applied sciences

MDPI

Article

Mixed Machine Learning Approach for Efficient Prediction of Human Heart Disease by Identifying the Numerical and Categorical Features

Ghulab Nabi Ahmad [1,*], Shafiullah [2,*], Hira Fatima [1], Mohamed Abbas [3,4], Obaidur Rahman [5], Imdadullah [6,*] and Mohammed S. Alqahtani [7,8]

[1] Institute of Applied Sciences, Mangalayatan University, Aligarh 202145, India; hirafatima2014@gmail.com
[2] Department of Mathematics, K.C.T.C. College, Raxaul, BRA, Bihar University, Muzaffarpur 842001, India
[3] Electrical Engineering Department, College of Engineering, King Khalid University, Abha 61421, Saudi Arabia; mabas@kku.edu.sa
[4] Computers and Communications Department, College of Engineering, Delta University for Science and Technology, Gamasa 35712, Egypt
[5] Department of Electrical Engineering, Jamia Millia Islamia, New Delhi 110025, India; obaidrahman47@gmail.com
[6] Electrical Engineering Section, University Polytechnic, Aligarh Muslim University, Aligarh 202002, India
[7] Radiological Sciences Department, College of Applied Medical Sciences, King Khalid University, Abha 61421, Saudi Arabia; qmalak46@gmail.com or mosalqhtani@kku.edu.sa
[8] BioImaging Unit, Space Research Center, Michael Atiyah Building, University of Leicester, Leicester LE1 7RH, UK
* Correspondence: ghulamnabiahmad@gmail.com (G.N.A.); shafi.stats@gmail.com (S.); imdadamu@gmail.com (I.)

Abstract: Heart disease is a danger to people's health because of its prevalence and high mortality risk. Predicting cardiac disease early using a few simple physical indications collected from a routine physical examination has become difficult. Clinically, it is critical and sensitive for the signs of heart disease for accurate forecasts and concrete steps for future diagnosis. The manual analysis and prediction of a massive volume of data are challenging and time-consuming. In this paper, a unique heart disease prediction model is proposed to predict heart disease correctly and rapidly using a variety of bodily signs. A heart disease prediction algorithm based on the analysis of the predictive models' classification performance on combined datasets and the train-test split technique is presented. Finally, the proposed technique's training results are compared with the previous works. For the Cleveland, Switzerland, Hungarian, and Long Beach VA heart disease datasets, accuracy, precision, recall, F1-score, and ROC-AUC curves are used as the performance indicators. The analytical outcomes for Random Forest Classifiers (RFC) of the combined heart disease datasets are F1-score 100%, accuracy 100%, precision 100%, recall 100%, and the ROC-AUC 100%. The Decision Tree Classifiers for pooled heart disease datasets are F1-score 100%, accuracy 98.80%, precision 98%, recall 99%, ROC-AUC 99%, and for RFC and Gradient Boosting Classifiers (GBC), the ROC-AUC gives 100% performance. The performances of the machine learning algorithms are improved by using five-fold cross validation. Again, the Stacking CV Classifier is also used to improve the performances of the individual machine learning algorithms by combining two and three techniques together. In this paper, several reduction methods are incorporated. It is found that the accuracy of the RFC classification algorithm is high. Moreover, the developed method is efficient and reliable for predicting heart disease.

Keywords: heart disease; mixed machine learning techniques; numerical features; categorical features; RFC; DT; LDA; QDA; Nu SVC; NB; NN; SVM; GBC; KNN; LR

check for updates

Citation: Ahmad, G.N.; Shafiullah; Fatima, H.; Abbas, M.; Rahman, O.; Imdadullah; Alqahtani, M.S. Mixed Machine Learning Approach for Efficient Prediction of Human Heart Disease by Identifying the Numerical and Categorical Features. *Appl. Sci.* **2022**, *12*, 7449. https://doi.org/10.3390/app12157449

Academic Editors: Atsushi Teramoto and Tomoko Tateyama

Received: 6 June 2022
Accepted: 21 July 2022
Published: 25 July 2022

1. Introduction

Cardiovascular diseases are a group of ailments that disrupt the proper functioning of the heart. Cardiac failure (HF), coronary artery disease (CAD), vascular disease, heart rhythm abnormalities, and other conditions are among the various types of heart disease. The condition where the blood channels constrict or block, resulting in a heart attack (myocardial infarction) and chest discomfort, is known as heart disease. Chest heaviness, shortness of breath, chest pain (angina), irregular heartbeats, and heart abnormalities are key symptoms of heart disease [1]. Heart failure is a long-term condition that damages the heart chambers. Cardiovascular illness interrupts the heart's natural function, which is to pump enough blood into the whole body without raising intracardiac pressure. When the heart fails to supply enough blood to the entire body, the fluid level of the body will be maintained by the kidney, producing lung obstruction and edema in the legs and arms. Congestive heart failure (CHF) is a major health problem in today's world, impacting 26 million people globally [2]. Around 17.9 million people are killed each year from cardiovascular disease, accounting for 31% of all deaths globally [3]. Many uncontrolled risk factors for heart failure include gender, family history, rising age, etc., whereas hypertension, high cholesterol, smoking, and obesity are in control label health conditions [4]. We research and review the most common types of heart failure problems in order to better understand heart failure. Figure 1 depicts the four right atria that are responsible for regulating blood pumping.

Recently, healthcare manufacturing has collected a vast amount of patient data. However, researchers and clinicians are not effectively utilizing this information for illness diagnosis. The healthcare industry is experiencing significant problems in terms of superiority of service (SS), which assures accurate and fast illness analysis and competent patient treatment. Improper diagnosis leads to negative outcomes that are unacceptable [5].

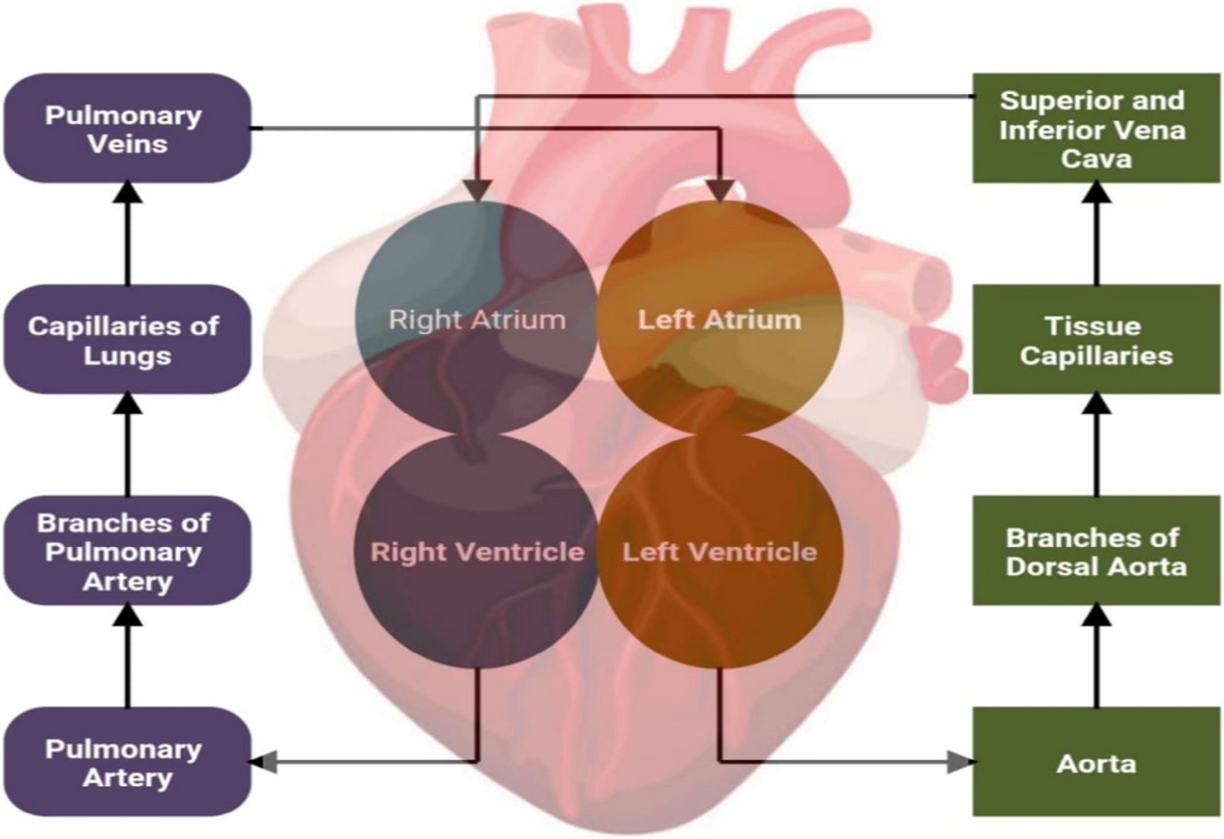

Figure 1. Schematic diagram of heart.

1.1. Main Kind of Cardiovascular Ailments

A coronary artery disease (CAD) is a kind of cardiovascular disease in which the narrowing or blockage of the coronary arteries is caused by plaque build-up. It is also known as coronary heart disease and ischemic heart disease. It is caused by fatty deposits (plaque) forming inside the arteries. As a result of the artery blockage, the blood supply to the heart muscles is restricted, causing the cardiac function to deteriorate. Coronary arteries are blood vessels that supply oxygen-rich blood to the heart muscle, allowing it to keep hiding. The coronary arteries cross the heart muscle in a straight line. The four principal coronary arteries are the right coronary artery (RCA), left coronary artery (LCA), left forward descending artery (LAD), and left circumflex artery. The medical name for this ailment is myocardial ischemia. A partial or total blockage of blood vessels causes irreversible damage to the heart, resulting in a heart attack. The four chambers of the human heart are the upper receiving chambers (right and left atria) and the bottom pumping chambers (true and left ventricle (LV)). The right atrium collects deoxygenated blood, while the right ventricle pumps it to the lungs to be oxygenated. The left atrium receives oxygenated blood from the lungs, which is subsequently delivered to all areas of the body by the left ventricle. Because of its size and purpose, the left ventricle chamber is an important and accountable component of the heart. As a result, left ventricle chamber injury is the most common cause of heart failure. By analyzing or monitoring the heart for the course of CAD and the formation of wall motion anomalies, echocardiography assists in the identification of CAD [6]. LV measurement and wall motion score can both be used to determine if a patient has CAD. As a result, LV monitoring is essential for preventing long-term damage to the LV's size, shape, and function. Ultrasound recordings are used in echocardiography to capture different viewpoints, structures, and motions of the heart. Echocardiography is a test that evaluates the functional and anatomical properties of the heart to detect cardiac disease [7]. In addition, echocardiography is utilized to assess the left ventricle discharge portion and cardiac output [8].

1.2. Congestive Heart Failure

Congestive Heart Failure (CHF) is a kind of heart failure in which the blood arteries become clogged. The cause of congestive heart failure is the insufficient pumping of blood through the heart muscle in the body. Due to this, blood and fluid collection in the lungs causes breathing problems. Because of this, the heart weakens or stiffens over time and causes cardiovascular diseases, such as clogged-up arteries in the heart (coronary artery disease) or extreme blood pressure. As a result, the blood will not be filled efficiently. The long-living life of a human being can be maintained by reducing the signs of cardiac attack and proper treatment. The quality of life can be improved by reducing weight, doing continuous exercise, applying restrictions on salt, and reducing the stresses of life. Congestive heart failure may cause severe symptoms and result in heart surgery, or a myocardial perfusion device may be required. One technique for avoiding heart failure is preventing and treating conditions that might contribute to it, such as atherosclerosis, hypertension, obesity, and being overweight.

1.3. Abnormal Heart Rhythms, Irregular Heartbeat, Cardiac Arrhythmia, Abnormal Electrocardiography

A person's description of symptoms can often aid clinicians in making an initial diagnosis and determining the severity of an arrhythmia. The most important things to look for are whether the palpitations are rapid or slow, regular or irregular, and whether they are caused by a problem with the heart's electrical system. The heart's electrical system controls when it beats and how much blood it pumps to each part of the body [9]. The most frequent symptoms of an irregular cardiac rhythm are palpitations, fatigue, loss of consciousness, dizziness, and shortness of breath. Because the signs of HF are challenging to detect, it is often known as the "silent killer", which is shown in Figure 2. For the diagnosis of heart failure, doctors offer a variety of medical tests [10], such as an echocardiogram, which uses ultrasound waves to evaluate blood flow through the heart. Another technique to identify cardiac abnormalities involving the heart's rhythm is an electrocardiogram (ECG). Holter monitoring is a portable gadget that records the patient's continuous ECG data. The heart failure of the patient's heart can be computed by tomography (CT) scans which provide an X-ray cross-sectional image of the patient's heart. Cardiovascular magnetic resonance imaging (MRI) uses powerful magnets and radio waves to create an image of the heart and its tissues.

