used. In addition, recent data and statistical records were collected from JSS Academy of Higher Education and Research (JSS AHER) and JSS Hospitals through authorized data sharing agreements. Evaluated each model's performance using different metrics. The results show that the three models could predict the heart disease cases with a reasonable accuracy. Among them, the KNN gets the best performance of 90.16% accuracy, subsequent to SVM with 86.89% and XGBoost with 81.97%. The superior performance of KNN can be due to the fact that it can group together similar patient profiles by proximity in the feature space. Overall, the results of the ML models enable to assist medical diagnosis and improve early detection strategies. Future initiatives will aim to scrutinize the recommended approach over extensive and diverse datasets, including model interpretability techniques, and investigate hybrid methods to enhance prediction accuracy and practical applicability in real-world applications.

Keywords: Heart Disease, Classification Models, Early Diagnosis, KNN, SVM, XGBoost.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 3 (September-2026), pp. 1756-1772. DOI: *10.58346/JOWUA.2026.I3.102*

*Corresponding author: Department of Cardiology, JSS Medical College, JSS Academy of Higher Education & Research, Mysuru, Karnataka, India.

1 Introduction

CVDs continue to be among the most prevalent and deadly health issues in the world, and they are difficult to resolve because they are closely related to various clinical risk factors (Alzboon & Al-Batah, 2026). In traditional diagnosis of heart-related diseases, the interpretation of medical parameters by experts is the basis for the diagnosis, and this process may be time-consuming and labor-intensive, and might depend on clinical judgment which varies between experts (Naser et al., 2024). As more structured clinical datasets are becoming available, the development of intelligent prediction systems that are able to uncover hidden associations between patient attributes and aid early disease identification is becoming possible. In this direction, the use of ML approaches has received great importance due to its capability of learning patterns from medical records from past and offering automated classification of CVDs.

Recent developments in medical analytics have shown how the use of ML models can be used to process medical attributes and other physiological signs for forecasting diseases. Compared to traditional statistical methods, ML methods are able to model a nonlinear relationship between multiple risks and enhance prediction capability. But the effectiveness of these models is critically dependent on the proper data preprocessing, feature selection, model optimization, and statistical validation. Thus, it is crucial to develop a reliable framework based on ML to assess the various learning strategies in terms of consistent evaluation criteria.

The present work depicts an ML based heart disease prediction framework that includes models of K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost). The study aims to create a Data-driven classification system using a publicly available clinical data set and evaluate the performance of various ML algorithms. The Aim is to determine which model gives better predictive ability and still performs well statistically.

This work has achieved the following major contributions:

1 An ML algorithm-based CVD prediction system is designed to automatically support clinical decisions with the KNN, SVM, and XGBoost classifiers.
2 A systematic experimental evaluation is carried out with utilization of a clinical data set after suitable pre-processing, training and testing strategies.
3 It includes multiple statistical validation measures such as accuracy, precision, recall, F1 score, ROC-AUC, confusion matrix and cross validation to get a holistic model validation.
4 The developed models are compared and KNN achieves the highest accuracy of 90.16% while SVM achieved 86.89% and XGBoost 81.97%.

The paper is structured as follows: Section I introduces problem domain and the motivation for developing ML-based prediction models for CVD. Section II reviews the literature with a focus on recent advances, methodologies, and research gaps in cardiovascular prediction systems. Section III details the methodology, mathematical modeling, and implementation of KNN, SVM, XGBoost, and the hybrid framework. Section IV presents the results analysis. Section V concludes the study.

2 Literature Review

Recently, heart disease prediction has improved greatly owing to the development of structured ML libraries. These were the early works that looked into the significance of early detection of heart disease using clinical datasets. In addition, these early works were mainly based on classical supervised learning algorithms, such as DT, SVM, LR, and KNN, by considering benchmark data sets. In addition, studies demonstrated that the implementation of ML techniques can efficiently uncover concealed patterns in

patient data, and these techniques can aid in clinical diagnosis. As an example, earlier in comparative studies, SVM and RF achieved good predictive accuracy, often above 90% accuracy, which established the trust of researchers in ML as a diagnostic tool (Singh et al., 2024; Kamal et al., 2024).

The subsequent research incorporated systematic review and analytical studies, which try to analyse various research works together to find the trend in the performance of these models. According to these reviews, DT, NB, RF, and (ANN) are the most used algorithms, with an accuracy range usually between 88 and 95%. In addition, interpretability was mainly the reason for the use of decision tree-based algorithms, while ensemble methods exhibit better robustness and generalization (Budihal & Kawale, 2025; Mao et al., 2025).

More recent developments included large comparative modeling frameworks comparing different algorithms under different pre-processing conditions. The studies used advanced scaling techniques, feature normalization, and cross-validation strategies to increase the reliability of the predictions by considering a large set of classifiers, including KNN, SVM, Gradient Boosting, Light-GBM, and ANN. The results demonstrate that pre-processing and feature engineering have an important influence on the model performance. These works focused on statistical validation using metrics such as AUC, Cohen's Kappa, and log-loss for more reliable and unbiased model comparisons (Alsabhan & Alfadhly, 2025).

Recently, the research has progressed towards ensemble models and hybrid models for classification, and their accuracy is nearly greater than 95%. Furthermore, the integration of ML models such as RF with sequential minimal optimization and Decision Tree has shown remarkable improvement in diagnostic accuracy, reaching up to 97% with high ROC values (Saedi et al., 2025). Further, advanced ensemble architecture such as Multi-Perspective Machine Learning (MPML) shows improvement in both prediction and accuracy with transparency in decision making by using structured feature grouping and interpretability mechanisms as reported in (Miller et al., 2026). The architecture and structure overcame the limitations of standalone ML models for any type of nonlinear data.

Simultaneously, the optimized ML has developed and shown a remarkable performance in achieving high prediction accuracy. Optimization with an ML approach is a profound way to solve the complex problem. The most popular optimization algorithms used in various applications are DE and PSO (Rehman et al., 2025). Hence, hybridized with these optimization algorithms shown greater achievement in terms of accuracy, as 96.5%, outperforming the traditional ML approaches (Yi et al., 2024). Various other hybridized frameworks with optimization demonstrated high predictive accuracy around 97%, with enhanced feature selection. Apart from these, Deep Learning approaches enable the model to give further insights into the input variables, thereby balancing the accuracy and data security, achieving improved prediction capability (Hassan et al., 2024).

