Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 104-114. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2753

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

MACHINE LEARNING FOR CORONARY HEART DISEASE RISK PREDICTION: A CONVENTIONAL AND ENSEMBLE CLASSIFICATION MODEL APPROACH

Ch V Raghavendran*[1], Kathi Chandra Mouli[2], S. Bhargavi Latha[3]

[1]Department of IT, Aditya University, Surampalem, AP, India.
[2]Department of CSE, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Bachupally, 500090, India.
[3]Associate Professor, School of CSE, Reva University, Yelahanka, Karnataka 560064, India.

[1]https://orcid.org/0000-0002-4423-9086, [2]https://orcid.org/0009-0001-0889-9263, [3]https://orcid.org/0000-0002-0102-3704

Email: *raghuchv@yahoo.com, kcm7961.1852@gmail.com, s.bhargavilatha@gmail.com

ARTICLE INFO

Article History
Received: September 28, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Coronary Heart Disease,
Classification,
Ensemble techniques,
Class imbalance,
Boosting,
Bagging,
Voting,
Stacking.

ABSTRACT

Coronary Heart Disease (CHD) is the leading cause of illness and mortality worldwide. This is driven by risk factors such as high cholesterol, hypertension, smoking, diabetes, obesity and genetic predisposition. Early detection and management of CHD are critical and involve lifestyle changes, pharmacological interventions, angioplasty, or bypass grafting. Traditional diagnostic methods require effective and timely medical interventions or the early detection of CHD. The proposed work suggests a stacking classifier for early prediction of CHD. Conventional classification models are typically simple, interpretable and efficient, making them suitable for classification problems. The performance of these models is limited when dealing with complex or high-dimensional data. Ensemble models address this limitation by combining the predictions of several models to improve accuracy. Bagging, Voting, and stacking classifiers are more sophisticated ensemble methods that use this concept further by using a metamodel to aggregate the predictions of multiple base models. In this study, several conventional models, such as the SGD Classifier, Logistic Regression, K-Nearest Neighbor, Naive Bayes, Decision Tree, and Random Forest Classifiers are applied to the dataset to predict whether the patient has 10 year risk of Coronary Heart Disease (CHD). Boosting models such as Gradient Boosting, AdaBoost, and XGBoost are also applied. The proposed voting classifier resulted in better accuracy, precision, recall, and other metrics compared with the conventional, boosting, and bagging models.

I. INTRODUCTION

Coronary Heart Disease (CHD), also known as ischemic heart disease, is one of the most common cardiovascular diseases and a major cause of death worldwide. Atherosclerosis, a disorder marked by the build-up of fatty deposits, cholesterol, and other materials in the coronary arteries, is the main cause of CHD, which is a condition that impairs blood flow to the heart muscle. The myocardium receives less oxygen when these arteries gradually narrow or get blocked, it may cause myocardial infarction, angina pectoris, and sudden cardiac death, among other clinical signs. [1]. With millions of deaths annually, CHD continues to be a major global public health concern. The World Health Organization (WHO) has reported that 32% of fatalities worldwide are attributable to cardiovascular illnesses, including CHD. Low- and middle-income countries have disproportionately high illness burdens because they have less access to healthcare and preventive measures. Globally, the incidence of CHD is on the rise owing to aging populations and the rising prevalence of risk factors including diabetes and obesity. The persistent inflammatory condition of atherosclerosis inside arterial walls functions as the main cause of CHD. Chemicals from LDL cholesterol enter the endothelium to form deposits inside artery walls during the first stage because of endothelial dysfunction. An immunological reaction gets launched after oxidized LDL draws monocytes and macrophages which later turn into foam cells that help to form fatty streaks [2]. Artery lumen becomes narrower because fibrous plaques form out of previously developed streaks and reduce blood flow. The formation of a thrombus after plaque rupture leads to total artery blockage thus causing heart attack or other acute coronary syndromes [3]. The development of CHD depends on both risk factors that people can change and those that cannot be changed. Conflicting variables that people can change include smoking alongside high blood pressure and high cholesterol together with diabetes and obesity while a sedentary lifestyle and poor eating choices complete the list [4]. Two important non-controllable CHD risk factors involve sex-based differences (men face greater risk) and the existence of cardiovascular disease in family history. The development of CHD depends on several newly emerging risk variables that include stress factors together with persistent inflammatory responses and particular gene sequences. The risk of CHD can be considerably decreased by effectively managing modifiable risk factors. Asymptomatic phases to severe, life-threatening illnesses are only a few of the many clinical presentations of congenital heart diseases. Palpitations, exhaustion, shortness of breath, and chest pain or discomfort (angina) are typical symptoms [5]. Stable angina is typified by consistent chest discomfort caused by emotional or physical stress, whereas unstable angina occurs suddenly and is related to a greater risk of myocardial infarction. Prolonged chest discomfort, nausea, sweating, and, in extreme situations, unconsciousness are typical symptoms of myocardial infarction, sometimes known as a heart attack.

Imaging methods, laboratory testing, and clinical evaluation are all used for the diagnosis of CHD. While biomarkers, such as troponin levels, aid in confirming myocardial injury, electrocardiography (ECG) and stress tests are frequently used to evaluate heart function. Imaging techniques that show the amount of blockage and provide thorough visualization of the coronary arteries include echocardiography, computed tomography (CT) angiography, and coronary angiography. Non-invasive techniques for risk assessment and early subclinical disease diagnosis are provided by emerging technologies, such as coronary artery calcium scoring. Medical treatment of CHD requires surgical interventions combined with drugs along with changes to individual lifestyles. Sustaining healthy changes in lifestyle including stopping smoking while starting physical exercise and maintaining proper weight and eating well remain fundamental for slowing disease advancement. Patients with CHD need beta-blockers and statins in combination with ACE inhibitors as well as antiplatelet medications to manage their symptoms while reducing health risks [6]. The medical procedure of both CABG and PCI serves to rebuild cardiovascular blood flow in conditions where substantial artery obstructions occur [7]. Current developments in personalized medicine united with risk-prediction tools have boosted the accuracy of managing CHD.

Primary prevention strategies aim at managing risk factors throughout the general population to reduce CHD incidence [8]. The primary prevention approach relies on three foundational elements that consist of prompt hypertension and hyperlipidemia testing along with community-wide healthy life education and expanded preventive health care facilities. The prevention of future cardiovascular events requires secondary prevention protocol to address patients with high-risk CHD profiles or established CHD through intense risk-factor management and scheduled follow-up. In eight villages in Indonesia’s East Java Province’s Malang District, Maharani A et al. [9] performed comprehensive household screenings for cardiovascular threat issues in persons aged 40 years and older. The World Health Organization/ International Society of Hypertension’s region-specific charts, which take into account factors such age, sex, blood pressure, diabetes status, and smoking behavior, were used to create 10-year cardiovascular risk scores. Adults in Indonesia who are 40 years or older are at high cardiovascular risk, and preventative treatment rates are poor. Population-based and clinical strategies for CVD prevention should be emphasized in both urban and rural regions. This work applies various classification algorithms and ensemble models for the early diagnosis of CHD based on the number of features including sex, age, smoking habit, diabetes, blood pressure, heart rate, and glucose. Section II summarizes the existing work in the field of CHD and the use of conventional and ensemble classification algorithms. Section III discusses the proposed model and the results of the data preprocessing techniques applied to the dataset. The experimental investigations and results are discussed in Section IV. Conclusions of the work are stated in Section V.