The most widespread socioeconomic and public health problem is heart disease, affecting both men and women and resulting in many fatalities and other impairments [11,12]. Despite being one of the most common chronic conditions, heart ailment is one of the most preventable and controlled diseases worldwide, causing a substantial percentage of disability and mortality globally, with 17.9 million fatalities each year, and each second, one person dies. Under the age of 70, one-third of these fatalities occur. Researchers from all across the globe are working hard to find ways to prevent, treat, and ultimately cure heart disease. In December 2020, Carmat, a French firm, is situated in Peris has obtained authorization to carried out its implant of an artificial heart in Europe for the first time [13].

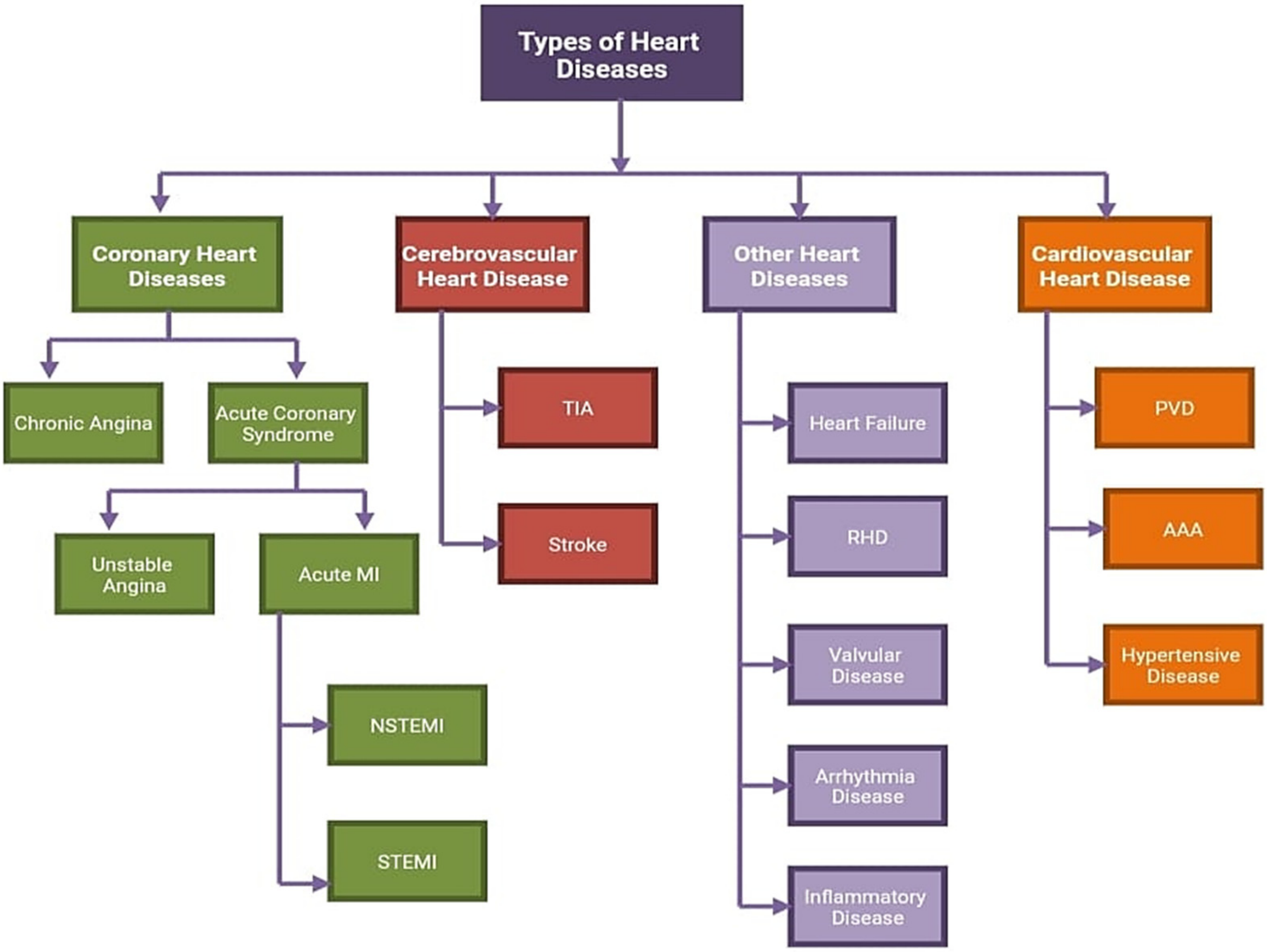

Figure 2. Types of Heart diseases.

Excess cholesterol, obesity, and high triglyceride levels are just a few of the harmful habits that people engage in [14]. Early cardiac disease detection reduces the risk of developing severe symptoms and complications [15,16]. These symptoms are similar to those of other diseases that affect the elderly, making a precise diagnosis challenging to obtain and potentially fatal. More study data and hospital patient records are becoming available as time goes on. There are multiple open sources for patient data, and a study might be done to investigate if different computer technologies can be utilized to diagnose the patients accurately and detect this problem before it becomes fatal.

2. Prediction of Heart Disease Using Machine Learning

A great variety of diagnostic systems for the automated diagnosis of various ailments, such as human heart disease, have recently been created. Methods of machine learning and optimization in machine learning have been effectively deployed on several datasets for automated heart disease identification, and it is now commonly considered to have a substantial impact in medical science. Several ML models are used to diagnose the disease and categorize or forecast the results. Large volumes of genetic data can be analyzed quickly using ML methods. To enhance projections, medical data may be used and analyzed more fully, and algorithms could be taught to predict pandemics [17,18]. Several insights are extracted from the dataset to understand the importance of each variable and how they are related to one another. Moreover, the primary aim of this work is to determine whether or not someone has a severe cardiac disease.

2.1. Linear Discriminants Analysis (LDA)

LDA is employed when all populations' variation covariance grids are homogenous. Our selection strategy in LDA is based on the linear score function, which is the population element, and the set difference covariance frames the linear score function's features.

2.2. Random Forest Classifier

The supervised learning approach is used by Random Forest, a well-known machine learning algorithm. It may be used for both classification and regression problems in machine learning. It is based on ensemble learning, which is a technique for combining a large number of classifiers to solve a difficult problem and improve the model's performance. "A random forest is a classifier that contains a number of decision trees on various subsets of a given dataset and takes the average to improve the dataset's prediction accuracy", as suggested by the name. Rather than relying on a single decision tree, the random forest considers the predictions from each tree and predicts the final output based on the majority votes of projections.

2.3. Gradient Boosting Classifier

Gradient boosting is a type of artificial intelligence (AI) that may be used to solve regression and classification problems. As a prediction model, it provides a set of overall prediction models and decision trees. It builds models in a stage-savvy approach, similar to previous improvement strategies, and summarizes them by permitting arbitrarily distinguishable unpleasant work. In order to minimize the target work, in each cycle, we adjust the basic learners to the negative angle of the negative gradient, progressively increasing the expected value and adding it to the previously emphasized incentives:

2.4. Decision Tree Classifier (DTC)

In summary, trees are a type of supervised machine learning in which data are split on a regular basis based on a parameter (in the training data, we define what the input is and what the related output is). The leaves symbolize the decisions or final results. This technique has a tree or structure like a flowchart, with the branches, leaves, nodes, and root node in a tree. The features are kept in the internal nodes, whereas the branches indicate the outcomes of each test on each node. DT is frequently used for classification applications since it does not need considerable field experience or parameter setting.

2.5. Support Vector Machine (SVM)

A support vector machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. A support vector machine is a statistical learning technique that classifies the datasets for solving both linear and nonlinear problems. The technique creates a line or a hyperplane which separates the data into classes.

2.6. Nu-Support Vector Classifier (Nu-SVC)

The sole distinction between the Nu-support vector classifier and the support vector classifier is the Nu parameter, which regulates the number of support vectors. In this lesson, we'll quickly go through how to use Python's NuSVC class from Scikit-learn to categorize data.

2.7. Logistic Regression (LR)

Logistic regression is a kind of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring. An example of logistic regression could be applying machine learning to determine if a person is likely to be infected with heart disease or not.

The nature of the target or dependent variable is dichotomous, which means there would be only two possible classes.

2.8. K-Nearest Neighbors

As the name suggests, it considers K-Nearest Neighbors (data points) to predict the class or continuous value for the new data point. The algorithm of the learning is:

- Select the number K of the neighbors
- Calculate the Euclidean distance of the K number of neighbors
- Take the K nearest neighbors as per the calculated Euclidean distance.
- Among these K neighbors, count the number of data points in each category.
- Assign new data points to that category for which the number of the neighbor is maximum.

2.9. Quadratic Discriminant Analysis (QDA)

It is a generative model which assumes Gaussian distribution for each class. The fraction of data points that belong to the class is all that the class-specific prior entails. The class-specific mean vector represents the average of the class-specific input variables. The class-specific covariance matrix is nothing more than the covariance of the class's vectors.

2.10. AdaBoost Classifier

A meta-estimator called an AdaBoost classifier starts by fitting a classifier to the initial dataset. It then fits additional copies of the classifier to the same dataset, but with the weights of instances that were incorrectly classified being changed so that later classifiers would concentrate more on challenging cases. The first truly successful boosting algorithm created for binary classification was called AdaBoost. Multiple "weak classifiers" are combined into a single "strong classifier" using the boosting approach known as "AdaBoost", which stands for "Adaptive Boosting".

2.11. Neural Network (NN)

An artificial intelligence technique called a neural network instructs computers to analyze data in a manner modeled by human beings. Deep learning is a sort of machine learning that employs linked neurons or nodes in a layered framework to mimic the human brain.

2.12. k-5 Fold Cross-Validation

The k-fold cross-validation method is widely used for calculating how well a machine learning algorithm performs on a heart disease dataset. Though five is a typical choice for k, that is how this arrangement is acceptable for our heart disease dataset. The algorithms are investigating the impact of various k values on the estimated model performance and contrasting it with the ideal test condition. This can help in selecting the right value for k. Whenever a k-value is determined, it can be used to assess a variety of algorithms on the heart disease dataset. The distribution of the results may then be compared to the results of an assessment of the same algorithms under the optimum test scenario to see whether or not they are strongly correlated. If the results are correlated, then it is confirmed that the configuration chosen is reliable for the optimum test situation.

2.13. Stacking CV Classifier

An ensemble machine learning approach called stacking generalization is used. The goal of stacking is to improve the analytical results of individual machine learning models by ensembling the different machine learning models in the best possible way. The most basic form of stacking can be characterized as an ensemble learning technique where the predictions of various classifiers (referred to as level-one classifiers) are used as new features to train a meta-classifier. Any classifier of the choice may serve as the meta-classifier.

- **Model building:** A machine learning model is developed during this phase to detect cardiac illness. A risk assessment model is also developed to aid doctors in developing an early prediction with excellent predictive power. This is binary classification (has-disease or no-disease cases). The machine learning algorithms such as logistic regression, K-nearest neighbors, support vectors machine, Nu support vectors classifier, decision tree classifiers, random forest, adaboost, gradient boosting, naive bayes, linear discriminant analysis, quadratic discriminant analysis, neural network (12 classifiers to select the optimal base learners), and k-5 fold cross-validation. For avoiding the overfitting of the base learners, the random forest classifier is used as the Meta Learner. The stacking model is implemented for CVD prediction and is inspired by other ML models. The proposed technique is tested against 12 individual ML classifiers for accuracy, precision, recall, F1-score, and AUC values by using a combined heart disease dataset from multiple UCI machine learning resources and another publicly accessible heart disease dataset.
- The data are divided into training and testing sets in 3:1, optimizing the trained model 75% and the tested model 25%. As a result, the random forest and decision tree models are chosen as the top classifiers. However, the default settings were used to achieve this. With tweaked settings, we should be able to enhance our model even further. The working diagram of proposed model is shown in Figure 3 and application of proposed model is shown in Figure 4. The distribution of the target variable of heart disease is shown in Figure 5.

The significant contributions of this paper include heart disease empirical and statistical analysis utilizing various graphics, charts, and diagrams. The following are the keys taken away from the heart disease prognosis:

- **Data insight:** Heart disease detection using a dataset is presented, generating intriguing inferences to arrive at useful conclusions.
- **EDA:** Exploratory data analysis is carried out in order to obtain useful results.

The Kernel Density Estimate (KDE) design is used to visualize the Probability Density (PD) of a continuous variable. At various levels, it depicts the PD of a continuous variable. Moreover, numerous samples are combined into a single graph, which speeds up the data visualization. Numerical features are plotted by the Kernel Density Estimation (KDE) shown in Figure 6, the categorical features count is plotted in Figure 7, pair graphs of numerical features are plotted in Figure 8, regression plots of selected features are plotted in Figure 9, the numerical features correlation (Pearson's) is plotted in Figure 10, and the categorical features correlation (Cramer's V) is plotted in Figure 11. The Cleveland heart disease dataset and the KDE Plot are used to show the dataset's findings in this paper.

- **Feature development**: It is needed to change the features when it acquires the insights from the data so that it can continue forward with the model development process.