Recent literature also discusses emerging trends such as explainable AI and quantum ML. Explainable models have improved interpretability using SHAP and feature selection strategies while maintaining high accuracy, with some models even achieving more than 97% accuracy (Orabe et al., 2026). Furthermore, comparative studies between classical and quantum models demonstrate the ability of quantum neural networks to outperform classical approaches when dealing with large-scale and complex datasets; however, scalability and implementation issues (El-Sofany et al., 2024). The literature is otherwise illustrated in the comparison form as represented in table 1.

Table 1: Literature review based on findings, models, and dataset

Literature	Data used	Model	Findings
(Singh et al., 2024)	UCI Heart Dataset	DT, RF, SVM, PCA ML models	RF and SVM showed strong performance; ML improves early diagnosis improves early detection and healthcare efficiency
(Kamal et al., 2024)	UCI dataset	KNN, DT, SVM, LR	Overall accuracy ~95.7%
(Budihal & Kawale, 2025)	Literature survey	SVM, RF, GB, LR	Accuracy range 88–95% across studies
(Mao et al., 2025)	Multi-database review	DT, LR, NB, RF, ANN	DT most commonly used; accuracy range 88–95%
(Alsabhan & Alfadhly, 2025)	Tabular clinical dataset	KNN, SVM, GBM, ANN, LGBM	Performance depends on preprocessing and scaling methods
(Saedi et al., 2025)	Clinical dataset	RF-SMO, J48-LR	RF-SMO achieved ~97% accuracy
(Miller et al., 2026)	Heart dataset	MPML Ensemble	Improved accuracy with interpretability enhancement
(Yi et al., 2024)	Standard ML dataset	RF, XGBoost, Transformer (PSO)	PSO-Transformer achieved 96.5% accuracy
(Rehman et al., 2025)	NHA NES dataset	PSO + ANN hybrid	Achieved ~97% accuracy with improved feature selection
(Orabe et al., 2026)	ECG + hospital datasets	ResNet, Transformer, LR	AUC ≈ 0.87–0.92; privacy-aware ML reduces performance slightly
(El-Sofany et al., 2024)	Public + private datasets	XGBoost, SVM, RF	XGBoost achieved ~97.57% accuracy
(Gul et al., 2026)	Multiple datasets	Ensemble ML models	Ensemble models improve robustness and generalization
(Eltawil et al., 2026)	CDC dataset	KNN	Extremely high accuracy (~99%) flagged due to data leakage issue
(Ramanathan V et al., 2024)	Clinical dataset	LR, RF, DT, SVM, NN	SVM achieved highest accuracy (~91.5%)
(Ravi et al., 2025)	Clinical dataset	Multiple ML models	ML improves early detection and healthcare efficiency
(Bao, 2025)	Heart dataset	LR, RF	Both models showed strong classification performance
(Subramani et al., 2023)	Combined dataset	ML + DL hybrid	Achieved ~96% accuracy using hybrid approach
(Wan et al., 2025)	Multi-source datasets	ML frameworks	ML improves diagnostic accuracy vs traditional methods
(Banerjee & Paçal, 2025)	Large-scale review (65 studies)	Ensemble, DL	Ensemble models dominate structured data prediction
(Kumar et al., 2025)	Combined datasets (918 samples)	Hybrid SVM + QML	Hybrid models improve robustness and accuracy

The reviewed literature shows that the ML-based prediction of the CVD has progressed from classical classifiers to ensemble, optimization based, explainable and deep learning methods. While more sophisticated models can produce better accuracy, they can be difficult to interpret, compute, and are still complex. The results inspire the creation of a comparative ML framework to assess multiple statistical

measures of the efficient classifiers. In this study, three models are chosen to set up a reliable prediction method: KNN, SVM, and XGBoost, in which the models are selected with a balanced perspective on the accuracy, validation performance, and applicability.

3 Methodology

3.1 CVD through SVM

The clinical risk attributes are used to model the decision boundary between healthy and unhealthy classes by using the SVM classifier. The SVM model is trained in the normalized feature space to find the optimal separating hyperplane that maximizes the margin between the two categories.

A kernel transformation projects the input feature vectors into a higher-dimensional space, allowing the classifier to distinguish patterns that cannot be separated using simple boundary line. In this study, the RBF kernel is used as it can capture the complex relationships among the clinical parameters. The classifier tries to maximize the margin between the support vectors, which are the most informative samples, closest to the decision boundary.

Mathematical Modelling of SVM Framework

Let X and Y be the input and output data set consists of N samples. The objective of SVM is to determine a separating hyperplane such that the margin between both classes is maximized. The optimization problem for the soft-margin SVM is expressed as equations (1), (2), and (3):

$$w^T x + b = 0 \quad (1)$$

$$min_{\omega,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \quad (2)$$

$$y_i(w^T x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \quad (3)$$

Where C is the regularization parameter controlling the penalty for misclassification, and ε_i denotes the slack variables.

For nonlinear datasets, the RBF kernel maps the feature space into a higher-dimensional domain and is defined as equation (4):

$$K(x_i, x_j) = exp\left(-\gamma\|x_i - x_j\|^2\right) \quad (4)$$

Where 'γ' controls the influence of individual samples. The final decision function is expressed as

$$f(x) = sign\left(\sum_{i=1}^{N} \propto_i y_i K(x_i, x) + b\right) \quad (5)$$

Where, in equation (5), α_i are the Lagrange multipliers corresponding to support vectors

Pseudo code for SVM

Input: Clinical dataset D = {X, Y}

Output: Trained SVM classifier and prediction results

1. Normalize feature values using z-score standardization

2. Perform stratified Train-Test partitioning

3. Select kernel function (RBF) and initialize parameters (C, γ)

4. Train an SVM classifier to compute the optimal separating hyperplane

5. Tune hyper parameters using validation accuracy and AUC

6. Evaluate the model on the test dataset:

7. Select the final SVM configuration with the best generalization

The SVM algorithm is applied to the clinical data by first normalizing the input attributes through z-score normalization and partitioning the data into stratified training and testing datasets. The SVM model with an RBF kernel is learned with optimal hyper-parameters (C, γ), resulting in the optimal decision boundary between classes. This model is further analyzed in terms of its performance based on validation metrics like accuracy and AUC, with the best model being chosen for predicting the results of the test set.