II. RELATED WORK

For Heart Disease Prediction (HDP), Jawalkar, A.P., et. al., [10] suggested a method using a Decision Tree-based Random Forest (DTRF) classifier with loss optimization. The findings show that the proposed HDP-DTRF method outperforms conventional techniques, achieving an 85% F1-score, and 96% accuracy on publicly accessible datasets. H. Yang et.al, [11] proposed the HY_OptGBM prediction model, which uses the optimized LightGBM classifier to predict CHD. The most sophisticated hyperparameter optimization framework (OPTUNA) was used in this study to optimize the hyperparameters of the prediction model. CHD data from the Framingham Heart Institute’s CHD data was used in this study to assess the prediction model. The proposed model outperformed other comparative models with an AUC score of 97.8%. C. Bemando et. al. [12] developed coronar, the model draws its information from public Cleveland heart disease patient data located at UCI repository. The study confirms that Random Forest achieves 75.00% accuracy while Gaussian Naïve Bayes and Bernoulli Naïve Bayes deliver 85.00% accuracy each. The data analysis demonstrates that the Gaussian and Bernoulli Naïve Bayes models show superior performance compared to Random Forest regarding precision and recall and F-measure which proves their worth for treatment forecasting. The research by Dulani et al. [13] provides an extensive system design approach which combines machine learning methods with Explainable AI (XAI) for developing dependable Coronary Artery Disease (CAD) prediction models.

The main purpose combines two critical tasks: prediction model development in conjunction with SHapley Additive exPlanations (SHAP) analysis which expands model understanding. Different machine learning algorithms were applied through careful system design to determine which model is best for CAD prediction. This study emphasizes the significance of interpretability and predicted accuracy in medical decision making. Shah, D., et. al. [14] discussed heart disease-related features and a model based on algorithms such as random forest, Naïve Bayes, decision tree, and K-nearest neighbor algorithms. This makes advantage of the Cleveland database, which is a UCI archive of people with cardiac illness with 303 samples and 76 features. Only 14 of these 76 characteristics were considered for testing, which is crucial for verifying the effectiveness of various algorithms. The findings showed that K-nearest neighbor yielded the highest accuracy score. In order to provide supplemental medical services, Shan Xu et. al. [15] concentrated on developing a more precise and useful risk prediction system. Both the Cleveland Heart-Disease Database (CHDD) and the PKU People’s Hospital Cardiology inpatient dataset were examined to verify their accuracy and usefulness. Our system’s accuracy of 91.6% in the CHDD test is noticeably greater than that of other approaches. With the exception of SVM (98.9%), it outperforms most other classifiers with an accuracy of 97% in the People’s Hospital dataset test; nevertheless, random forest takes half as long as SVM. A. N. Repaka et.al., [16] focused on diagnosing cardiac disease by taking into account prior data and knowledge to build Smart Heart Disease Prediction (SHDP) using Navies Bayesian that predicts heart disease risk variables. Dataset collection, application-based user registration and login, Navies Bayesian classification, prediction, and secure data transfer using AES (Advanced Encryption Standard) are the steps in the suggested methodology. The results show that the current diagnostic approach predicts risk factors for cardiac illnesses in an effective manner. Zhenzhen Du et.al., [17] sought to develop a high-precision CHD forecast method using big data and machine learning techniques, based on a sizable group of hypertension patients in Shenzhen, China. With an area under the receiver operating characteristic curve (AUC) score of 0.943, the ensemble approach XGBoost demonstrated significant accuracy in predicting the beginning of CHD within three years for the independent test dataset. On the same datasets, comparison study revealed that machine-learning techniques greatly outperformed fixed models or traditional risk scales, while nonlinear models (K-nearest neighbor with AUC 0.908, random forest with AUC 0.938) perform better than linear models (logistic regression with AUC 0.865). The development of an accurate predictive model by Bhatt CM et.al, [18] aims to reduce cardiovascular disease death numbers. The accuracy rates can rise through implementation of k-modes clustering with Huang beginning as a suggested technique. The analysis utilizes XGBoost (XGB) and Multilayer Perceptron's (MP) as well as Random Forest (RF) and Decision Tree classifier (DT). The evaluation of Area Under the Curve show 0.94 for decision trees alongside 0.95 for XGBoost and 0.95 for random forests and 0.95 for multilayer perceptron's. The accuracy analysis of multilayer perceptron with cross-validation has demonstrated the superior performance among all employed methods reaching 87.28% accuracy.

A combination of a pre-trained Deep Neural Network (DNN) followed by Principal Component Analysis (PCA) for dimensionality reduction through which the prediction occurs by a Logistic Regression (LR) model was presented by Diman Hassan et.al., [19] as a novel approach for Heart Disease (HD) prediction. A DNN + PCA + LR combination was tested on Cleveland HD dataset which is available to the public. The proposed methodology achieved superior accuracy rates compared to existing methods through most measurements according to evaluation results. Mert Ozcan and Serhat Peker analyzed the Classification and Regression Tree (CART) algorithm that serves as a supervised machine learning method for finding decision rules which connect input data to output values and performs cardiac disease predictions [20]. Furthermore, the study's conclusions prioritize the characteristics that affect heart disease. The model's dependability is confirmed by the prediction's 87% accuracy when all performance metrics are taken into account. On testing data, which makes up 85.70% of the entire NHANES dataset, Aniruddha Dutta et.al., [21] developed a CNN architecture with a classification power of 77% to properly classify the existence of CHD patients and 81.8% to successfully classify the nonappearance of CHD cases. This outcome shows that other healthcare studies with a comparable sequence of features and imbalances can use the suggested architecture. Although our suggested CNN model's recall values are similar to those of other machine learning techniques like SVM and random forest, our model is more accurate at predicting negative (non-CHD) cases. C. Beulah Christalin Latha and S. Carolin Jeeva [22] studied ensemble classification, a technique that combines several classifiers to increase the accuracy of weak algorithms. The study's findings show that ensemble methods, such bagging and boosting, perform satisfactorily in determining heart disease risk and are useful in increasing the prediction accuracy of weak classifiers. Ensemble classification helped weak classifiers achieve an accuracy gain of up to 7%.

III. DATA PRE-PROCESSING AND PROPOSED MODEL

The machine learning pipeline requires data pre-processing as its essential first step in order to increase algorithm performance while promoting data quality standards. Data pre-processing requires the transformation of unorganized data into ready-to-train format which organizes information into a structured shape. Regular data pre-processing duties involve encoding categorical data into numbers together with outlier adjustment and feature scaling to normalize ranges and different methods to handle missing data values [23] [24]. Multiple techniques such as feature selection and data augmentation and dimensionality reduction help both boost model effectiveness and stop overfitting phenomena. The pre-processing of data enhances machine learning algorithm efficiency and prediction accuracy because it helps find meaningful patterns [25]. This research delivers a system to detect early coronary heart disease onset among patients and non-patients. Multiple ensemble techniques enhanced both accuracy and prediction stability throughout refined classification modelling processes on the Otto dataset. Bagging, voting, and stacking techniques were used to build ensemble frameworks, and the outcomes of these techniques were carefully examined. The process flow of the suggested model is shown in Figure 1.