The remainder of the paper is arranged as follows: overview of present machine learning approaches was presented in Section 2. Section 3 outlines the study's objectives, whereas Section 4 gives the outlines of the techniques used. Section 5 gives details of data collecting, Section 6 presents experimental results and findings, and Section 7 wraps up the conclusions of the study and suggests further research.

3. Literature Review

The auscultation method was the most common method used by clinicians to distinguish between normal and abnormal heart sounds [19]. Doctors listening to the sounds of the heart with stethoscopes detected every cardiac ailment [20]. Professional doctors utilize auscultation to identify cardiac illness, although it has several disadvantages. The ability and practices of doctors, which are developed during extensive tests, are linked to the explanation and classification of various noises in the heart [21]. Aside from the manual technique, different machine learning algorithms for CVD detection have been presented. Amin et al. [22] did research to categorize the most important features of heart disease prediction. Spencer et al. [23] achieved an accuracy of 85%, the model was built using chi-square feature selection and the Bayes Net approach. Khan et al. [24] presented a new IOT framework based on deep convolutional neural networks. The system is coupled to a wearable detecting device that monitors the patient's vital signs. In terms of accuracy, our method outperforms existing DL neural networks and LR algorithms by 98.2%. Mehmood et al. [25] used Cardio Help which is a method that combines Neural and deep learning methods to use CNN for HF prediction and temporal model creation at an early stage.

Our technology outperforms all existing state-of-the-art techniques with a 97% accuracy rate. Budholiya et al. [26] used a strategy for recognizing essential cardiovascular disease risk factors that was developed by using a mean Fischer-based and an accuracy-based feature selection algorithm. The descriptive analysis was utilized to improve the selected feature subset, which was then used to classify the data using an RBF-based SVM. Martins et al. [27] used a Bayesian optimization XGBC, and a one-hot encoding technique is utilized by the researchers to predict heart disease. The Cleveland cardiovascular diseases dataset is used to assess the model's performance, and the results are comparable to other models. Miranda et al. [28] foresee this health hazard; they used the Naive Bayes algorithm and observed the risk levels for everyone. The test results of blood and urine of clinical laboratory tests served as the study of training datasets.

The problem with this study was that the authors deserted to see ECG and echocardiography studies, both of which are critical in diagnosing heart disease; hence the accuracy of 80% is regarded as poor. Because all of Naive Bayes' qualities must be mutually independent, it is difficult to use it to predict heart disease in practice due to the difficulty of generating a group of predictors that are totally independent of one another. Pandey et al. [29] developed a model for predicting cardiac disease to help doctors determine the severity of the condition. The J38 decision tree, which is based on a collection of clinical criteria, is used to categorize heart disease using the Cleveland Heart Disease dataset. Fasting blood glucose levels are the most important factor in heart disease, according to the model's findings. Mienye et al. [30] discovered a two-stage approach for forecasting cardiac disease that is successful. To begin, the researchers developed an enhanced sparse autoencoder (SAE), an unsupervised neural network that searches for the best description of the training data. Based on the learning records, they used an artificial neural network (ANN) to predict patient health.

The experimental findings produced utilizing the proposed approach improved the ANN classifier's performance. Siontis et al. [31] examined the current and future status of AI-assisted electrocardiography (ECG) in the diagnosis of heart disease in high-risk regions, as well as the implications for treatment decisions in patients with coronary heart disease. Anitha et al. [32] studied if learning vector quantization methods may be used to predict heart disease. The accuracy of her method was 85.55%. The datasets for her analysis are from the University of California, Irvine (UCI) machine learning collection, which had 303 entries and 76 attributes. The data was pre-processed due to missing values, providing a sample of 302 records with just 14 heart disease features. The data are separated into two categories: 70% of the expenditure goes to model training, whereas 30% goes to model testing. Kumar et al. [33] used a variety of machine learning methods for predicting if cardiovascular disease has been developed. According to the proposed model, RFC gives the highest accuracy of 85.71% when compared to other ML techniques. Chowdhury et al. [34] proposed that patients' heart sounds may be monitored in real time and any anomalies can be found using a digital stethoscope prototype. Negi et al. [35] used a mixture of uncorrelated discriminant analysis and PCA. The optimum features for regulating upper limb movements were determined, and the results were outstanding. Linda et al. [36] suggested that patients with heart disease who are advised to exercise benefit from a one-of-a-kind health information system. On the basis of previous data, clinicians are challenged to design the prescription of exercise using Mobile Information Systems for patients with a variety of CVD risk factors. The system offered to the patients is an evidence-based, easy-to-use, and time-saving guide. G. N. Ahmad et al. [37] applied machine learning techniques with and without a sequential feature selection and gave a comparative study of the optimal medical diagnosis of human heart disease.

4. Proposed Methodology

The major purpose of this study is to create a more accurate and better model for predicting heart illness. The specific goals are identifying new patients quickly, reducing diagnostic time, reducing heart attacks, and saving lives. In this Section, the techniques are suggested and describe how the future stages define them. As seen in Figures 3 and 4, the suggested technique is depicted as a block diagram.

(a) Choosing a dataset from the online machine learning repository is the initial step. The Cleveland, Switzerland, Hungarian, and Long Beach VA datasets are only a handful of the online repositories that provide 14 variables related to patients' vital signs and cardiac disease. With 13 out of the 14 variables working as predictor variables, the last trait acts as the target. Sex, age, kind of chest pain, serum cholesterol, resting blood pressure, fasting blood sugar, resting maximum heart rate, electrocardiography, and ST-segment elevation are all factors to consider. Segment elevation is one of the study's 13 predictor variables. Exercise-induced angina, depression, slope, and the outcome of a thallium test are all predicted, as are the number of vessels harmed by fluoroscopy and the diagnosis. The dataset has 1025 instances. Because there were no biases in the data, this strategy had no influence on the remaining data utilized in the experiment. Table 1 lists the datasets and their descriptions.

(b) Establishing links between various heart disease risk factors. The reciprocal link between the heart disease features in this study is assessed using Pearson's correlation. The results of the applied Pearson's correlation coefficients amongst the heart disease risk a variable which is shown in Figures 8 and 9 as a heatmap. The heatmap grid depicts the relationship among heart disease variables and their related coefficients. After completing the heatmap analysis, we noticed that independent characteristics are loosely associated with one another, which is a good indicator that the model's performance may be improved. However, if the attributes in a dataset are tightly correlated, a change in one variable can cause a change in another, lowering the algorithm's performance. The substantial association among qualities should be studied significantly because correlation does not imply causality. Because of some neglected elements, a link between qualities may appear causative through significant correlation.

(c) The gathered data sets were improved and standardized in the second stage. These datasets were not collected in a controlled setting and contained incorrect information. As a result, data pre-processing is a necessary step in data analysis and machine learning. The various values of the dataset of risk factors are referred to as data normalization; for instance, Celsius and Fahrenheit are two different temperature measuring units. Data standardization requires scaling risk factors and producing values that indicate the difference between standard deviations (SD) from the mean value. The performance of machine learning classifiers can be improved by rescaling the risk factor value with SD as 1 and mean as 0. The standardizing formula is as follows:

$$\text{Standardization of X} = \frac{\text{X} - \text{Mean of X}}{\text{Standard deviation X}} \quad (1)$$

(d) Applying machine algorithms to the dataset produced in step 2 (logistic regression, k-nearest neighbors, support sectors machine, nu support vectors classifier, decision tree, random forest, adaboost, gradient boosting, naive bayes, linear discriminant analysis, quadratic discriminant analysis, neural net, k-fold cross validation and ensemble technique).

(e) The prediction model's accuracy, precision, recall, and F-measure are all assessed at this step. It is decided on the model with the greatest accuracy, precision, recall, and F-measures. The accuracy metric is used to assess the precision or exactness of the MLC or model's predictions. In mathematics, it is supplied by an equation.

$$\text{Accuracy} = \frac{\text{true possitive}(\text{tp}) + \text{true negative}(\text{tn})}{\text{true possitive}(\text{tp}) + \text{true negative}(\text{tn}) + \text{false possitive}(\text{fp}) + \text{false negative}(\text{fn})} \tag{2}$$

$$\text{Precision} = \frac{\text{true possitive}(\text{tp})}{\text{true possitive}(\text{tp}) + \text{false possitive}(\text{fp})} \tag{3}$$

$$\text{Recall} = \frac{\text{true possitive}(\text{tp})}{\text{true possitive}(\text{tp}) + \text{false negative}(\text{fn})} \tag{4}$$

$$\text{F1} - \text{Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

Table 1. Details of the features.

S. No.	Attribute	Details
1	Age	Age in years
2	Sex	Male and female
3	Cp	Type of chest discomfort
4	Trestbps	Blood pressure at rest (in mm Hg on admission to the hospital)
5	Chol	Cholesterol levels in the blood in milligrams per decilitre
6	Fbs	After a fast, check your blood sugar levels.
7	Restecg	At rest, electrocardiography produces the following values. 0 indicates that the ST-T wave is normal; 1 indicates that the ST-T wave is aberrant;
8	Thalach	The highest heart rate was attained.
9	Exang	Angina due to exercise
10	Oldpeak	ST depression is a kind of depression that occurs when the J point is displaced below baseline
11	Slope	The slope of the ST half of the peak workout
12	Ca	A number of major vessels (0–3) have been colored using fluoroscopy.
13	Thal	The result of a thallium stress test
14	Target	1 denotes heart illness, while 0 denotes the absence of heart disease.

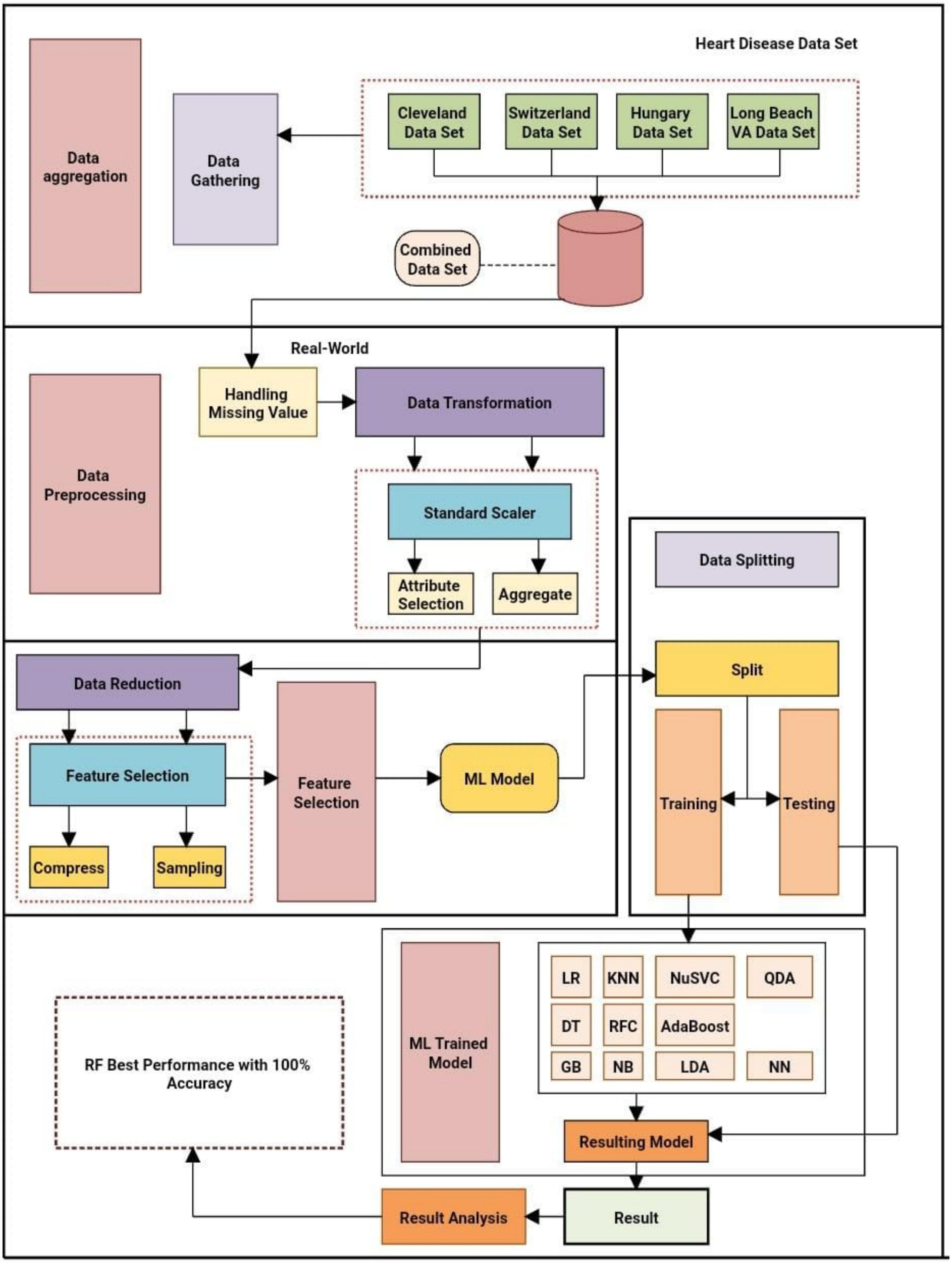

Figure 3. Working diagram of proposed model.