3.2 XGBoost for Heart Disease Prediction

An XGBoost classifier is employed to model complex relationships among medical risk determinants for CVD. The XGBoost model is trained on the processed feature space to construct an ensemble of decision trees that collaboratively improve classification accuracy.

XGBoost operates on the principle of gradient-boosted decision trees, where weak learners are added sequentially, and each new tree attempts to correct the prediction errors of the previous trees. Such a bootstrapping procedure allows proposed model to discover high-level feature interactions that are difficult to learn by a standard classifier. Regularization terms are added to the objective function of the algorithm to control the complexity and to avoid overfitting on the data set. Additionally, the XGBoost supports weighted decision learning and subsampling, making it robust against noisy features and small variations in patient records.

Following model training, the classifier outputs a probability score indicating the likelihood of heart-disease risk, which is converted into a binary diagnostic decision. All ML models have the hyper parameters tuned to obtain the best-performing configuration.

The XGBoost model consists of an ensemble of K decision trees, and the prediction for a sample xi is expressed as equation (6):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \;\; f_k \in F \tag{6}$$

Where F represents the set of regression trees utilized by the model. The learnig objective is formulated by integrating a differentiable loss function with a regularization component to handle complexity of the model is shown in equation (7).

$$L = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{7}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{8}$$

Here, in equation (8), T specifies the total number of leaf nodes present in the decision tree, while ‘w’ corresponds to the assigned weights to those leaves. The parameter act as regularization factors that help constrain model complexity and improve generalization performance. The optimization process adds new trees by minimizing the second-order Taylor approximation of the objective using gradient and Hessian statistics, which enables fast and stable convergence. The ultimate class designation is derived by imposing a threshold on the anticipated likelihood.

Pseudo code for XGBoost

Input: Clinical dataset D = {X, Y}.

Output: Trained XGBoost model and prediction results

1. Normalize feature values using z-score standardization

2. Perform stratified Train-Test partitioning

3. Initialize XGBoost parameters:

4. Train a gradient-boosted tree ensemble on training data.

5. Tune hyper parameters using validation accuracy and AUC.

6. Evaluate the model on the test dataset:

7. Select the final XGBoost configuration with the best generalization.

XGBoost algorithm works on the clinical dataset through normalizing the features and splitting the data into training and testing datasets. XGBoost initializes the model with parameters including the learning rate, tree depth, the number of estimators, subsampling, feature subsampling, and regularizers. Then an ensemble of gradient boosted trees is fitted iteratively to reduce the prediction errors. Tuning the hyperparameters is carried out using the validation accuracy and AUC. The XGBoost model with optimal performance is selected for the test dataset.

3.3 KNN for Heart Disease Prediction

The unique feature of KNN is that it serves as a distance-based ML method for any application. In this, heart disease prediction is done based on clinical attributes. The model categories the patients with similar characteristics as a separate class and others as distinct categories.

A stratified train-test split to preserve the proportion of heart disease and non-heart disease samples in both training and testing data. It provides the classifier with diverse samples to learn from and ensures the least bias at prediction time. In the data set, each row corresponds to a patient's medical record. The columns constitute the feature vector consisting of clinical attributes. Given a patient's feature vector, KNN predicts if that patient has heart disease or not by comparing it to already seen data. The technique is simple: calculate the distance of a new unseen patient to all other rows in the training data and "vote" on the majority label among its k neighbors. KNN assigns the new patient to the class that maximizes this count. The distance can be measured using Euclidean distance or any other appropriate metric. The larger the value of k, the more neighbors are considered for voting. There is no training process per se for KNN. It "remembers" the dataset and does computation at prediction time. This is best suited for when you have structured data, as in a clinical dataset where attributes relate to each other meaningfully in the feature space. At the end of running this model with the data set, it is assessed based on accuracy, ROC-AUC, F1 score, confusion matrix, and cross-validation accuracy of proposed final model.

Mathematical Modelling of KNN Framework

The clinical feature vector for a patient is represented as equation (9):

$$X = [x_1, x_2, x_3, \dots x_n] \tag{9}$$

Where 'n' indicates the number of clinical attributes used in the prediction model. Euclidean distance is used to determine the similarity between two patient records X_i and X_j, which is represented via equation. (10)

$$d(X_i, X_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \quad (10)$$

For a given test sample 'X_t', the algorithm identifies the set of 'k' nearest neighbors based on the smallest computed distances as expressed mathematically as equation (11):

$$N_k(X_t) = \{X_{(1)}, X_{(2)}, \dots, X_{(k)}\} \quad (11)$$

The predicted class label is then determined by majority voting among the neighbors as in equation (12):

$$\hat{y} = mode\{y_{(1)}, y_{(2)}, \dots, y_{(k)}\} \quad (12)$$

Equation. (13) determines the probability of heart disease presence to ensure effectiveness of the model, which categorizes whether the patient has the disease or not.

$$P(y = 1|X_t) = \frac{1}{k}\sum_{i=1}^{k} y_i \quad (13)$$

Pseudo Code for KNN Modeling

Input: Clinical dataset $D = \{X, Y\}$

Output: Trained KNN model and prediction scores.

1. Normalize all numerical features using z-score scaling.

2. Apply a stratified Train-Test split to maintain class distribution.

3. Select the optimal number of neighbors (k) using validation.

4. For each testing instance (Xt):

a.Calculate the Euclidean distance between the testing instance and each sample in the training data set.

b. Sort the data in ascending order.

c. Select the nearest neighbors (k).

d. Retrieve the class labels of the selected neighbors.

e. Predict the final class label based on the majority voting principle among the selected neighbors.