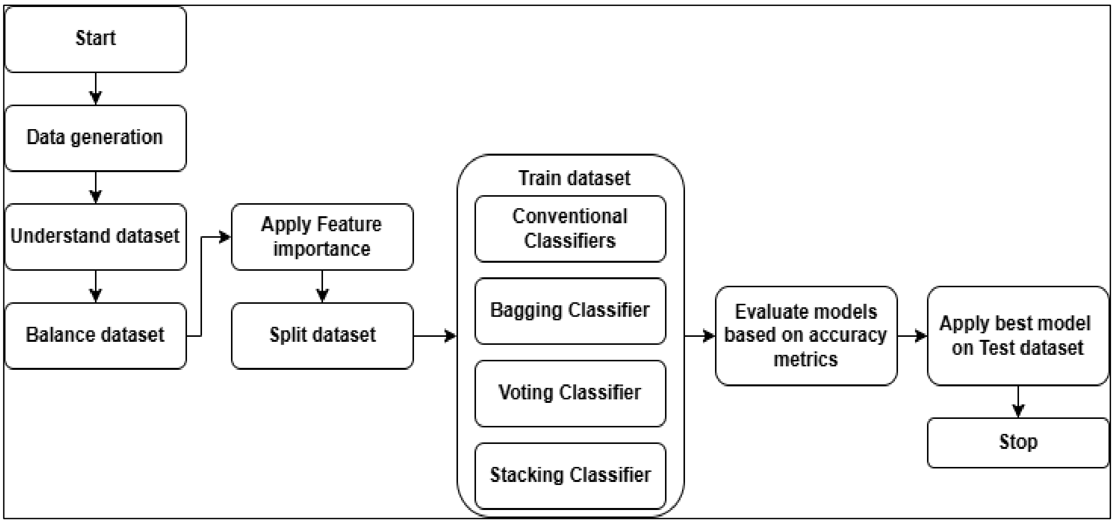

Figure 1: Steps in the proposed model.
Source: Authors, (2026).

III.1 DATASET AND ITS FEATURES

The “Framingham” dataset, which comes from a current cardiovascular study on the people of Framingham, Massachusetts, is openly accessible on the Kaggle website [26]. Determining the patient’s 10-year risk of increasing coronary heart disease is the classification goal. Figure 2 displays details about the five patients in a dataset consisting of 4,240 records and 15 features.

male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0

Figure 2: Sample dataset.
Source: Authors, (2026).

III.2 PERFORMANCE METRICS

Classification algorithm performance measures are used to assess the performance of the model with respect to predicting the target classes. The most popular metrics are as follows:

Accuracy – In this measure the ratio of successful instance classifications reflects all total instances.

$$Accuracy = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{Total\ Instances}$$

Precision – The proportion of correctly predicted positive instances among all instances predicted as positive.

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}$$

Recall (Sensitivity or True Positive Rate) – The proportion of correctly predicted positive instances among all actual positive instances.

$$\mathrm{Recall} = \frac{\mathrm{True\ Positives\ (TP)}}{\mathrm{True\ Positives\ (TP) + False\ Negatives\ (FN)}}$$

F1-Score – The harmonic mean of precision and recall, providing a balanced metric.

$$\mathrm{F1 - Score} = 2\mathrm{x}\frac{\mathrm{Precision\ x\ Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

Specificity (True Negative Rate) – The rate at which negative cases are predicted correctly out of all the real cases of negative.

$$\mathrm{Specificity} = \frac{\mathrm{True\ Negatives\ (TN)}}{\mathrm{True\ Negatives\ (TN) + False\ Positives\ (FP)}}$$

ROC-AUC (Receiver Operating Characteristic – Area Under the Curve) – This measurement evaluates how effectively the model manages true positive recognition rates (Recall) and wrong positive classifications. A model achieves higher performance when it displays increased AUC values.

Log Loss (Logarithmic Loss) – Quantifies the uncertainty of the model’s predictions. Lower values indicate better performance.

$$\mathrm{Log\ Loss} = -\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}[\mathrm{y_i}\log(\mathrm{p_i}) + (1-\mathrm{y_i})\log(1-\mathrm{p_i})]$$

Confusion Matrix **–** The following Table 1 shows a tabular representation of predictions vs. actuals, showing TP, FP, TN, and FN.

Table 1: Confusion matrix.

		Predicted Value		
		Yes	No	
Actual Value	Yes	True Positives (TP)	False Negatives (FN)	**Recall**
	No	False Positives (FP)	True Negatives (TN)	**Specificity**
		Precision	**Negative Predictive Value**	**Accuracy**

Source: Authors, (2026).

III.3 MODELLING AND COMPUTATIONAL ENVIRONMENT

A correlation matrix is a table that illustrates the correlation coefficients among variables in a dataset. It is frequently used in machine learning to find redundant or strongly correlated variables as well as to comprehend the links between various aspects. This improves interpretability by using colour gradients and annotation to visualize the matrix and is shown in Figure 3. Compared to all the independent data, the correlation coefficient between education and target variable TenYearCHD is very low and actually negative.

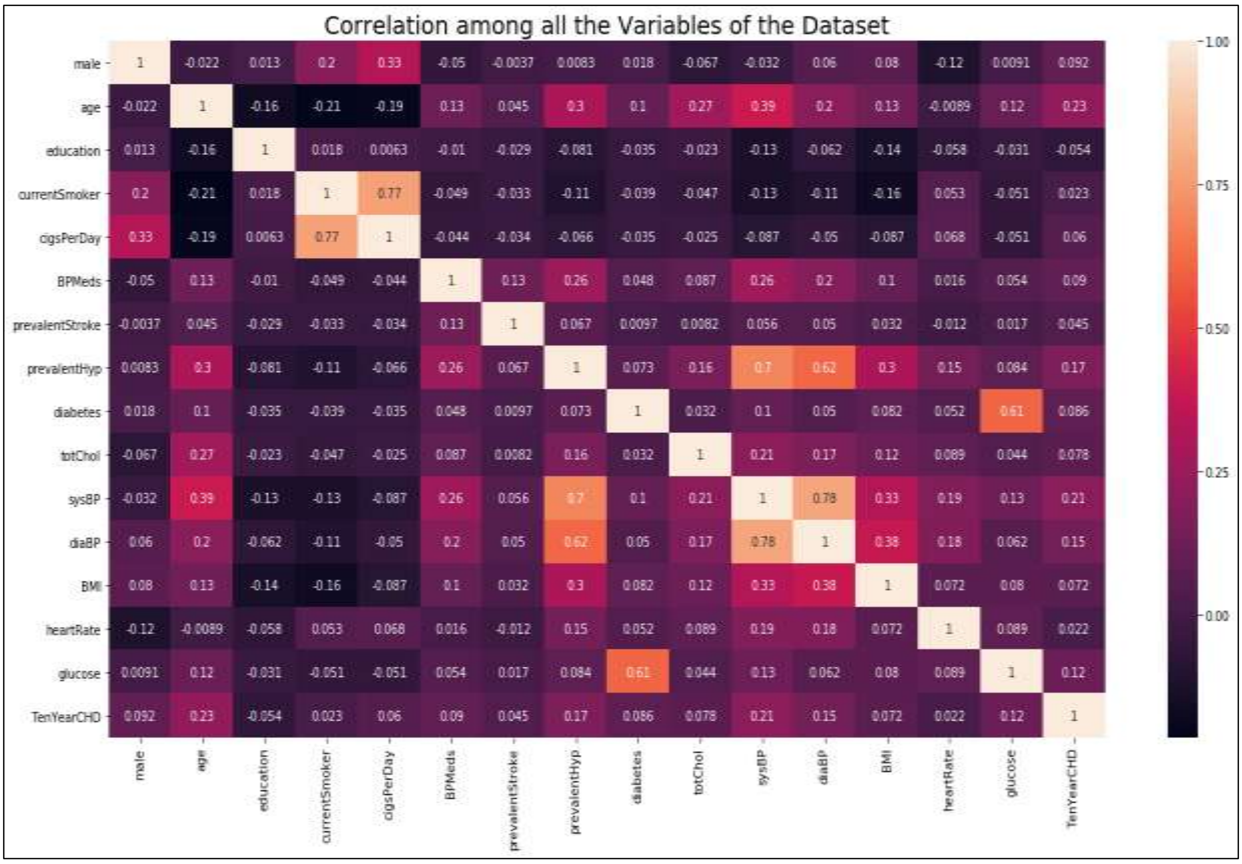

Figure 3: Correlation heatmap for dataset.
Source: Authors, (2026).