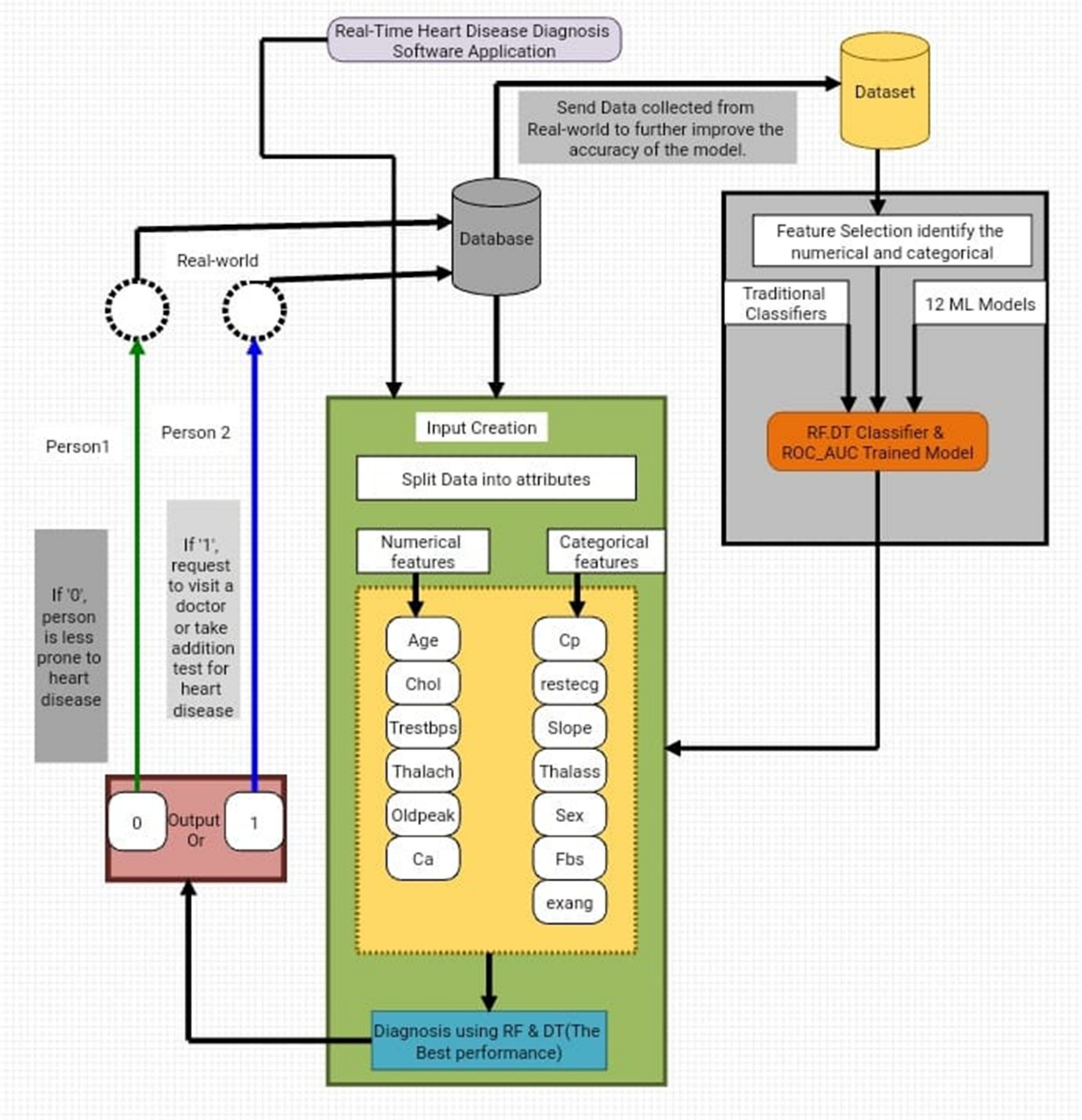
Real-Time Heart Disease Diagnosis Software Application
Dataset
Send Data collected from Real-world to further improve the accuracy of the model.
Database
Real-world
Feature Selection identify the numerical and categorical
Traditional Classifiers
12 ML Models
RF.DT Classifier & ROC_AUC Trained Model
Person1
Person 2
Input Creation
Split Data into attributes
Numerical features
Categorical features
If '0', person is less prone to heart disease
If '1', request to visit a doctor or take addition test for heart disease
Age
Chol
Trestbps
Thalach
Oldpeak
Ca
Cp
restecg
Slope
Thalass
Sex
Fbs
exang
0
Output Or
1
Diagnosis using RF & DT(The Best performance)

Figure 4. Application of proposed model.

5. Data Collection

Two databases dedicated to heart illnesses include the Cleveland database and the National Cardiovascular Disease Surveillance (NCDS) System's heart disease database. The Kaggle [38] heart disease dataset utilized in this work is made up of four combined databases: Cleveland, Hungary, Switzerland, and the VA Long Beach. The dataset has 14 properties, each of which has a value given to it. It has 1025 patient records spanning a wide range of ages, with 713 male and 312 female records among them [39] in a subset of this dataset.

The process of putting data into a visual framework, such as a map or graph, to make it easier for the human brain to analyze and extract insights from it is known as data visualization. Data visualization's major purpose is to make it simpler to spot cardiac disease, patterns, and outliers in massive data sets.

5.1. Checking the Distribution of the Data

The allotment of the data is decisive for predicting or classifying an issue. Recently, it was observed that 50.80% of the dataset featured the occurrence of heart disease, and 49.20% of the dataset featured no heart disease. For avoiding overfitting, the dataset must be balanced. As seen in Figure 5, this will aid the model in identifying a data trend that indicates the presence of heart disease versus one that does not.

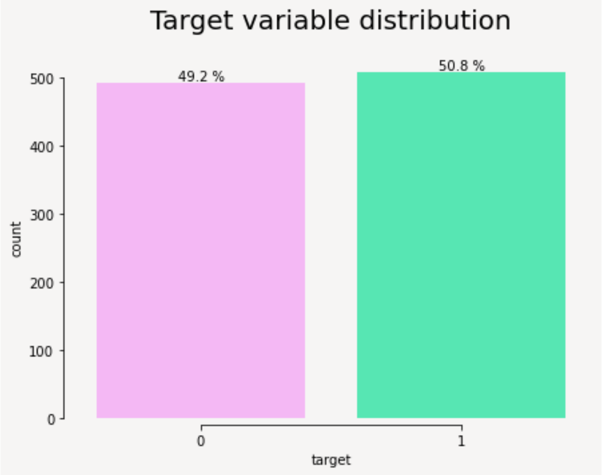

Figure 5. Heart disease target variable distribution.

5.2. Checking the Categorical and Numerical Features Pair Plots, Count Plots of the Data

To evaluate the data, many distribution plots are constructed to confirm attribute values and look at the data's skewness (the asymmetry of a distribution). To offer an overview of the data, a number of graphs are presented. Figures 5 and 6 depict the results of analyzing the distributions of age and sex, chest pain, and trestbps, cholesterol and fasting blood, ECG resting electrode and thalach, exang and old peak, slope and ca, and thal and goal. The numerical feature pair graphs are shown in Figures 7 and 8. A five-point biserial correlation and a point-biserial correlation are used to assess the strength and direction of a relationship between two continuous variables and one dichotomous variable. Pearson's correlation of numerical variables, the correlation between two points, is a point-biserial correlation that is used to determine the strength and direction of a relationship between two continuous variables and one dichotomous variable. It is a variant of Pearson's product-moment correlation, which is used to compare two continuous variables. The Cramer's V statistic is a statistical measure of the connection between two nominal variables that range from 0 to 1. Pearson's chi-squared statistic provides the foundation for it.

It is clear from the distribution plots that the data are overfitted or underfitted. It is shown in Figure 5 that the target variable distribution is 50.8% for abnormal heart disease and 49.2% for normal heart disease. The majority of abnormal heart patients are aged 55–65. This also seems to be a balance in disease patients who have experienced cholesterol and max-heart rate achieved.

Chest Pain (cp) is described in Figures 6 and 7. Heart disease is more common in those with a cp of 1, 2, or 3 compared to those with a cp of 0.

Resting electrocardiographic: People with value 1 are more likely to have abnormal cardiac disease, which can manifest as anything from minor symptoms to serious complications.

Exercise-induced angina: People with a value of 0 (no exercise-induced angina) have heart disease more than people with a value of 1 (exercise-induced angina positive)

Slope {the slope of the peak exercise ST-segment}: People with a slope value equal to 2 (down sloping: signs of an unhealthy heart) are more likely to have heart disease than people with a slope value equal to 0 (upsloping: better heart rate with exercise) or 1 (Flatsloping: minimal change (typical healthy heart)).

Thalassemia {thalium stress result}: People with a thal value equal to 2 (fixed defect: used to be defective but OK now) are more likely to have heart disease.

Trestbps: Resting blood pressure (in mm Hg on admission to the hospital). Anything above 130–140 is typically a cause for heart disease.

Chol {serum cholesterol in mg/dl}: The value of **Chol** above 200 is a cause for heart disease.

Thalach {maximum heart rate achieved}: People who achieved a **thalach** value higher than 150-160 are more likely to have heart disease. Oldpeak ST depression induced by exercise relative to rest looks at the stress of the heart during exercise. An unhealthy heart will stress more.

CA {number of major vessels (0–3) colour by fluoroscopy}: The high movement of the blood in the blood vessels is the reason for good health, and people with the lowest value of CA, which is equal to 0, are more likely to have heart disease.

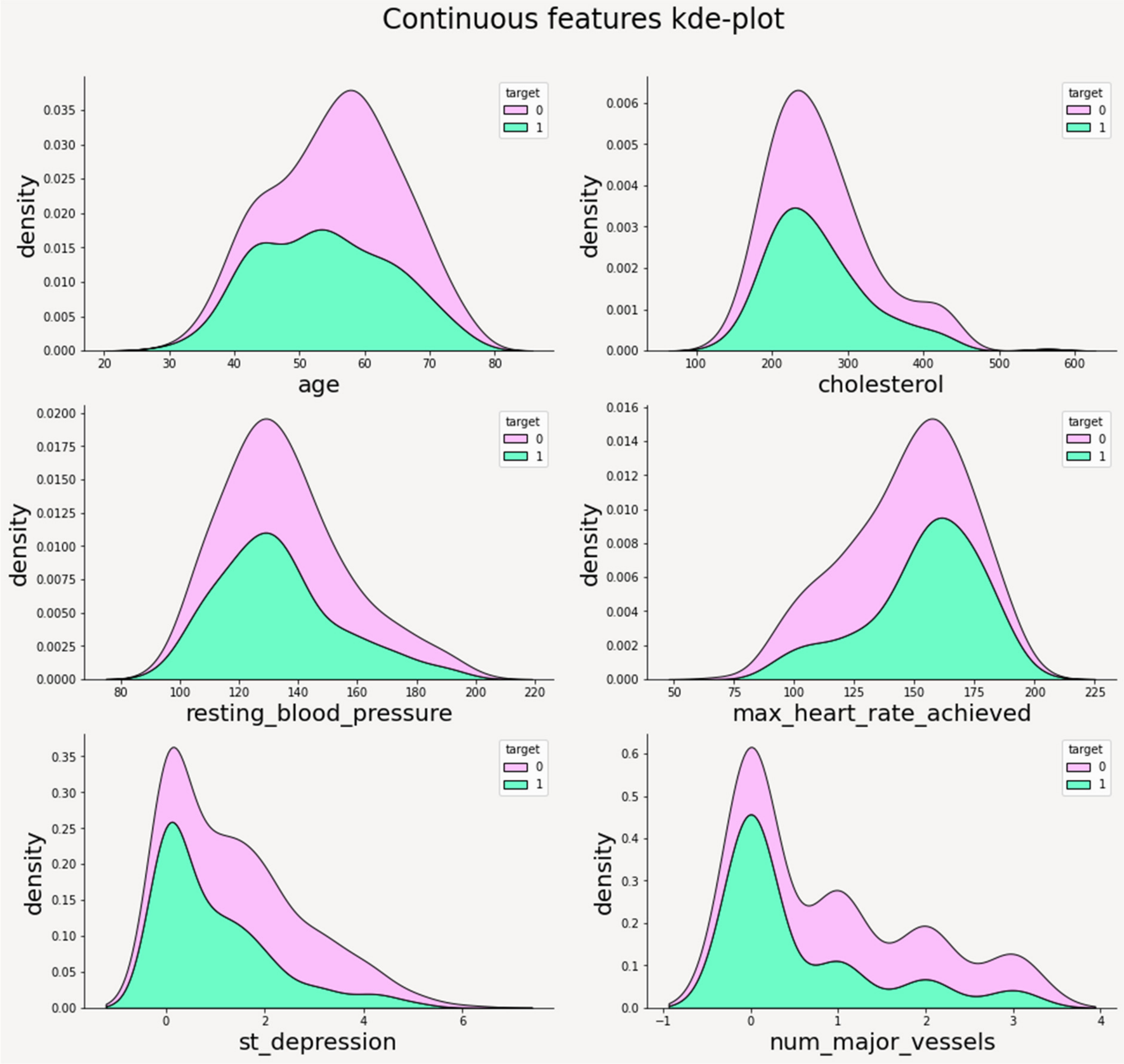

Figure 6. Continuous feature Kernel Density Estimate Plot.