5. Assess the model performance via performance metrics.

6. Perform cross-validation to verify model stability.

The KNN algorithm works on the clinical data set by first scaling the numerical values using z-score scaling and then separating the data into stratified training and testing sets. The optimal value of (k) is found through validation, and the distance between the test and training data is computed using the Euclidean distance. Next, the nearest neighbors are determined and the class is found based on majority voting. The performance of the model is measured through various metrics, and cross-validation is used to validate the KNN classifier.

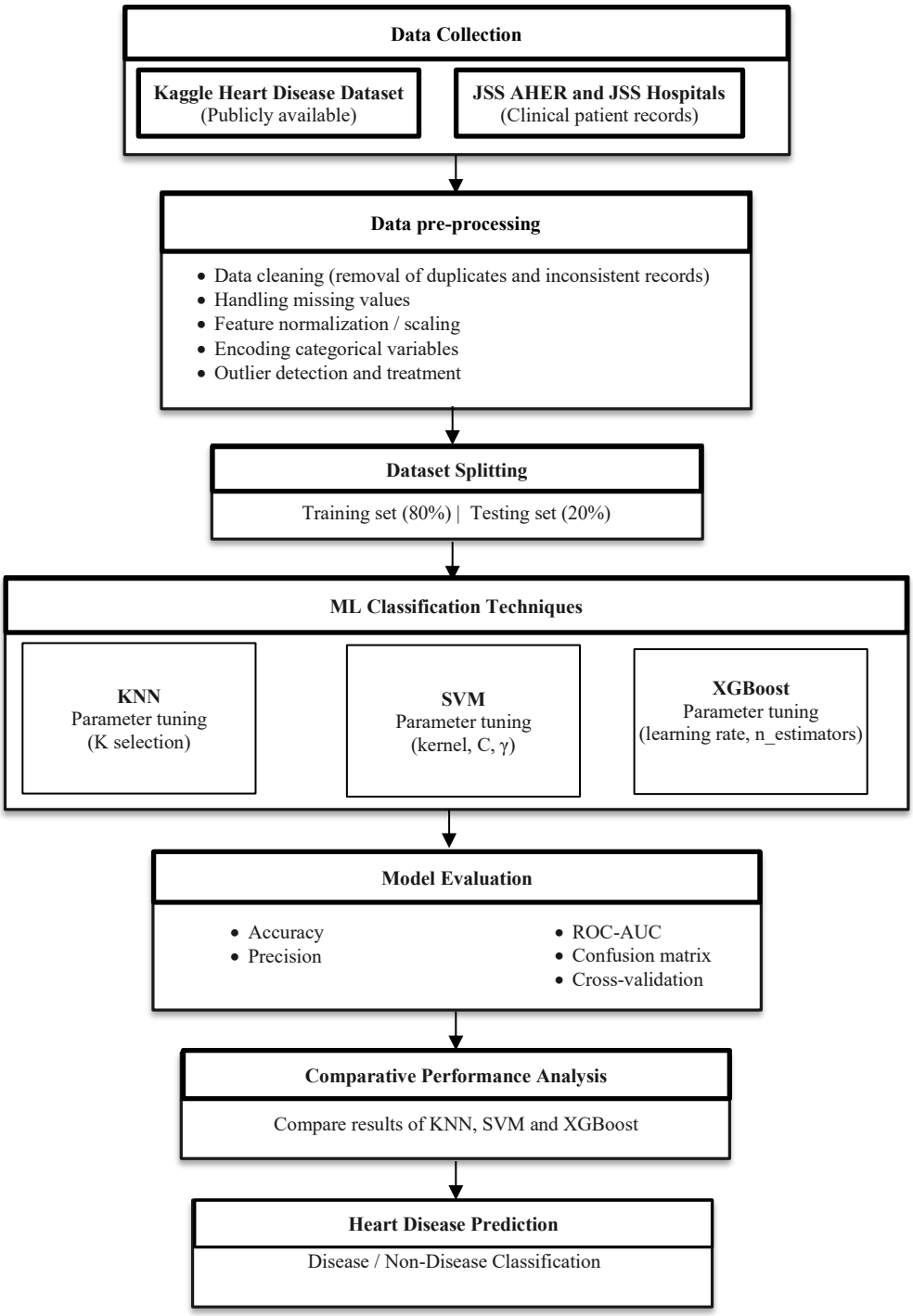


Figure 1: Methodological framework of ML-based heart disease prediction using Knn, Svm, and XGBoost classifiers

Figure 1 shows the entire process flow of the designed heart disease prediction system, consisting of data gathering, Data Preprocessing, Feature Generation, *ML* Model Design, Performance Measurement, And Comparison. The Processed Clinical Data Are Classified Using KNN, SVM, And XGBoost Algorithms, And Their Performance Is Tested Through Various Evaluation Criteria For Precise Disease and Non-Disease Classification.

4 Results and Discussions

The model performance for the heart disease prediction is evaluated in a structured manner, including confusion matrix analysis, statistical metrics, data visualization, and feature importance.

4.1 Experimental Setup and Parameter Initialization

The suggested framework is developed by utilizing Python along with libraries like NumPy, Pandas, Scikit-learn, XGBoost, and Matplotlib. Evaluation of the suggested framework was done using Kaggle Heart Disease dataset (https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset) and clinical data from JSS AHER and JSS Hospitals. This dataset contains various clinical attributes like age, blood pressure, cholesterol, heart rate, ECG features, and chest pain features.

Parameters for each algorithm were set in the following manner: the KNN model had its neighbor value (k = 5) set to the optimal one after the validation process. The SVM algorithm was fitted to use an RBF kernel with C = 1.0 and gamma = scale. XGBoost had its parameters learning rate = 0.1, max_depth = 5, and n_estimators = 100. The optimization was done based on validation accuracy to obtain accurate classification results.

In medical prediction, false negatives are important because it indicate missed cases of the disease, and false positives can lead to unnecessary tests.