III.4 DATA ANALYSIS

III.4.1 Univariate Analysis

The study and analysis of a single variable within a dataset is known as univariate analysis. Summarizing and comprehending the properties of individual variables, such as their distribution, central tendency, dispersion, and possible anomalies, is a basic stage in exploratory data analysis (EDA). Observations from the categorical features

- BPmeds, prevalentStroke and diabetes are highly imbalanced.
- There are four levels of education whereas the rest categorical features are all binary.
- CurrentSmoker has nearly an equal number of Smokers and non-Smokers.

For the plots shown in Figure 4 observations made for numerical features are

- cigsPerDay does not have even distribution but the majority of the data is at 0
- The majority portions of the following columns lie in the range:
  - totChol: 150 to 300
  - sysBP: 100 to 150
  - diaBP: 60 to 100
  - BMI: 20 to 30
  - heartRate: 50 to 100
  - glucose: 50 to 150

The class distribution of data in the dependent feature TenYearCHD is highly imbalanced. As it were, the number of negative cases are more than the number of positive cases. This would cause class imbalance issue during the process of fitting our models. Consequently, this issue needs to be managed and handled. So, the dataset is resampled by oversampling positive cases and the classes are now balanced for model fitting. The Fig. 5 shows the distribution before and after resampling.

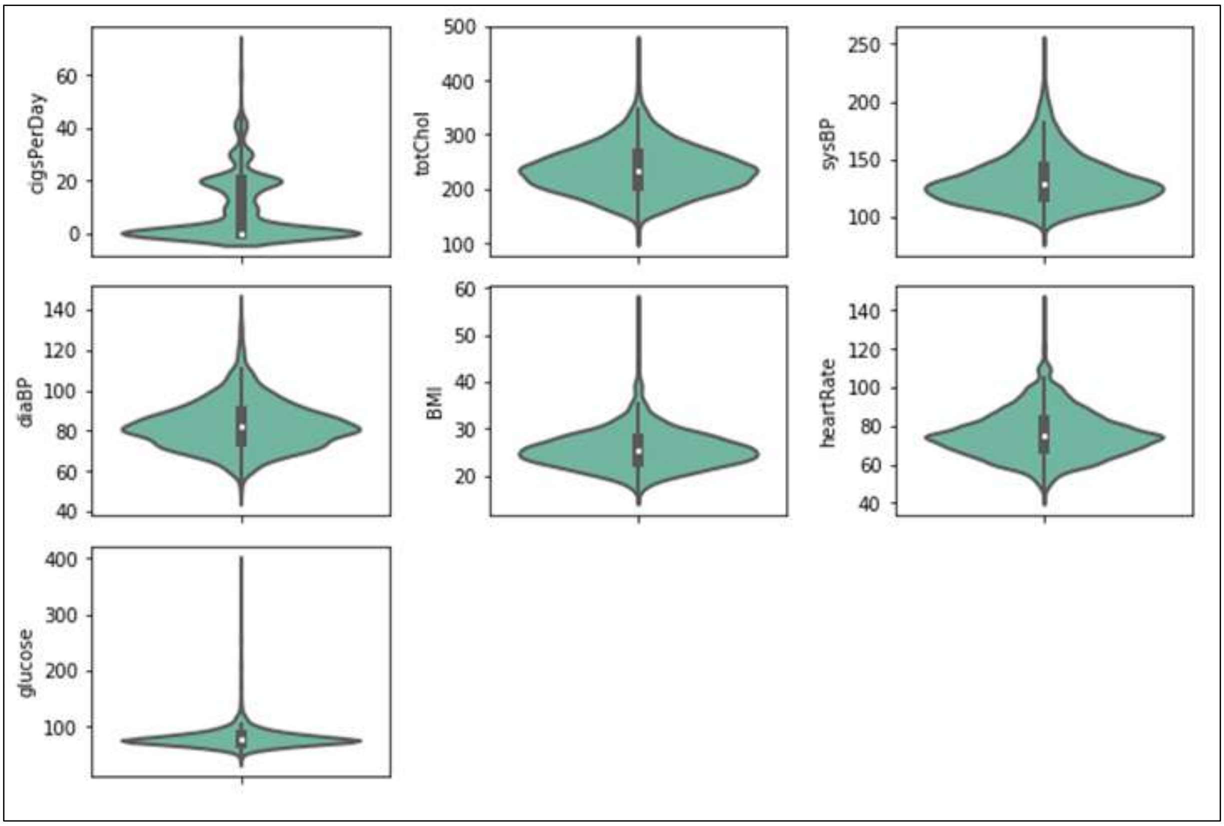

Figure 4: Data distribution in numerical features.
Source: Authors, (2026).

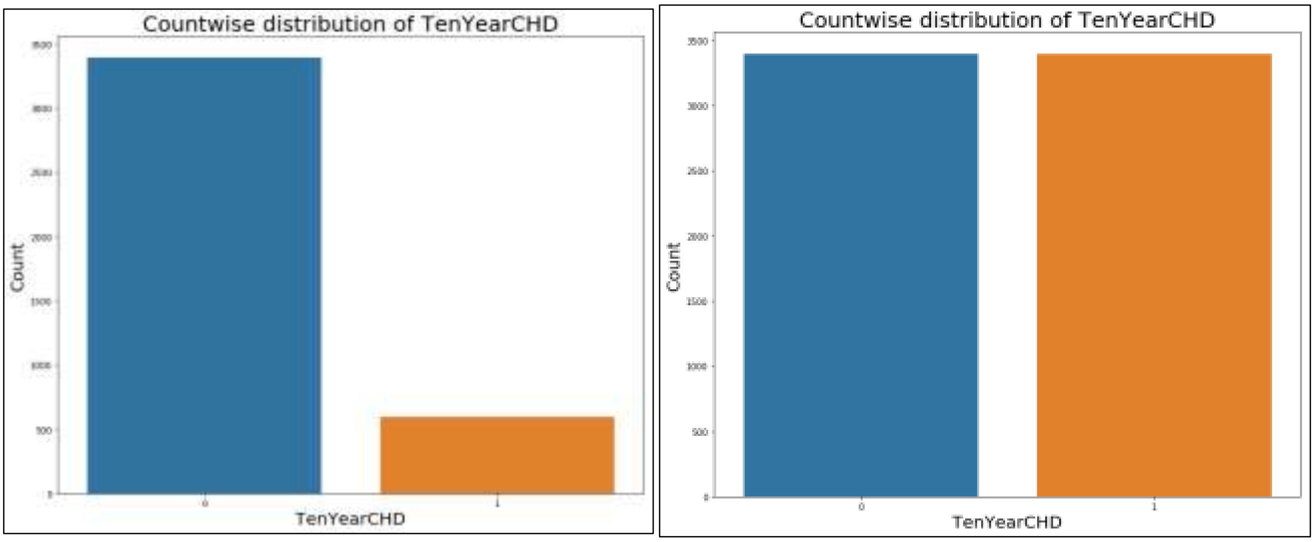

Figure 5: Before and after resampling TenYearCHD.
Source: Authors, (2026).