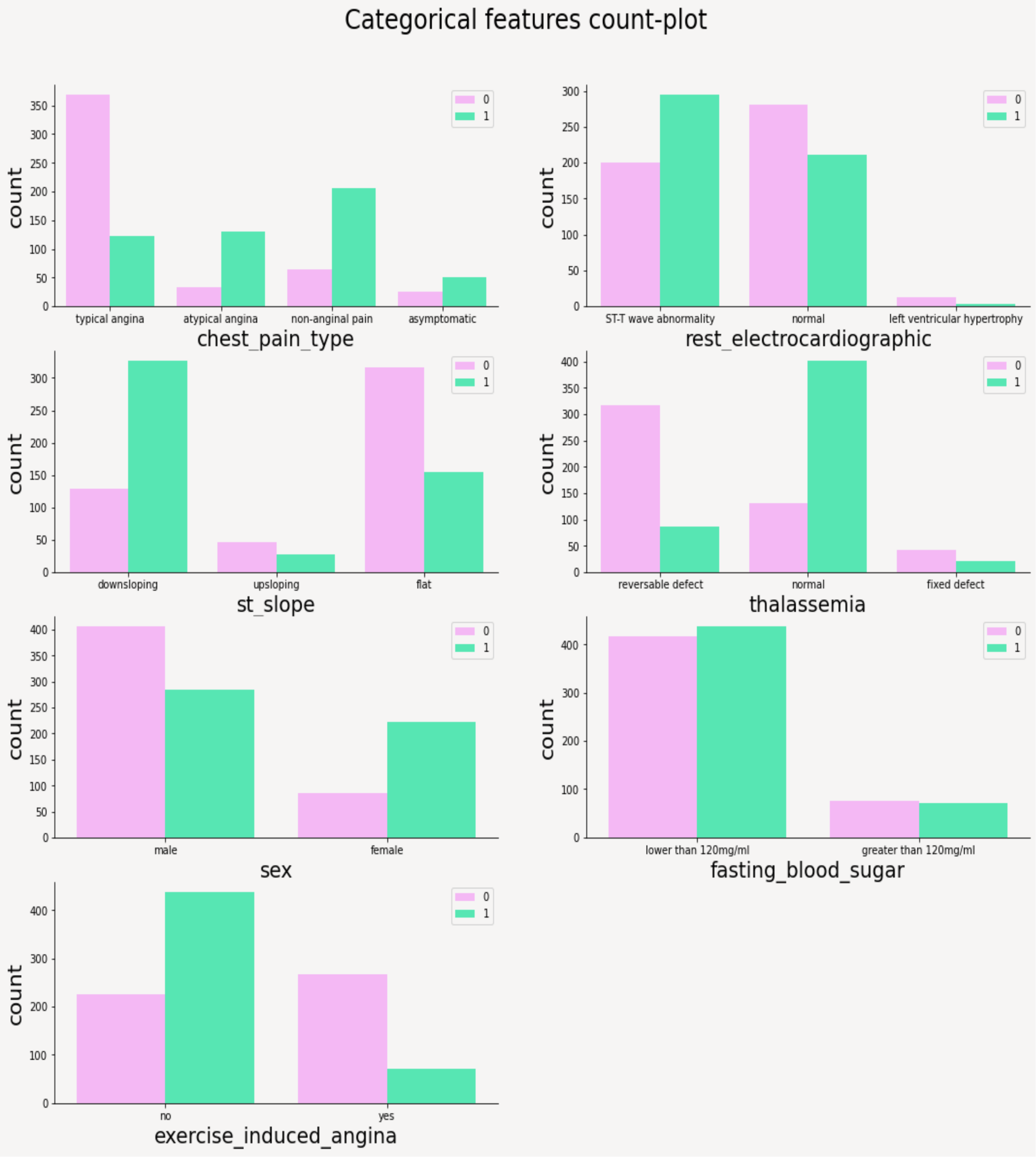

Figure 7. Categorical features count plots of the Data.

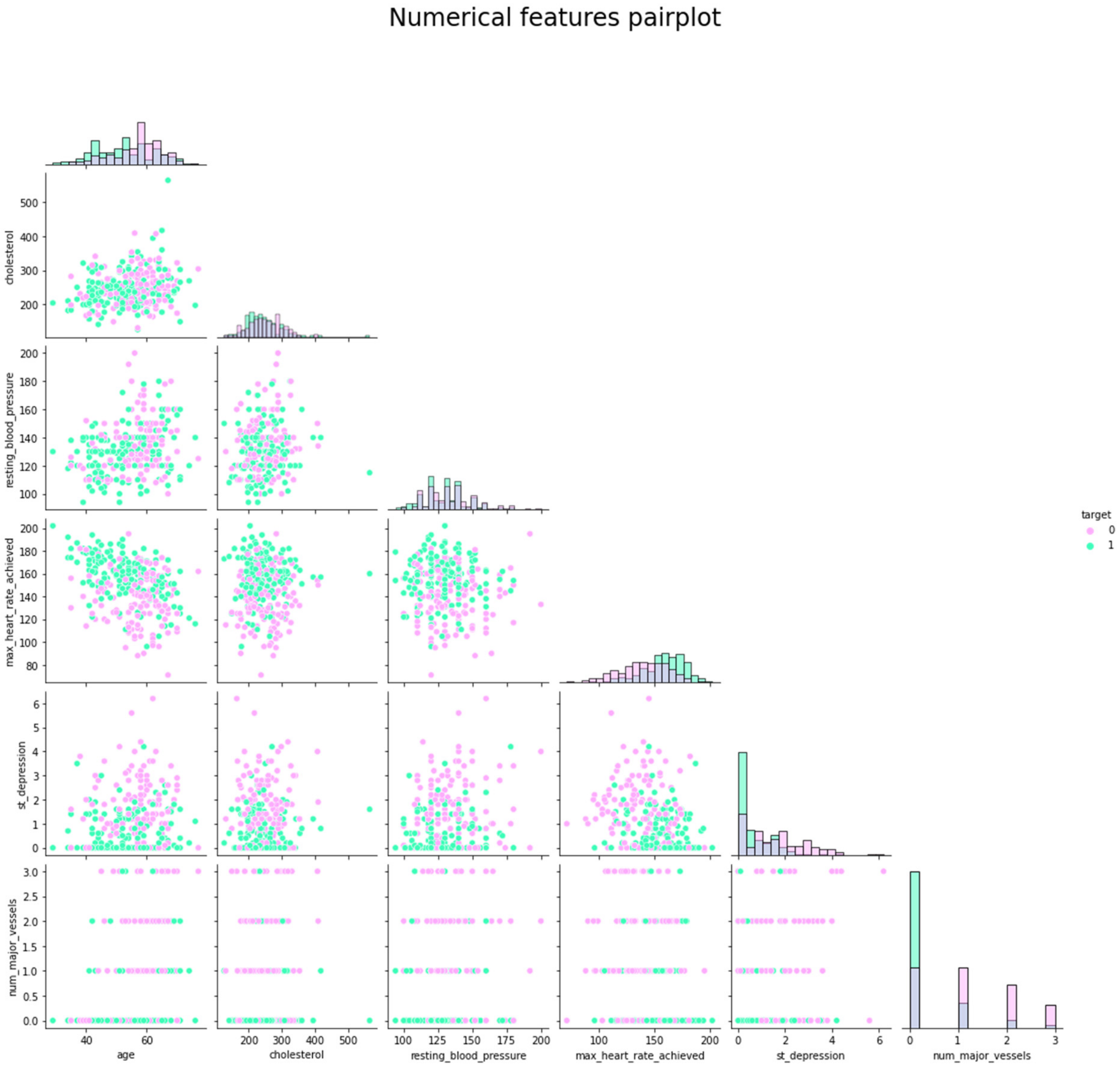

Figure 8. Numerical feature pair plots.

It is shown in the given Figures 8–10 that as the goal varies, chest pain or angina (cp), fluctuates. Angina is a serious chest pain condition caused by the heart muscles not receiving enough oxygen-rich blood. Pain in the shoulders, arms, and neck is also caused by angina. While negative patients have lower levels, positive patients have an elevated median for ST depression. The objectives of the outcomes for men and women are also similar, with the exception that men tend to have slightly wider ranges of ST depression.

The heatmaps are shown in Figures 10 and 11. These may be defined as the illustrations of correlation matrices, which illustrate the connections between various variables. The correlation coefficients might have any value between −1 and 1. A correlation is a statistical word describing a connection when two variables are linearly related. It is also known as a measure of correlation between two variables. In this case, the objective is to organize the findings after determining a correlation between several factors. Information was stored in this case using a matrix data structure. Figures 10 and 11 illustrate this feature-by-feature. Many facts are provided by the figures. First, with correlations of 0.2, 0.28, −0.39, −0.27, and 0.37, respectively, the five variables with the greatest class-feature dependency are cholesterol, resting blood pressure, maximum heart rate, ST depression, and the number of main arteries. The second fact refers to the feature correlations between the variables cholesterol age, resting blood pressure and cholesterol, ST depression and maximum heart rate achieved, and the number of major blood vessels and maximum heart rate achieved, with correlations of 0.2, 0.13, −0.35, and −0.27, respectively.

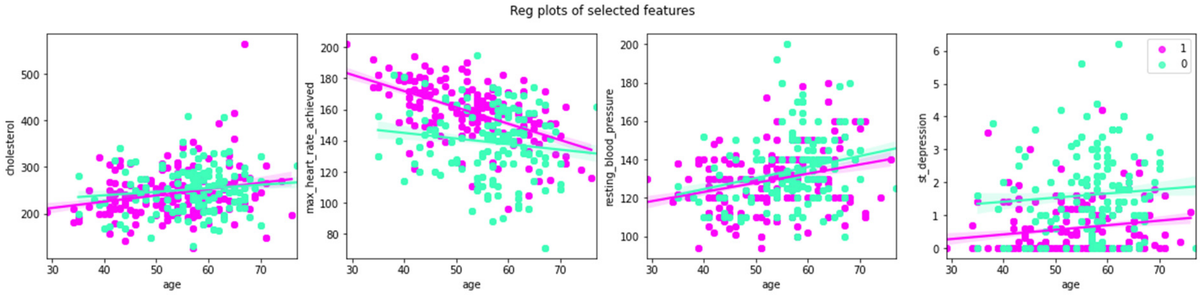

Figure 9. Regression plots of selected features.

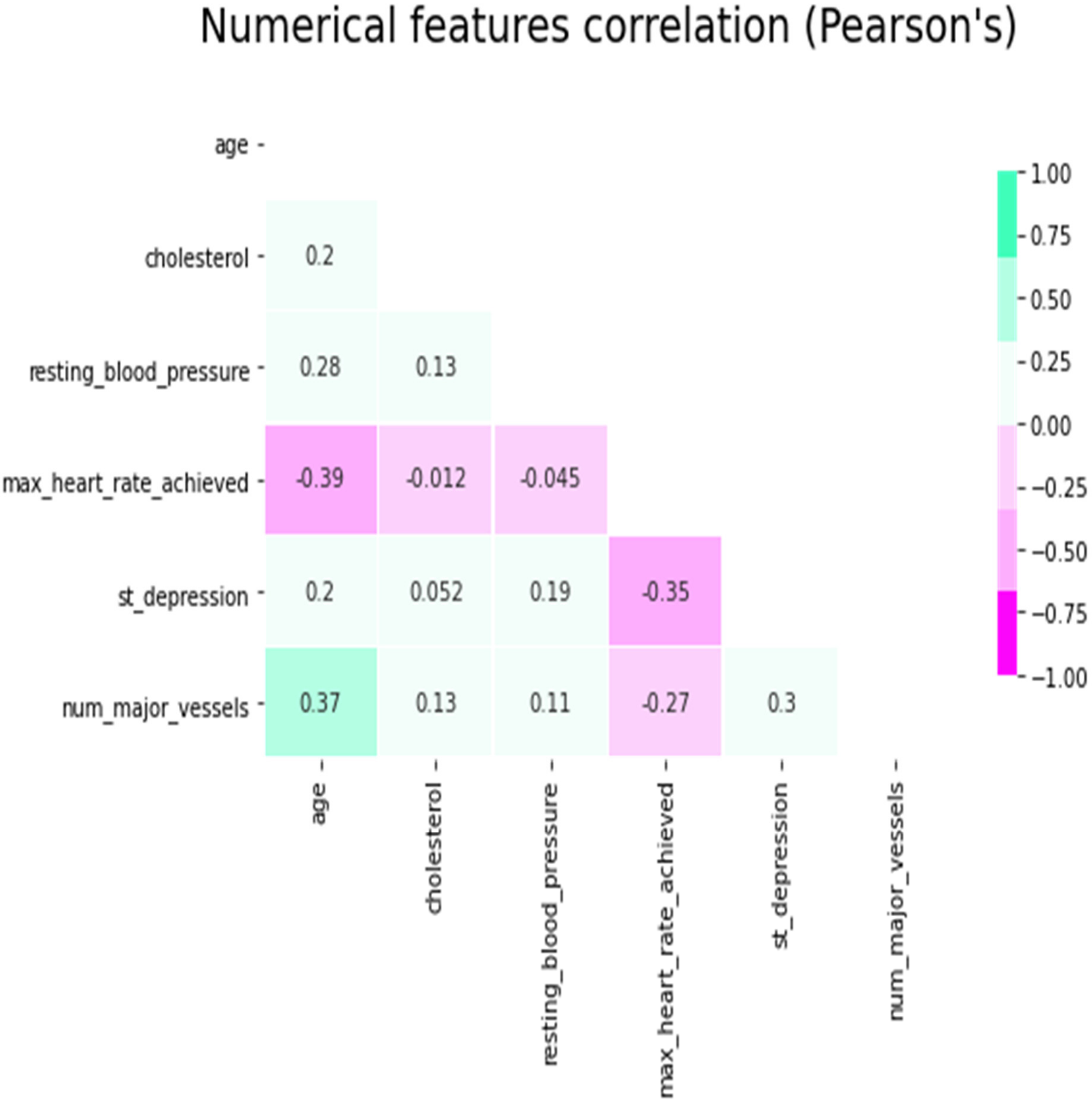

Figure 10. Numerical features correlation (Pearson's).