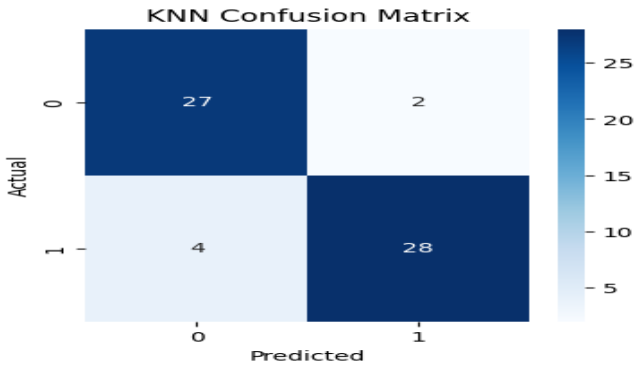


Figure 2: KNN confusion matrix

Figure 2 shows the prediction efficiency of the KNN algorithm with 27 correctly classified class-0 samples and 28 correctly classified class-1 samples. There are very few errors of predictions (2 and 4).

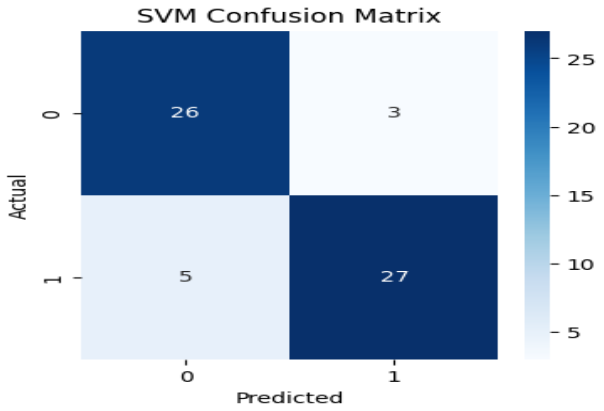


Figure 3: SVM confusion matrix

Figure 3 illustrates the findings of classification accuracy obtained through SVM, where there are 26 correct predictions of class-0 data, 27 correct predictions of class-1 data, and very few mistakes (3 and 5).

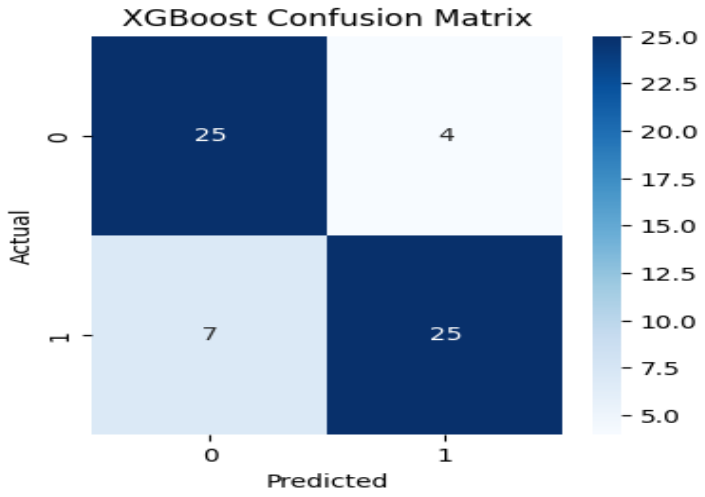


Figure 4: XGBoost confusion matrix

Figure 4 shows the effectiveness of classification using the XGBoost model, as it has successfully classified 25 samples from class 0, 25 samples from class 1, and a small number of samples were misclassified (4 and 7 samples).

An analysis of the models depicted in figure 2-4 reveals that the KNN model excels in classification accuracy. The KNN model's confusion matrix reveals that the majority of cases are accurately categorized by the model. The forecast of 27 adverse and 28 favorable instances is precise and demonstrates robust categorization proficiency. A limited quantity of misclassifications is noted: 2 false positives and 4 false negatives. The occurrence of false negatives indicates that the model does not identify certain patients with heart disease, meaning some individuals with the condition are erroneously categorized as healthy. This is a serious problem in medical applications. But the number is not too high, which means that the model can detect positive cases pretty well. The low number of false positives suggests that the model does not tend to classify healthy people as having heart disease.

All performance metrics can be obtained from the confusion matrix. Equation (14) shows that accuracy is the overall correctness, Equation (15) indicates that precision represents the ratio of accurately predicted cases. Equation (16) illustrates that recall is significant as it assesses the model's capacity to identify people with heart disease. Equation (17) illustrates that the F1-score represents the equilibrium between precision and recall, while AUC indicates the model's capacity to differentiate across classes. The evaluation metrics are determined utilizing the equation from existing literature, while the AUC is computed by library functions, with the resultant data organized in table 2.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

$$Precision = \frac{TP}{TP+FP} \quad (15)$$

$$Recall = \frac{TP}{TP+FN} \quad (16)$$

$$F1-score = \frac{2.(Precision.Recall)}{Precision+Recall} \quad (17)$$

Table 2: Comparative analysis of ML models

S. No	Model	Accuracy	Precision	Recall	F1	AUC
1	KNN	90.16	0.93	0.88	0.90	0.92
2	SVM	86.89	0.90	0.84	0.87	0.93
3	XGBoost	81.97	0.86	0.78	0.82	0.91

Figure 5 illustrates the model comparison in terms of accuracy. Figure 6 illustrates that precision, recall, and F1 score, based on actual values, while table 2 summarizes the same results in values rounded to two decimal places. Figure 7 represents the importance of different clinical features to predict heart disease. Among the different variables, oldpeak, ca, and cp have the strongest influence. Thalach and age also significantly contribute, whereas fbs and restecg have the least influence. This shows the prediction model depends more strongly on certain clinical variables.