III.4.2 Bivariate Analysis

Bivariate analysis looks at how two variables in a dataset relate to one another. It is a crucial phase in exploratory data analysis (EDA) and aids in identifying patterns, trends, or relationships. Different methods and visualizations are employed depending on the kind of variables (numerical, categorical). From the visualization shown in Figure 6, it can also be seen that Mid-age groups between or after age 38-46 contain more currentSmokers. There is no below 32 year old currentSmokers and the maximum age of a currentSmokers is 70. The upward shift of the boxplots in Figure 7 indicates that adults had higher levels of cholesterol (bad cholesterol in general). Other observations made in this bivariate analysis are

- Non-smokers, i.e. a cigsPerDay of 0.0, are at a very low risk of developing the disease.
- Those with sysBP between 72-130 have a reduced risk of getting the disease.
- Minor relation found between higher risks of TenYearCHD with higher diaBP.
- Individuals who have diaBP of up to 80.0 are less likely to be affected by the disease.

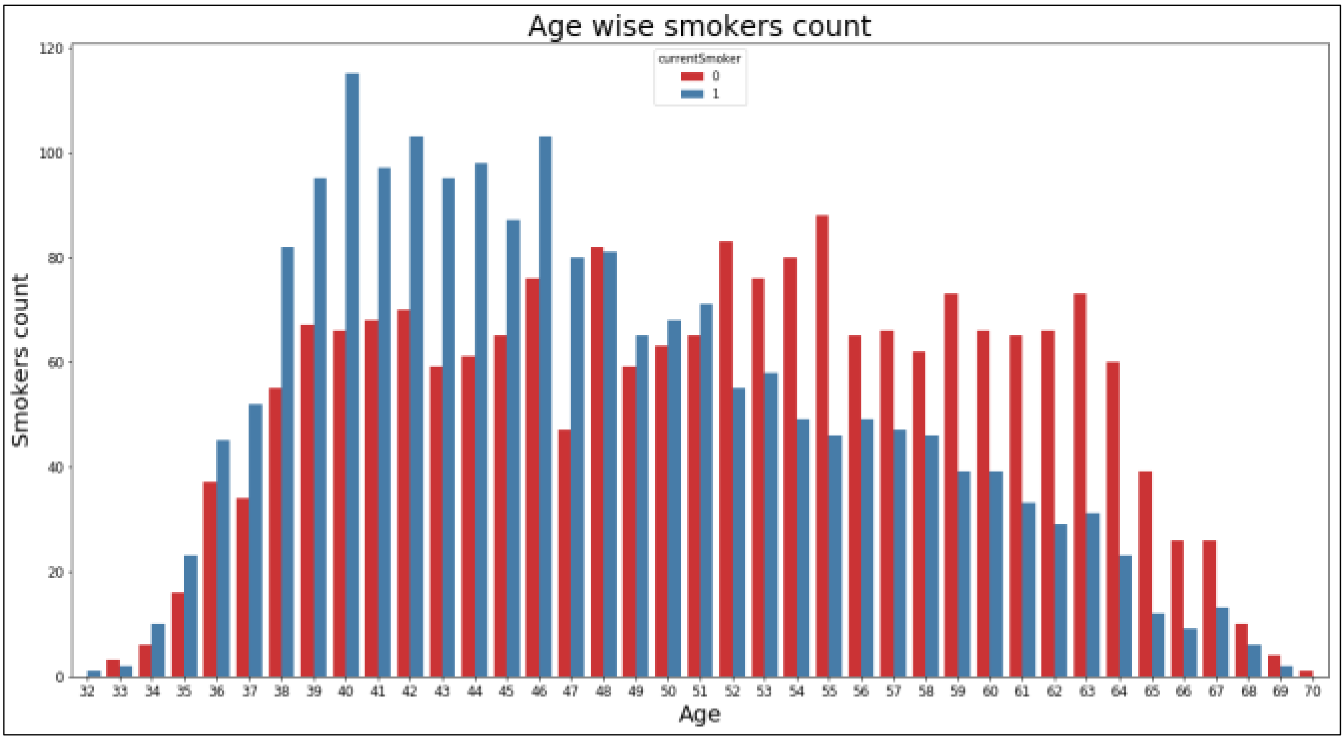

Figure 6: Age vs Smokers count.
Source: Authors, (2026).

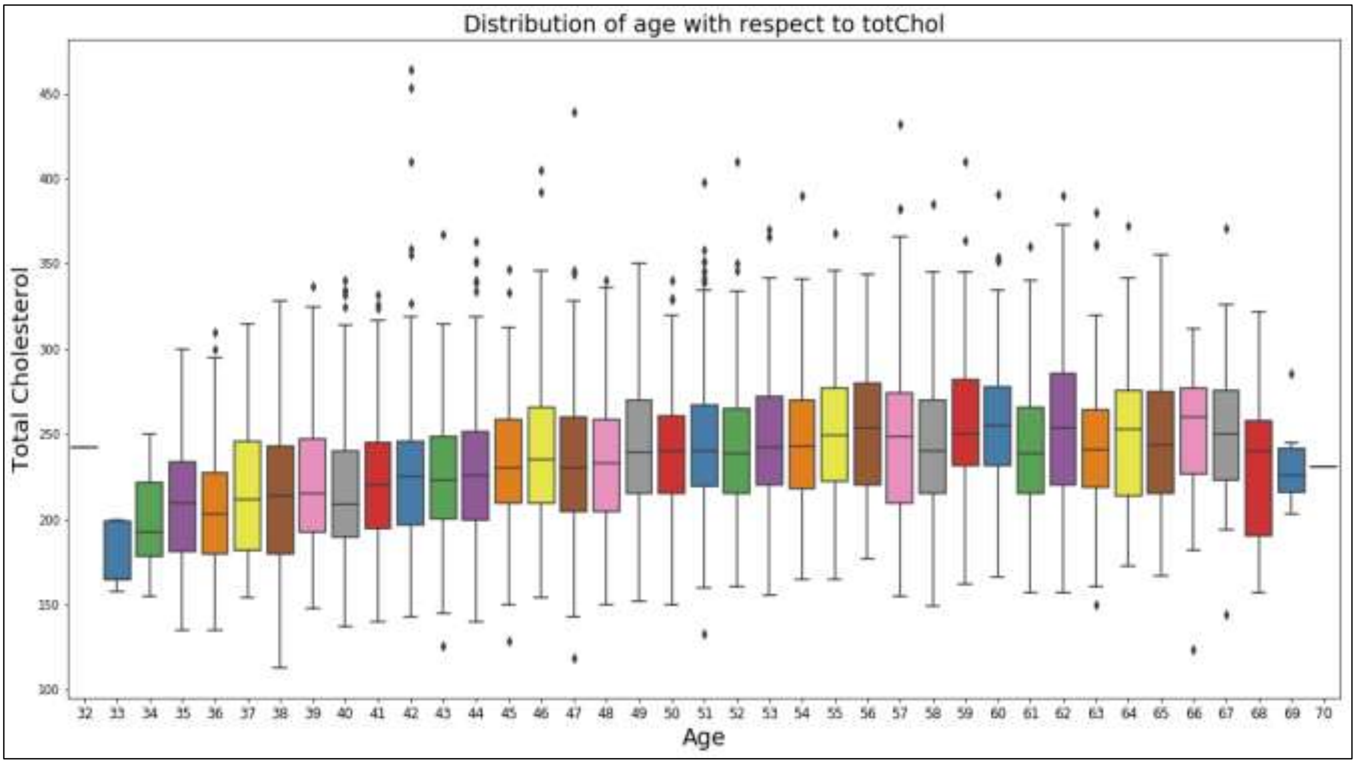

Figure 7: Distribution of Age with respect to Total Cholesterol.
Source: Authors, (2026).