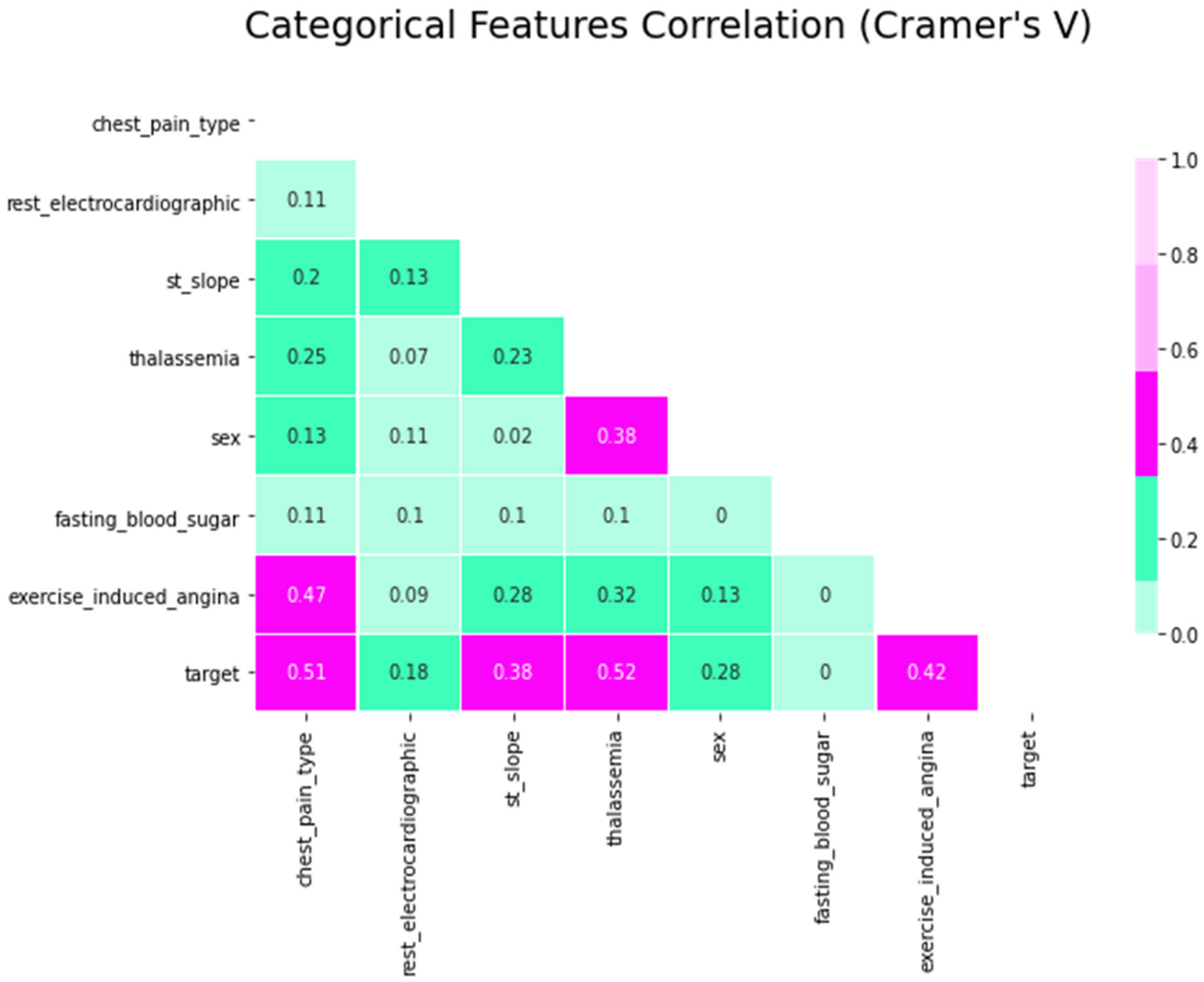

Figure 11. Categorical features correlation (Cramer'sV).

6. Experimental Results and Discussion

In this section, the findings of our experiments are given. The online machine learning repository is used to gather the data, which is then polished and standardized. Following standardization, hyperparameter and machine classifiers are used. All of the classifiers are made up of 75% training data and 25% testing data. Before and after standardized datasets, the accuracy of classifiers is also examined. The accuracy of the selected classifiers is shown for assessment purposes. Figure 11 depicts the accuracy of classifiers before and after the standardization of the data. Actually, the majorities of listed algorithms are not based on neural networks (LR, KNN, SVM, Nu SVC, DT, RFC, AdaBoost, GBC, NB, LDA, Q DA, NN, k-fold cross validation and ensemble technique) and increased their accuracy, as shown in Figure 8. On the standardized dataset, the accuracy of Naive Bayes and Support Vector Machine classifiers fell. On the standardized dataset, certain classifiers, including RF, DT, GB, and NN, exhibited considerable accuracy gains. The RF and DTC reach the maximum prediction accuracy of 100% and 98.80%, respectively, as shown in Figure 10. The Support Vector Machine has the lowest overall presentation and the lowermost accuracy of 68.80%. We also look at the dataset's correctness before and after normalization. The Random Forest and Decision Tree classifiers reach an accuracy of 100% and 98.80%, respectively, demonstrating the influence of the dataset normalized.

Figures 12 and 13 and Table 2 are for positive classes. The random forest classifier achieves 100% recall, precision, F-measure, and accuracy, whereas, for negative classes, the decision tree classifier achieves 100% recall, precision, F-measure, and accuracy. However, random forest has the best memory, precision, F-measure, and accuracy of 100%. All classifiers have a precision of 100%, whereas all classifiers have a recall better than any other which is of 100%. Random forest and decision tree both have a maximum F1-score of 100%. SVM has a low recall, precision, and F1-score of 69% and it has the lowest recall, F-measure, and accuracy of 69%, and random forest and the gradient boosting classifier achieve a maximum ROC-AUC of 100%. As a result, the negative class had a ROC-AUC of 73%, which was much lower than the positive class. Random forest has a fairly good performance, with 100% recall, precision, F-measure, and ROC-AUC, as well as 100% accuracy.

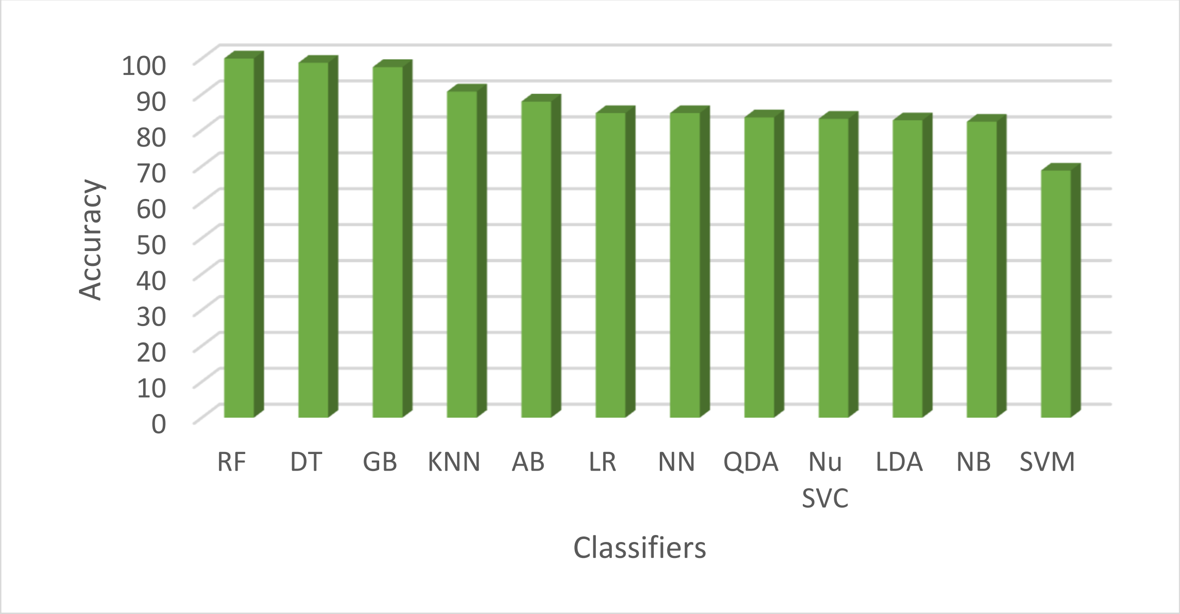

Figure 12. Classifiers Accuracy.

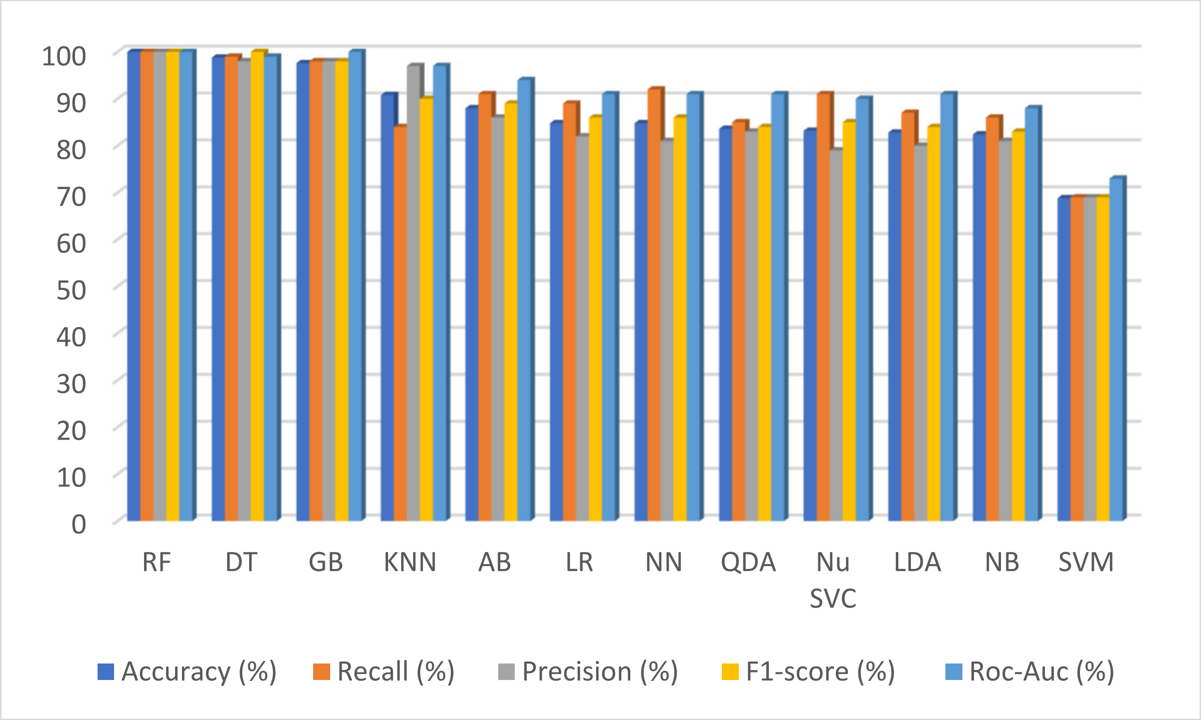

Figure 13. Results of different machine learning models.

Table 2. Results of different machine learning models.