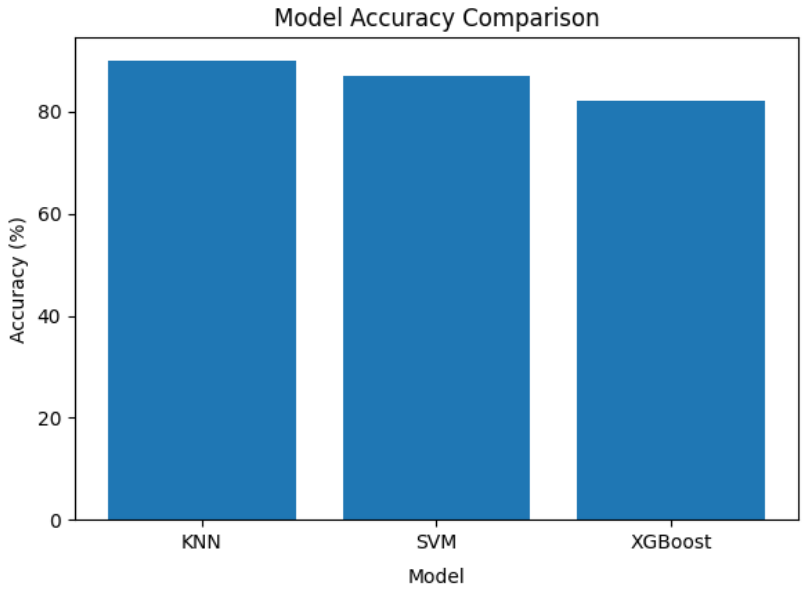


Figure 5: Accuracy comparison of models

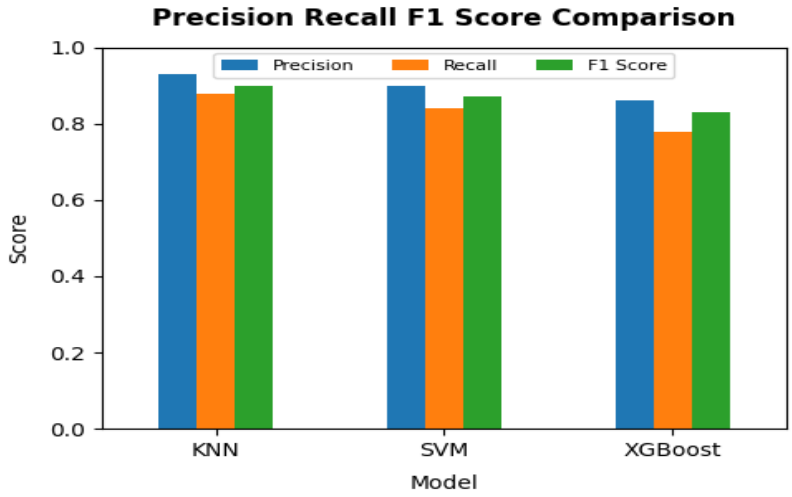


Figure 6: Comparison of statistical measures of models

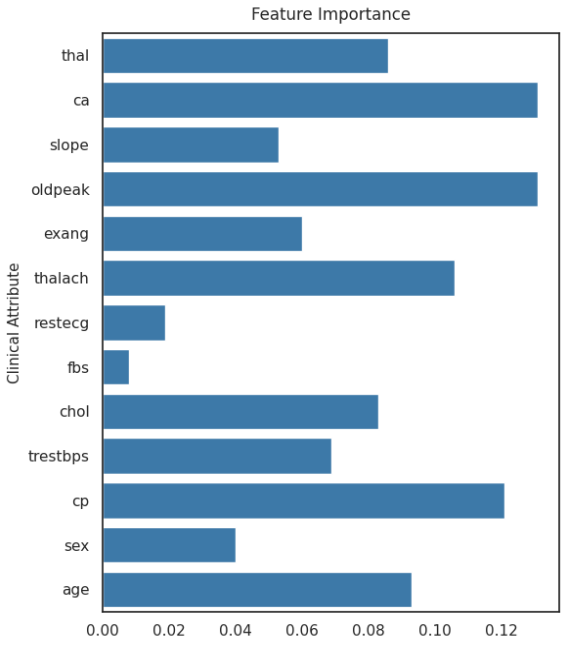


Figure 7: Feature importance plot of the dataset

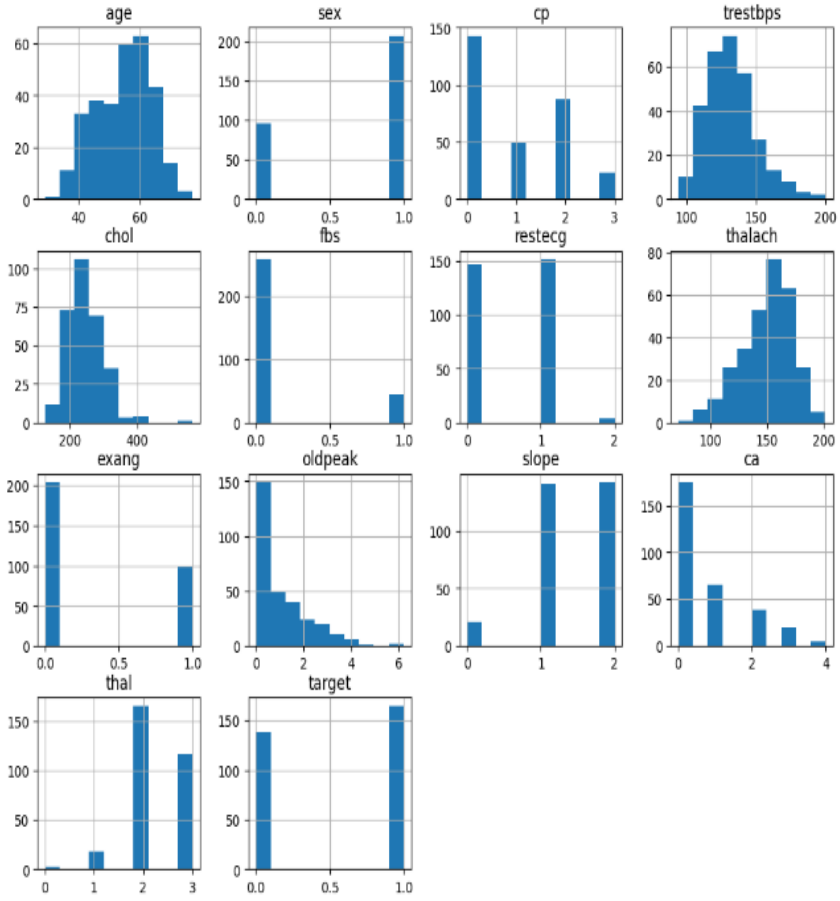


Figure 8: Feature distribution plot of the dataset

Figure 8 depicts feature distribution, illustrating how data varies among different variables. Features such as age, cholesterol, trestbps, and thalach all continue to show different ranges and different patterns. Some of the patterns show slight skewness. Some features, such as oldpeak, show distinct, large outliers. This may lead to problems in model learning. In the categorical variables, such as sex, cp, exang, and target, different class-based separations occur in the data. This shows differing class separations in the dataset.