III.4.3 Feature Selection

Using machine learning requires identifying and choosing the most important features contained in a dataset in order to enhance model accuracy as well as lower its complexity. It eliminates irrelevant or redundant data, which can lead to overfitting, increased computational cost, and decreased interpretability. The plot in Figure 8 shows the features and their respective chi-square test scores. The top ten features are selected for implementing the algorithms.

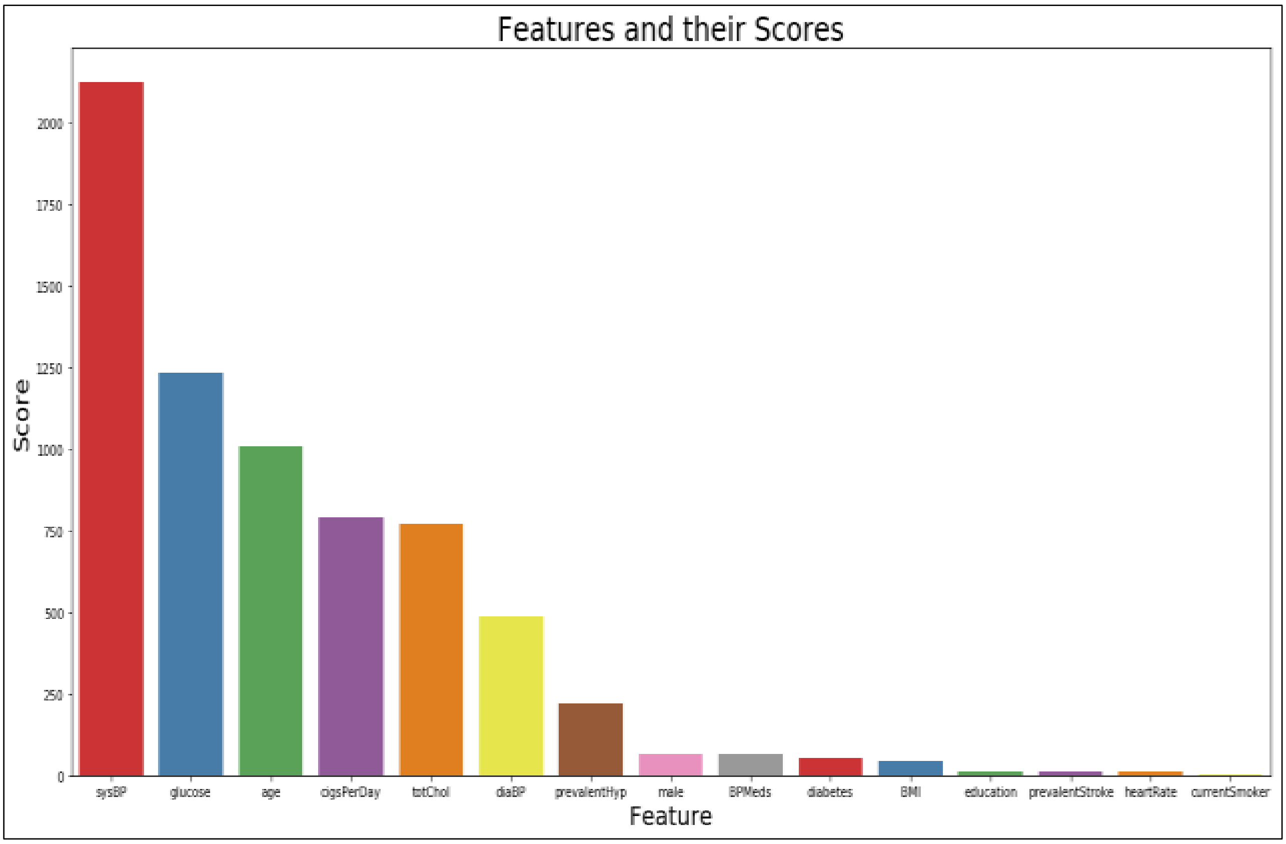

Figure 8: Features and their Chi-scores.
Source: Authors, (2026).

III.5 DATA PREPARATION

In general, the data in numerical features are of different scales and it needs to ensure that all features are on the same scale, making the model training process more stable and effective. In data science, normalization is the process of standardizing numerical data to fall into a predetermined range, usually between 0 and 1. It ensures that every feature makes an equal contribution to the analysis and keeps models from favouring features with larger scales. Min Max Scalar is applied to scale the data. The dataset is split into train and test in the ratio of 80:20.

IV. RESULTS ANALYSIS AND DISCUSSION

IV.1 CLASSIFICATION ALGORITHMS RESULTS

In this study, the dataset is trained using ten classification algorithms. Performance of these algorithms is related with respect to the metrics – Accuracy, Specificity, Precision, Recall, F1-Score and F2-Score. The Table 2 and Table 3 illustrates the results of classification algorithms on the train and test dataset respectively. From the Table 2 it is evident that, KNN Classifier, Decision Tree Classifier and XG Boost are giving 100%, 100% and 98.95% accuracy and it needs to be verified for test data also. For the test data, these algorithms are giving 93.37%, 90.86% and 91.23% accuracy. These results shows that KNN Classifier not fall under overfit problem. The confusion matrices for these three classifiers are shown in Figure 9 and KNN Classifier has made 90 wrong predictions out of 1358 samples.

Table 2: Train set results.

S.No.	Model	Accuracy	Precision	Recall	F1 Score	F2 Score
1	SGD Classifier	0.671823	0.674597	0.671823	0.670644	0.670743
2	Logistic Regression	0.677164	0.677183	0.677164	0.677163	0.677161
3	KNN Classification	1.000000	1.000000	1.000000	1.000000	1.000000
4	Naïve Bayes	0.609761	0.647355	0.609761	0.582372	0.589847
5	Decision Tree	1.000000	1.000000	1.000000	1.000000	1.000000
6	Random Forest	0.984530	0.984569	0.984530	0.984530	0.984526
7	Gradient Boosting	0.756538	0.756821	0.756538	0.756489	0.756473
8	Ada Boosting	0.689134	0.689150	0.689134	0.689134	0.689133
9	XGBoost	0.989503	0.989504	0.989503	0.989503	0.989503

Source: Authors, (2026).

Table 3: Test set results.

S.No.	Model	Accuracy	Precision	Recall	F1 Score	F2 Score
1	SGD	0.653903	0.656044	0.653903	0.652067	0.652482
2	Logistic Regression	0.667158	0.667190	0.667158	0.666999	0.667052
3	KNN Classification	0.933726	0.939146	0.933726	0.933468	0.932920
4	Naïve Bayes	0.609720	0.646893	0.609720	0.586412	0.592023
5	Decision Tree	0.908689	0.919147	0.908689	0.908010	0.907037
6	Random Forest	0.921944	0.924964	0.921944	0.921757	0.921465
7	Gradient Boosting	0.712813	0.712862	0.712813	0.712712	0.712743
8	Ada Boosting	0.679676	0.679654	0.679676	0.679656	0.679666
9	XGBoost	0.912371	0.916567	0.912371	0.912085	0.911691

Source: Authors, (2026).