S. No.	Classifier	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)	ROC-AUC (%)
1	RF	100	100	100	100	100
2	DT	98.80	99	98	100	99
3	GB	97.60	98	98	98	100
4	KNN	90.80	84	97	90	97
5	AB	88.00	91	86	89	94
6	LR	84.80	89	82	86	91
7	NN	84.80	92	81	86	91
8	QDA	83.60	85	83	84	91
9	Nu SVC	83.20	91	79	85	90
10	LDA	82.80	87	80	84	91
11	NB	82.40	86	81	83	88
12	SVM	68.80	69	69	69	73

As shown in Figures 14 and 15 and Table 3, the decision tree classifier achieved the highest accuracy of 99.60 and precision of 100%, the random forest classifier achieved 100% recall, and the gradient boosting classifier achieved 100% ROC-AUC. NB has a low accuracy of 82.15%, precision of 86.60%, F1-score of 83.12%, and ROC-AUC of 90.67%.

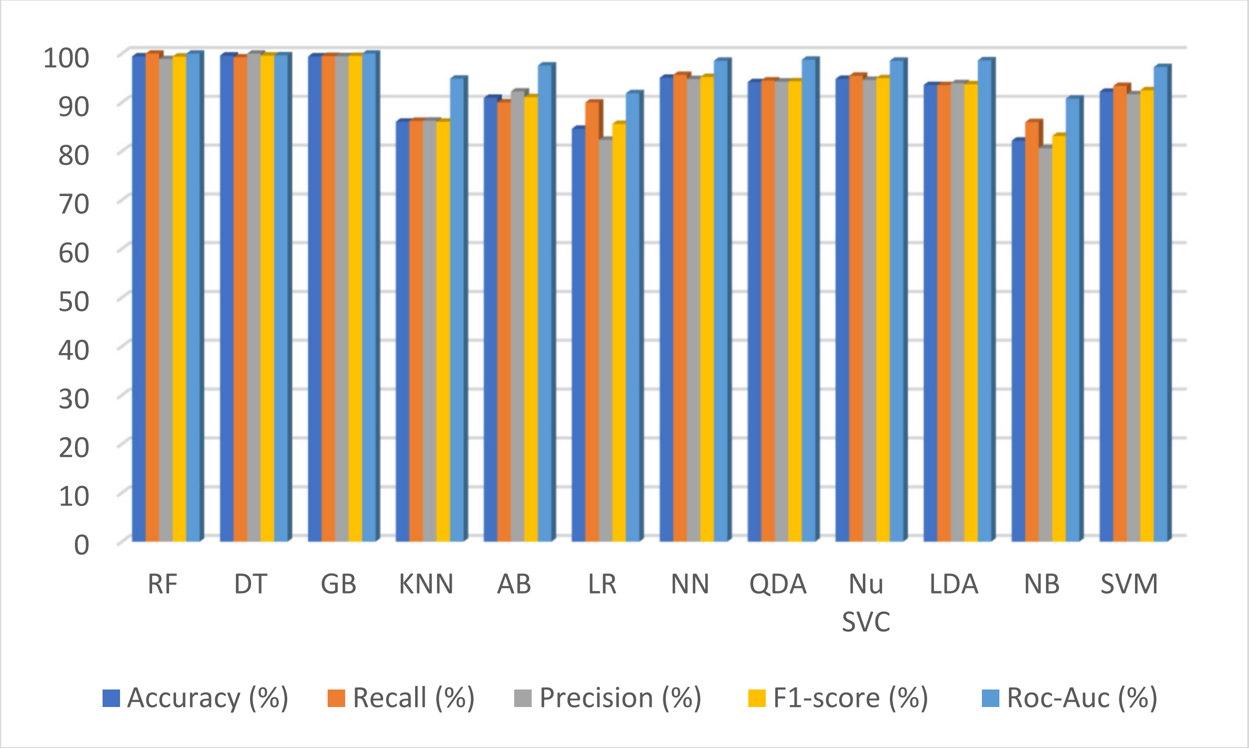

Figure 14. Performance of ML classifier using 5-fold cross-validation.

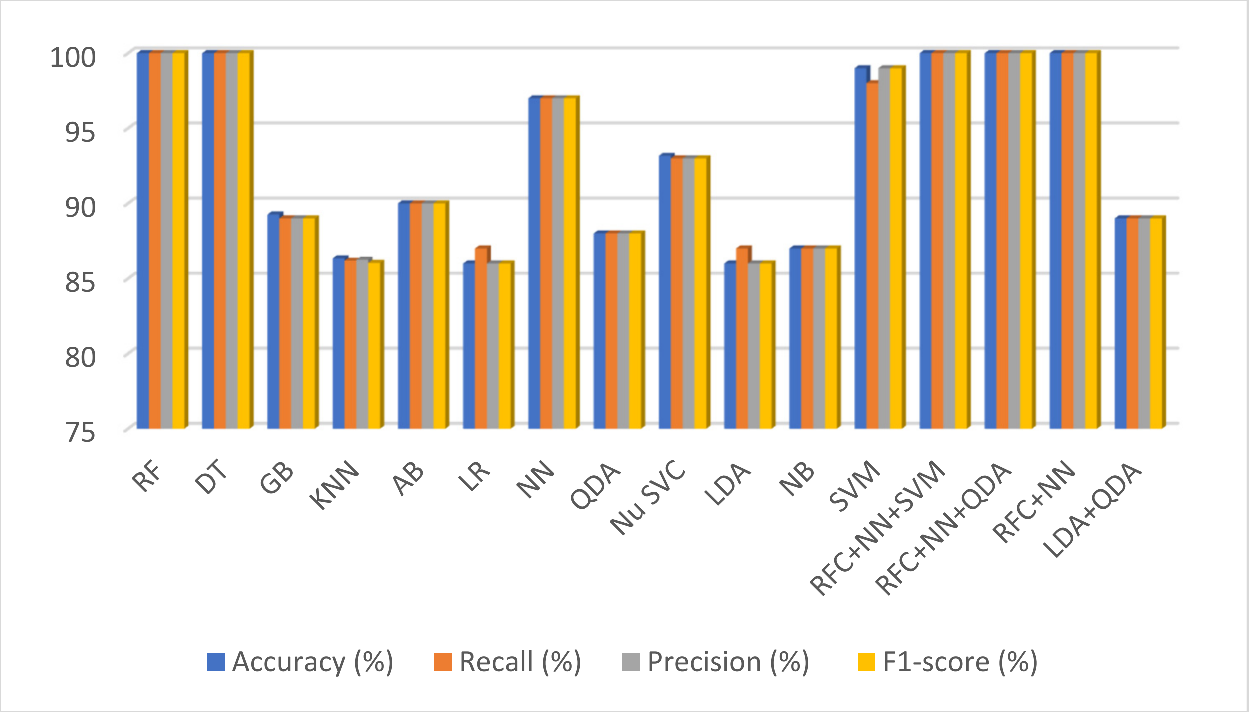

Figure 15. Performance of ML classifier using StakingCV Classifier.

Table 3. Performance of ML classifier using 5-fold cross-validation.

S. No.	Classifier	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)	ROC-AUC (%)
1	RF	99.41	100	98.89	99.40	99.97
2	DT	99.60	99.23	100	99.61	99.62
3	GB	99.41	99.50	99.45	99.53	100
4	KNN	86.06	86.19	86.25	86.04	94.86
5	AB	90.93	89.99	92.24	91.05	97.57
6	LR	84.59	89.98	82.31	85.59	91.88
7	NN	95.03	95.64	94.76	95.19	98.54
8	QDA	94.16	94.49	94.27	94.33	98.74
9	Nu SVC	94.83	95.45	94.60	94.97	98.54
10	LDA	93.56	93.54	93.93	93.71	98.60
11	NB	82.15	85.94	80.63	83.12	90.76
12	SVM	92.19	93.36	91.68	92.49	97.24

The confusion matrix of an experiment's expected outcomes is depicted in Figure 16. We also have analyses of the model's performance using ROC-AUC. The receiver operating characteristic curve (ROC) is a curve with a true positive rate on the ordinate and a false positive rate on the abscissa [40]. It is created from a series of varied boundary values. The AUC, or area under the ROC curve, indicates the likelihood that the positive sample's calculated score will be greater than the negative sample's computed value. When the samples are chosen at random, the advantages and disadvantages of the prediction model may be investigated. As illustrated in Figure 17, which also illustrates an experiment's ROC curve, our model's average AUC value is 100%.

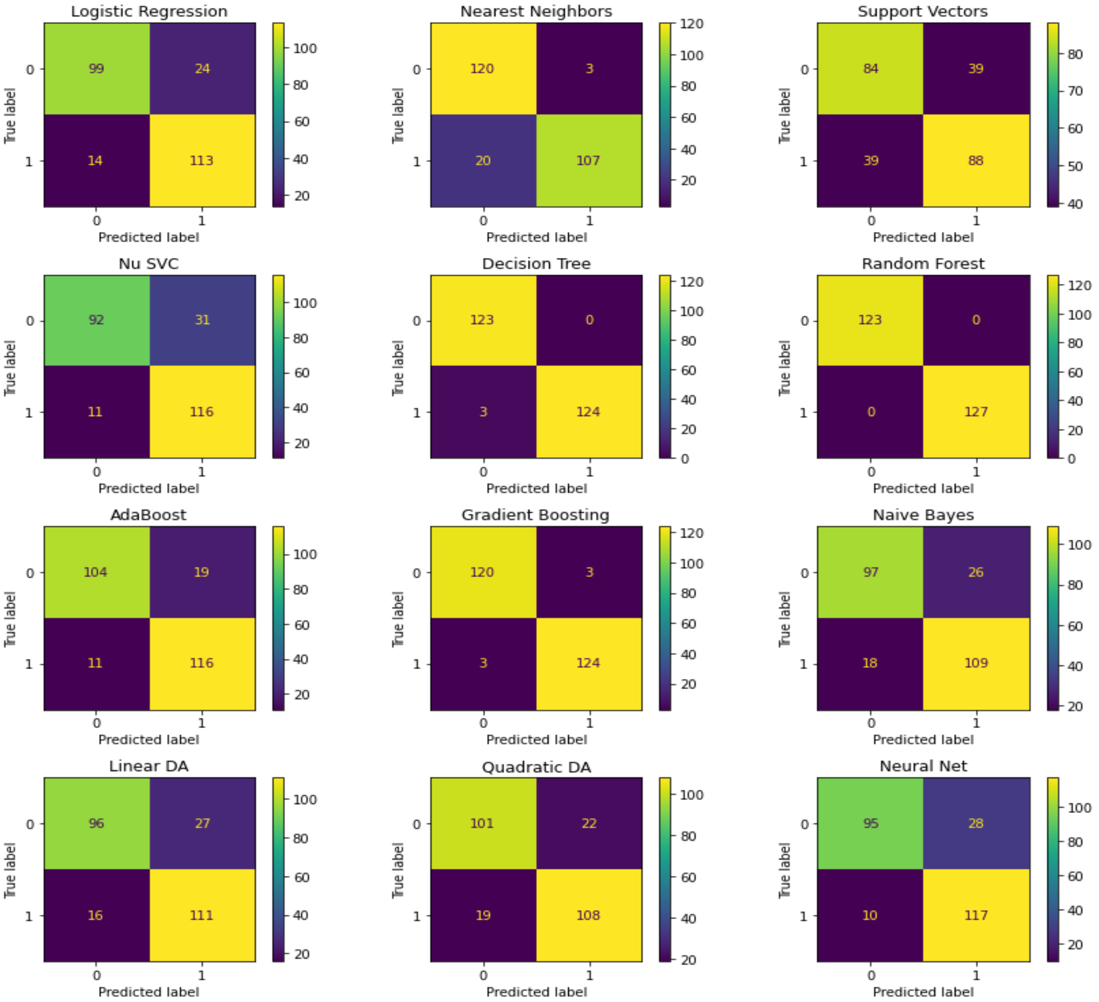

Figure 16. The confusion matrix of the predicted results.

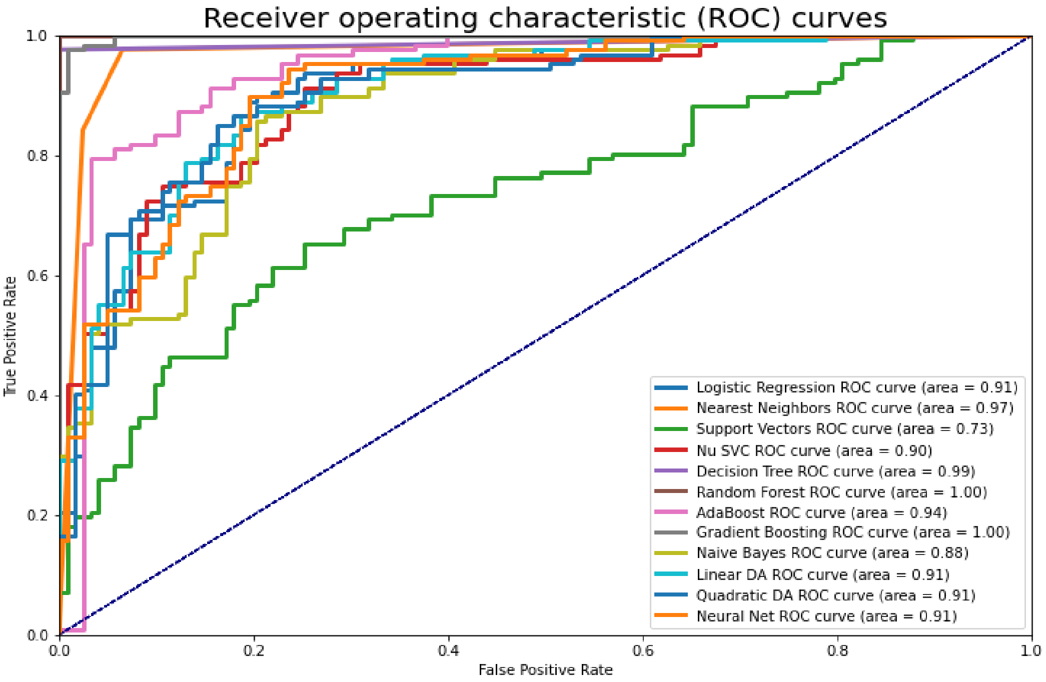

Figure 17. Combine plots for ROC-AUC.