5 Conclusion

The current research effectively illustrates the efficacy of ML methodologies in forecasting cardiovascular ailments utilizing the Kaggle dataset, as well as data from JSS Academy of Higher Education and Research (JSS AHER) and JSS Hospitals over recent years. A comparative framework incorporating KNN, SVM, and XGBoost was developed and evaluated using a standardized Kaggle dataset and real time data collected. The analysis reveals that all three models are capable of identifying patterns associated with cardiac risk, and their performance varies depending on their learning mechanisms and data handling capabilities. Out of three models, the KNN model produces a high classification accuracy of 90.16% among other models, showcasing similarity-driven relationships among clinical attributes effectively. SVM comes very close to KNN with an accuracy of 86.89% by learning the best decision boundary in higher dimensions. Generally, XGBoost shows excellent results as an ensemble model, but in the case, it lags with an accuracy of 81.97%. This may be because of the dataset size and features. The evaluation of models is based on statistical measures, through which it is concluded that the proposed framework is consistent and effective. Also, on observing the confusion matrix, it can be interpreted that there are fewer misclassifications done by the KNN model. Hence, it is more trustworthy for early-stage diagnosis. But there are still some false negatives that can be further reduced. Overall, It is reasonable to conclude that with the use of ML models, it can assist doctors by providing them with quick predictions based on data. There is future scope for improvement by increasing the size of the dataset by including more features, modelling hybrid models, DL models, and planning for the explainability of proposed model for real-time deployments.

Declaration of Competing Interests

The authors declare no financial or other relationships that are potential conflicts of interest or might be interpreted as influencing the findings of this report.

Acknowledgment

We would like to acknowledge JSS MAHAVIDYAPEETA, Mysure, JSS Academy of Higher Education and Research, Mysuru, JSS Academy of Technical Education, Bangalore, Karnataka, INDIA for all the support provided.

Data availability

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

[1] Alsabhan, W., & Alfadhly, A. (2025). Effectiveness of machine learning models in diagnosis of heart disease: a comparative study. *Scientific Reports*, *15*(1), 24568. https://doi.org/10.1038/s41598-025-09423-y

[2] Alzboon, M. S., & Al-Batah, M. S. (2026). A clinically accessible heart disease prediction framework: multi-model evaluation with ensemble learning and low-code deployment. *Discover Applied Sciences*, *8*(3), 339. https://doi.org/10.1007/s42452-026-08353-2

[3] Aniruth Ramanathan, V., Yerra, S., Swetha, N. G., & Gandikota, P. (2024). Advancements in heart disease prediction: A machine learning approach for early detection and risk assessment. *International Journal of Science and Research (IJSR), 13*(10), 774–780. https://dx.doi.org/10.21275/SR241004171836

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

[5] Bao, Y. (2025). Research and application of heart disease prediction model based on machine learning. In *ITM Web of Conferences* (Vol. 70, p. 04023). EDP Sciences. https://doi.org/10.1051/itmconf/20257004023

[6] Budihal, S., & Kawale, S. R. (2025). A Systematic Review of Machine Learning Models for Cardiac Disease Prediction. *International Journal of Computer Applications*, *186*(63), 27-33. https://doi.org/10.5120/ijca2025924439

[7] El-Sofany, H., Bouallegue, B., & El-Latif, Y. M. A. (2024). A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. *Scientific reports*, *14*(1), 23277. https://doi.org/10.1038/s41598-024-74656-2

[8] Eltawil, M., Byham-Gray, L., Jia, Y., Mistry, N., Parrott, J., & Gohel, S. (2026). Comment on Iacobescu et al. Evaluating Binary Classifiers for Cardiovascular Disease Prediction: Enhancing Early Diagnostic Capabilities. J. Cardiovasc. Dev. Dis. 2024, 11, 396. *Journal of Cardiovascular Development and Disease*, *13*(1), 46. https://doi.org/10.3390/jcdd13010046

[9] Gul, G., Korejo, I. A., Hakro, D. N., Alqahtani, H., Abbasi, A., Babar, M., ... & Ali, N. I. (2026). Machine learning and ensemble methods for cardiovascular disease prediction: A systematic review of approaches, performance trends, and research challenges. *Computers*, *15*(1), 25. https://doi.org/10.3390/computers15010025

[10] Hassan, M. M., Zaman, S., Rahman, M. M., Bairagi, A. K., El-Shafai, W., Rathore, R. S., & Gupta, D. (2024). Efficient prediction of coronary artery disease using machine learning algorithms with feature selection techniques. *Computers and Electrical Engineering*, *115*, 109130. https://doi.org/10.1016/j.compeleceng.2024.109130

[11] Kamal, H., Bakhtawar, Z., Hussain, M. Z., Hasan, M. Z., Mustafa, M., Atif, M., & Yaqoob, M. (2024). Heart disease prediction using machine learning. *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)*, 1–6. https://doi.org/10.1109/IDICAIEI61867.2024.10842908

[12] Kumar, A., Dhanka, S., Sharma, A., Bansal, R., Fahlevi, M., Rabby, F., & Aljuaid, M. (2025). A hybrid framework for heart disease prediction using classical and quantum-inspired machine learning techniques. *Scientific reports*, *15*(1), 25040. https://doi.org/10.1038/s41598-025-09957-1

[13] Mao, Y., Jimma, B. L., & Mihretie, T. B. (2025). Machine learning algorithms for heart disease diagnosis: A systematic review. *Current problems in cardiology*, *50*(8), 103082. https://doi.org/10.1016/j.cpcardiol.2025.103082

[14] Miller, S. T., Logan, K. A., Anderson, R., Cowell, P. E., Busby-Earle, C., & Morris, L. D. (2026). Multi-Perspective Machine Learning MPML: A High-Performance and Interpretable Ensemble Method for Heart Disease Prediction. *Machine Learning with Applications*, 100836. https://doi.org/10.1016/j.mlwa.2026.100836