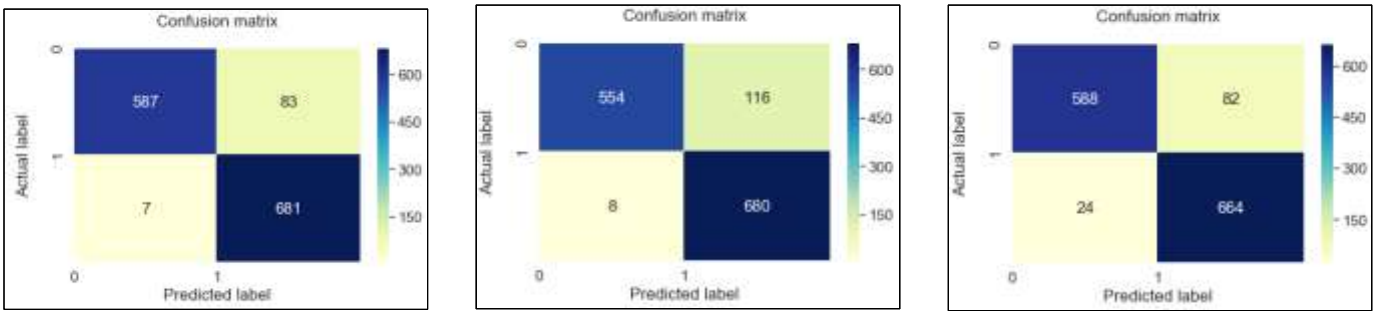

Figure 9: Confusion matrices generated for test data by a) KNN b) Decision Tree c) Random Forest.

Source: Authors, (2026).

IV.2 ENSEMBLE MODEL RESULTS

Ensemble models are a dominant tool in machine learning that combine multiple individual models to increase overall performance and robustness. By leveraging the strengths of diverse algorithms, ensemble methods can attain higher accuracy and superior generalization compared to single models. Techniques such as bagging, boosting, and stacking help reduce variance and bias, leading to more reliable predictions. The model creation process of bagging produces several models with different training data segments while boosting applies successive model optimization to tackle past prediction mistakes. A stacking method consists of creating an additional model which learns to combine predictions made by various base models. Ensemble models which use the Voting Classifier produce improved performance through the combination of multiple machine learning models. The model assembles multiple model predictions into one prediction through a majority vote for classifications and average value computation for regression problems. The two voting forms include hard voting which depends on the models' majority class predictions as well as soft voting that utilizes the individual models' averaged prediction probabilities.

Other than their practicality Ensemble models become the preferred framework for complex prediction problems which span across different application domains. Several techniques under ensemble models work together to improve predictive performance while maintaining unique methods for uniting multiple models. The current research applied three ensemble models which included bagging and stacking in addition to voting. Tables 4 and 5 exhibit the results for train and test data after the top-performing algorithms from Table 3 are chosen. XG Boost, KNN, and Random Forest classifiers are used in the suggested voting classifier, which produced an accuracy of 94.25% for test data. XG Boost, Random Forest, and Decision Tree are used as base classifiers for stacking classifier, while KNN is used as the final classifier. For both train and test data, this classifier produced 100% and 92.19% accuracy, respectively. Fig. 10 displays the confusion matrices for the three proposed ensemble models. The voting classifier outperforms the KNN classifier in terms of accuracy. Table 6 presents the results of other performance metrics and voting classifier resulted with 98.87 for ROC AUC mean.

Table 4: Train data set results for Ensemble models.

S.No.	Model	Accuracy	Precision	Recall	F1 Score	F2 Score
1	KNN Bagging	0.989871	0.989995	0.989871	0.989871	0.989856
2	Voting Classifier	0.999632	0.999632	0.999632	0.999632	0.999632
3	Stacking Classifier	1.000000	1.000000	1.000000	1.000000	1.000000

Source: Authors, (2026).

Table 5: Test data set results for Ensemble models.

S.No.	Model	Accuracy	Precision	Recall	F1 Score	F2 Score
1	KNN Bagging	0.919735	0.924276	0.919735	0.919459	0.919021
2	Voting Classifier	0.942563	0.944736	0.942563	0.942464	0.942240
3	Stacking Classifier	0.921944	0.922341	0.921944	0.921942	0.921895

Source: Authors, (2026).

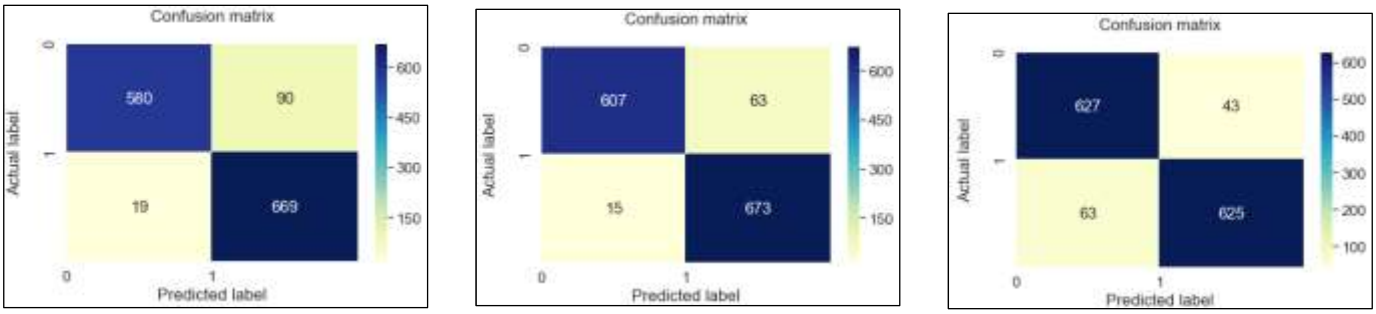

Figure 10: Confusion matrices generated for test data by a) KNN Bagging b) Voting Classifier c) Stacking Classifier.
Source: Authors, (2026).

Table 6: Accuracy metrics for Ensemble models.

S.No.	Model	ROC AUC Mean	ROC AUC STD	Accuracy Mean	Accuracy STD
1	KNN Bagging	97.23	0.99	94.88	1.38
2	Voting Classifier	98.87	0.51	96.02	1.31
3	Stacking Classifier	94.98	1.15	94.97	1.14

Source: Authors, (2026).

V. CONCLUSION

A unique ensemble approach for categorizing patients and predicting their risk for CHD was proposed in this paper. Make sure to perform pre-processing in order to comprehend the data, identify correlations between features, eliminate noisy data, balance data in two classes in the dependent feature, and extract contributing features. To manage the entire experiment with different evaluation factors such as specificity, recall, sensitivity, precision, accuracy, F-score, and confusion matrix, a Python implementation platform was used. The dataset was integrated with 4240 patients who may have a risk of coronary heart disease. KNN Classification has the best accuracy among the eight classification algorithms: SGD Classifier, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, Gradient Boosting, Ada Boosting, and XGBoost. Its training and testing accuracies are 1.0 and 0.93, respectively. The dataset is subjected to ensemble techniques like bagging, voting, and stacking classifiers. Using the best-performing classifiers – KNN, Random Forest, and Decision Tree—a voting classifier is suggested. For both train and test data, the suggested approach achieved model accuracy of 0.99 and 0.94. The ROC AUC Mean and Accuracy Mean for the work are 98.87 and 96.02, respectively.