[15] Naser, M. A., Majeed, A. A., Alsabah, M., Al-Shaikhli, T. R., & Kaky, K. M. (2024). A review of machine learning's role in cardiovascular disease prediction: recent advances and future challenges. *Algorithms*, *17*(2), 78. https://doi.org/10.3390/a17020078

[16] Orabe, Z., Vasankari, A., Pahikkala, T., Kaisti, M., & Airola, A. (2026). Securing deep learning models with differential privacy for cardiovascular disease prediction. *Biomedical Signal Processing and Control*, *112*, 108502. https://doi.org/10.1016/j.bspc.2025.108502

[17] Ravi, C. S., Mamidala, J. V., Kurma, J., Attipalli, A., & Chundru, S. K. (2025). Early heart diseases prediction using advanced machine learning algorithms. *Journal of Contemporary Education Theory & Artificial Intelligence*, *4*(1), 1–9. https://doi.org/10.47991/2996-4954/JCETAI-120

[18] Rehman, M. U., Naseem, S., Butt, A. U. R., Mahmood, T., Khan, A. R., Khan, I., ... & Jung, Y. (2025). Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment. *Scientific Reports*, *15*(1), 13361. https://doi.org/10.1038/s41598-025-96437-1

[19] Saedi, A. Y. F., Hussein, R. T., & Ghrairi, D. A. M. (2025). Enhancing Heart Disease Diagnosis with Hybrid Machine Learning Models: A Case Study. *Engineering, Technology & Applied Science Research*, *15*(5), 27194-27199. https://doi.org/10.48084/etasr.12260

[20] Singh, A., Mahapatra, H., Biswal, A. K., Mahapatra, M., Singh, D., & Samantaray, M. (2024). Heart disease detection using machine learning models. *Procedia Computer Science*, *235*, 937-947. https://doi.org/10.1016/j.procs.2024.04.089

[21] Subramani, S., Varshney, N., Anand, M. V., Soudagar, M. E. M., Al-Keridis, L. A., Upadhyay, T. K., ... & Rohini, K. (2023). Cardiovascular diseases prediction by machine learning incorporation with deep learning. *Frontiers in medicine*, *10*, 1150933. https://doi.org/10.3389/fmed.2023.1150933

[22] Wan, S., Wan, F., & Dai, X. J. (2025). Machine learning approaches for cardiovascular disease prediction: A review. *Archives of cardiovascular diseases*, *118*(10), 554–562. https://doi.org/10.1016/j.acvd.2025.04.055

[23] Yi, J., Yu, P., Huang, T., & Xu, Z. (2024, December). Optimization of Transformer heart disease prediction model based on particle swarm optimization algorithm. In *2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC)* (pp. 1109-1113). IEEE. https://doi.org/10.1109/ICFTIC64248.2024.10913096

Authors Biography



Sandhyarani is Research Scholar in the Department of Computer Science and Engineering at JSS Academy of Technical Education Bangalore. She obtained her Bachelor of Engineering (B.Tech.) degree in 2016 and Master of Technology (M.Tech.) degree in 2018 from Visvesvaraya Technological University (VTU), Belagavi, Karnataka, India. Her research interests include Machine Learning, Deep Learning, Data Pre-processing, Artificial Intelligence, and Data Analytics, with a particular focus on developing intelligent and data-driven solutions for real-world applications. Actively engaged in research and academic activities, she regularly participates in national and international conferences, workshops, faculty development programs, and seminars to stay abreast of the latest advancements in emerging technologies. Her commitment to continuous learning and scholarly excellence contributes to her growing expertise in advanced computational methods and intelligent systems.



Dr.B. Nethravathi is currently working as Associate Professor, Department of Information Science and Engineering in JSS Academy of Technical Education Bangalore. She received B.E degree in Computer Science and Engineering and M.Tech. degree in Computer Network Engineering from Visvesvara Technological University Belagavi, India, in 2003 and 2010 respectively. She received Ph.D degree from VTU Belagavi in 2019. Six Research scholars working under her. She is IEEE senior member, Life member of ISTE and IAENG. Her research interests include 5G networks, IoT, Cloud Computing Artificial Intelligence and Machine Learning. She attended many National Conferences and International Conferences and received 2 Best Paper awards. She has published more than 45 International Journal papers. She has filed and published 2 Indian Patents. She reviewed for many National and International conferences.



Syed Abdul Hafeez, Sr. Cardiac Sonographer and PhD scholar, JSS Academy of higher education and research - Mysore - Karnataka - India. Seving as a Sr. technologist in the department of non-invasive cardiology. As a research scholar actively participated in all research forums. Expertise in both paediatric and adult echocardiography. Member of Indian academy of echocardiography and also actively involve in teaching and clinical services.



Dr.S. **Sunil Kumar** is a Professor and Head of the Department of Cardiology at JSS Medical College, JSS Academy of Higher Education and Research (JSS AHER), Mysuru, India. He is an accomplished interventional cardiologist with expertise in coronary interventions and adult and paediatric structural heart interventions. He is actively involved in undergraduate and postgraduate teaching and training and serves as a PhD guide. He is also actively engaged in clinical research and has contributed numerous publications to national and international peer-reviewed journals.



Dr. Vikram Patil is the Vice Principal of JSS Medical College and Professor of Radio-Diagnosis at JSS Academy of Higher Education and Research (JSS AHER), Mysuru, India. Serving as a key clinical and administrative leader, his work focuses on institutional growth, internationalization strategies, and bridging the gap between clinical research and commercial healthcare applications. He is actively champions the breaking of traditional academic silos. He facilitates high-impact, multi-disciplinary networks connecting clinicians, engineers, and data scientists to validate and implement AI and digital health-tech solutions for complex diagnostic challenges. His research interests lie at the intersection of translational medicine, future healthcare delivery, and indigenous medical device innovation, supported by a registered Indian and international patent for non-invasive medical ablation systems. Dedicated to fostering robust industry-academia linkages, He is actively leads collaborative.