VI. AUTHOR’S CONTRIBUTION

Conceptualization: Ch V Raghavendran, Kathi Chandra Mouli, S. Bhargavi Latha.
Methodology: Ch V Raghavendran, Kathi Chandra Mouli.
Investigation: S. Bhargavi Latha.
Discussion of results: Ch V Raghavendran, Kathi Chandra Mouli, S. Bhargavi Latha.
Writing: Review and Editing: Ch V Raghavendran, Kathi Chandra Mouli, S. Bhargavi Latha.
Resources: Kathi Chandra Mouli, S. Bhargavi Latha.
Supervision: Ch V Raghavendran, Kathi Chandra Mouli, S. Bhargavi Latha.
Approval of the final text: Ch V Raghavendran, Kathi Chandra Mouli, S. Bhargavi Latha.

VII. REFERENCES

[1] Foxwell R, Morley C, Frizelle D. Illness perceptions, mood and quality of life: a systematic review of coronary heart disease patients. J Psychosom Res. 2013 Sep;75(3):211-22. doi: 10.1016/j.jpsychores.2013.05.003.

[2] Price KJ, Gordon BA, Bird SR, Benson AC. A review of guidelines for cardiac rehabilitation exercise programmes: Is there an international consensus? Eur J Prev Cardiol. 2016 Nov;23(16):1715-1733. doi: 10.1177/2047487316657669.

[3] Candelaria D, Randall S, Ladak L, Gallagher R. Health-related quality of life and exercise-based cardiac rehabilitation in contemporary acute coronary syndrome patients: a systematic review and meta-analysis. Qual Life Res. 2020 Mar;29(3):579-592. doi: 10.1007/s11136-019-02338-y.

[4] Byrne M, Walsh J, Murphy AW. Secondary prevention of coronary heart disease: patient beliefs and health-related behaviour. J Psychosom Res. 2005 May;58(5):403-15. doi: 10.1016/j.jpsychores.2004.11.010.

[5] Bane C, Hughes CM, McElnay JC. The impact of depressive symptoms and psychosocial factors on medication adherence in cardiovascular disease. Patient Educ Couns. 2006 Feb;60(2):187-93. doi: 10.1016/j.pec.2005.01.003.

[6] Brydon L, Magid K, Steptoe A. Platelets, coronary heart disease, and stress. Brain Behav Immun. 2006 Mar;20(2):113-9. doi: 10.1016/j.bbi.2005.08.002.

[7] Juergens MC, Seekatz B, Moosdorf RG, Petrie KJ, Rief W. Illness beliefs before cardiac surgery predict disability, quality of life, and depression 3 months later. J Psychosom Res. 2010 Jun;68(6):553-60. doi: 10.1016/j.jpsychores.2009.10.004.

[8] Sebregts, E. H. W. J., Falger, P. R. J., & Bär, F. W. H. M. (2000). Risk factor modification through nonpharmacological interventions in patients with coronary heart disease. Journal of Psychosomatic Research, 48(4-5), 425–441. https://doi.org/10.1016/S0022-3999(99)00113-0.

[9] Maharani A, Sujarwoto, Praveen D, Oceandy D, Tampubolon G, Patel A (2019) Cardiovascular disease risk factor prevalence and estimated 10-year cardiovascular risk scores in Indonesia: The SMARThealth Extend study. PLoS ONE 14(4): e0215219. https://doi.org/10.1371/journal.pone.0215219.

[10] Jawalkar, A.P., Swetcha, P., Manasvi, N. et al. Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting. J. Eng. Appl. Sci. 70, 122 (2023). https://doi.org/10.1186/s44147-023-00280-y.

[11] H. Yang, Z. Chen, H. Yang and M. Tian, "Predicting Coronary Heart Disease Using an Improved LightGBM Model: Performance Analysis and Comparison," in IEEE Access, vol. 11, pp. 23366-23380, 2023, doi: 10.1109/ACCESS.2023.3253885.

[12] C. Bemando, E. Miranda and M. Aryuni, "Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms," 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), Pekan, Malaysia, 2021, pp. 232-237, doi: 10.1109/ICSECS52883.2021.00049.

[13] Dulani, H. et al. (2024). Heart Disease Prediction Using Ensemble Techniques and Explainable AI Validation. In: Dehuri, S., Cho, SB., Padhy, V.P., Shanmugam, P., Ghosh, A. (eds) Machine Intelligence, Tools, and Applications. ICMITA 2024. Learning and Analytics in Intelligent Systems, vol 40. Springer, Cham. https://doi.org/10.1007/978-3-031-65392-6_24.

[14] Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 (2020). https://doi.org/10.1007/s42979-020-00365-y.

[15] S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan and T. Zhu, "Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework," 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), Beijing, China, 2017, pp. 228-232, doi: 10.1109/ICBDA.2017.8078813.

[16] A. N. Repaka, S. D. Ravikanti and R. G. Franklin, "Design And Implementing Heart Disease Prediction Using Naives Bayesian," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 292-297, doi: 10.1109/ICOEI.2019.8862604.

[17] Du Z, Yang Y, Zheng J, Li Q, Lin D, Li Y, Fan J, Cheng W, Chen XH, Cai Y, Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation, JMIR Med Inform 2020;8(7):e17257, doi: 10.2196/17257.

[18] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms, 16(2), 88. https://doi.org/10.3390/a16020088.

[19] Diman Hassan, Haval I. Hussein, Masoud M. Hassan, Heart disease prediction based on pre-trained deep neural networks combined with principal component analysis, Biomedical Signal Processing and Control, Volume 79, Part 1, 2023, 104019, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2022.104019.

[20] Mert Ozcan, Serhat Peker, A classification and regression tree algorithm for heart disease modeling and prediction, Healthcare Analytics, Volume 3, 2023, 100130, ISSN 2772-4425, https://doi.org/10.1016/j.health.2022.100130.

[21] Aniruddha Dutta, Tamal Batabyal, Meheli Basu, Scott T. Acton, An efficient convolutional neural network for coronary heart disease prediction, Expert Systems with Applications, Volume 159, 2020, 113408, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2020.113408.

[22] C. Beulah Christalin Latha, S. Carolin Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Informatics in Medicine Unlocked, Volume 16, 2019, 100203, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2019.100203.

[23] Mouli, K. C., Raghavendran, Ch. V., Bharadwaj, V. Y., Vybhavi, G. Y., Sravani, C., Vafaeva, K. M., … Hussein, L. (2024). An analysis on classification models for customer churn prediction. Cogent Engineering, 11(1). https://doi.org/10.1080/23311916.2024.2378877.

[24] Mouli, K. C., Raghavendran, Ch. V., Mallikarjuna Rao, Ch., Ushasree, D., Indupriya, B., Vatin, N. I., & Negi, A. S. (2024). Performance analysis of linear and non-linear machine learning models for forecasting compressive strength of concrete. Cogent Engineering, 11(1). https://doi.org/10.1080/23311916.2024.2368101

[25] B. Narsimha, Ch V Raghavendran, Pannangi Rajyalakshmi, G Kasi Reddy, M. Bhargavi and P. Naresh (2022), Cyber Defense in the Age of Artificial Intelligence and Machine Learning for Financial Fraud Detection Application. IJEER 10(2), 87-92. DOI: 10.37391/IJEER.100206.

[26] https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset/data.