As shown in Figures 18–29, the gradient boosting classifier achieved 100% ROC-AUC. NB has a low ROC-AUC of 90.67%.

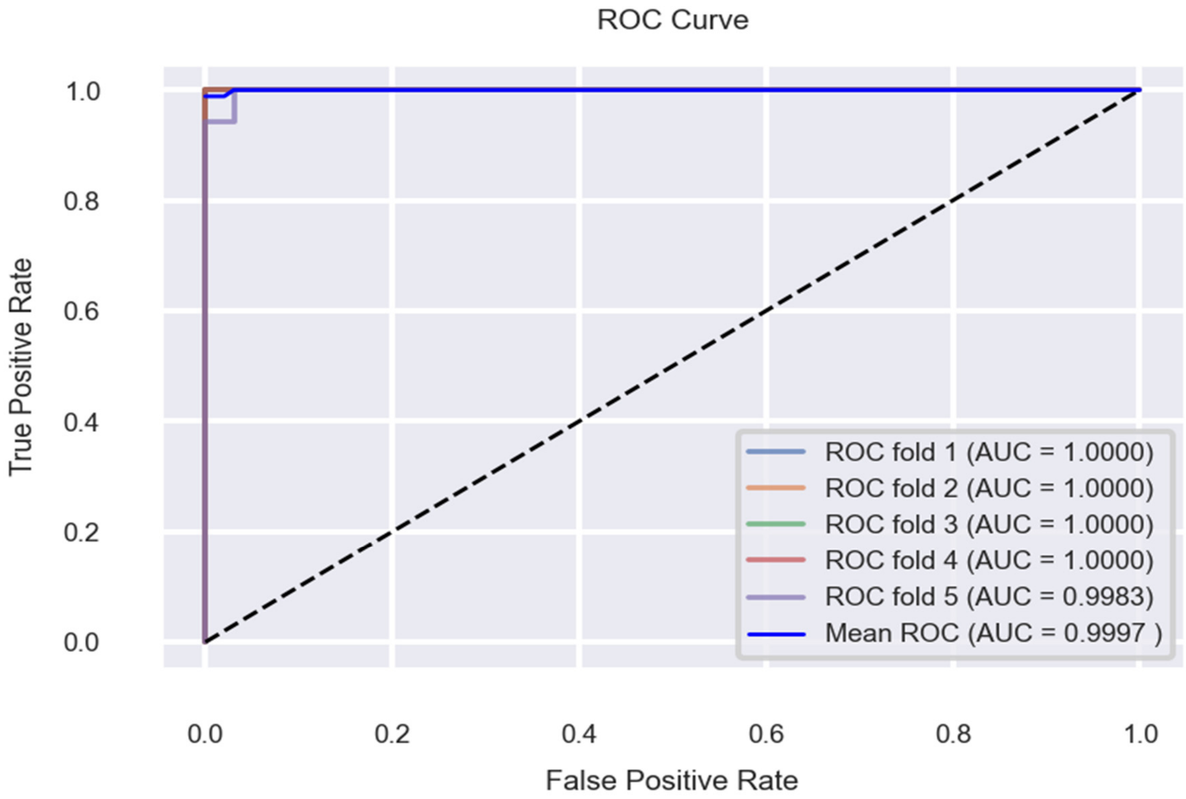

Figure 18. Stratified (K = 5)-Fold Cross-validation Results for Random Forest Classifier (ROC-AUC).

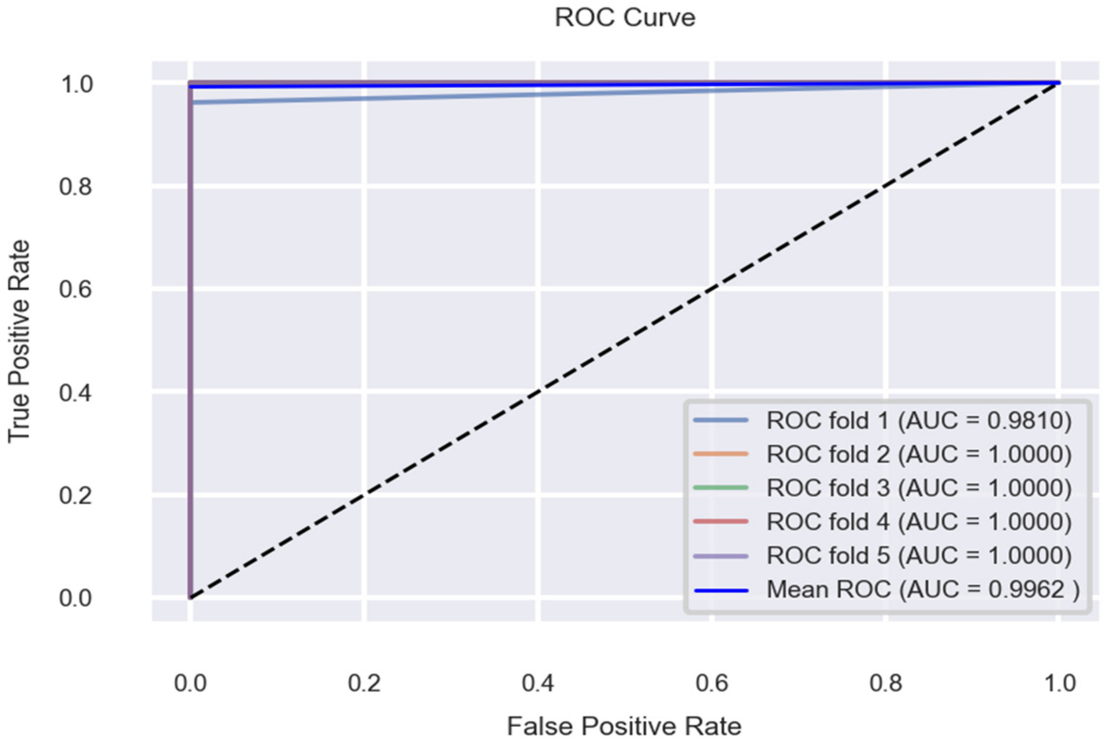

Figure 19. Stratified (K = 5)-Fold Cross-validation Results for Decision Tree Classifier (ROC-AUC).

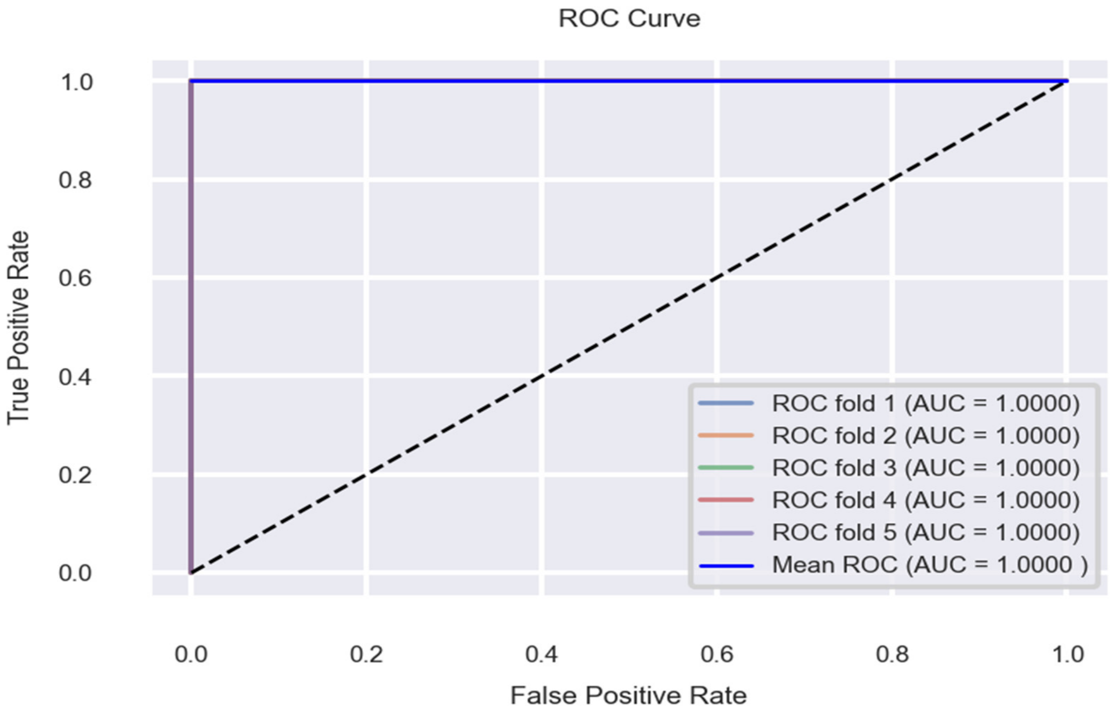

Figure 20. Stratified (K = 5)-Fold Cross-validation Results for Gradient boosting Classifier (ROC-AUC).

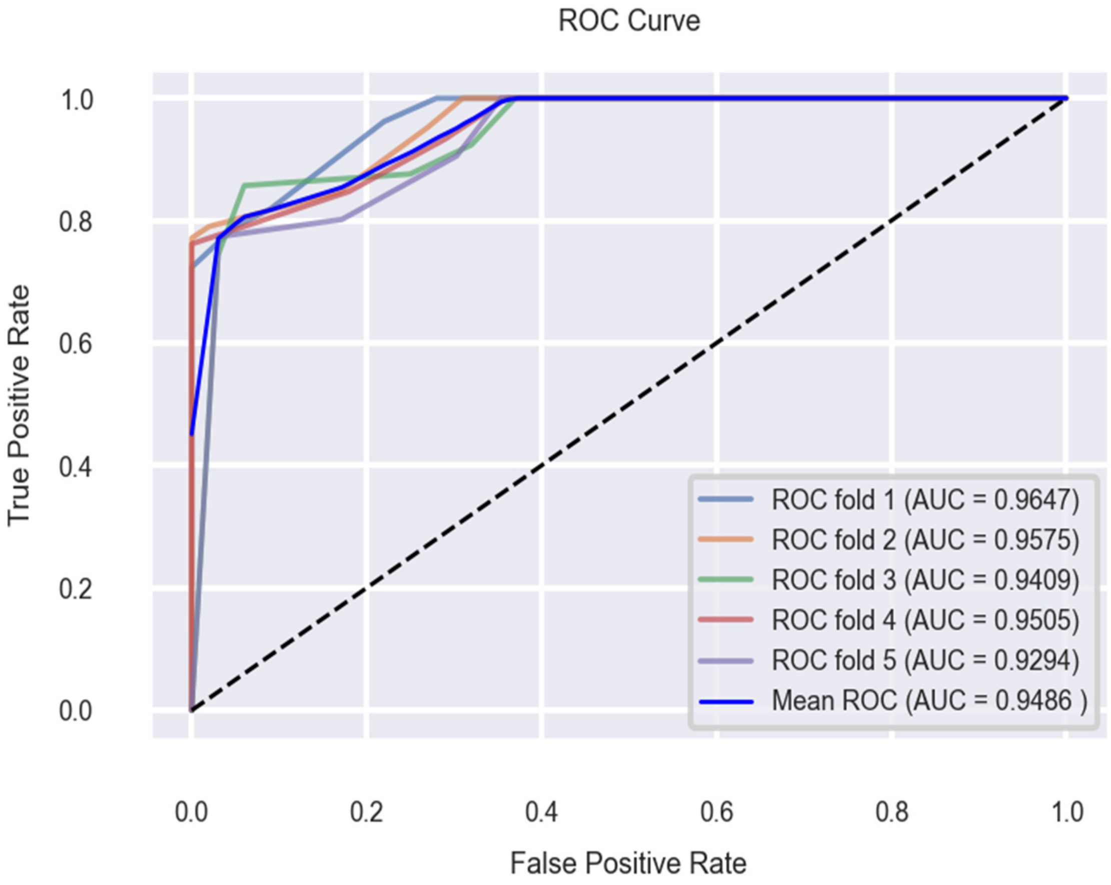

Figure 21. Stratified (K = 5)-Fold Cross-validation Results for KNN (ROC-AUC).

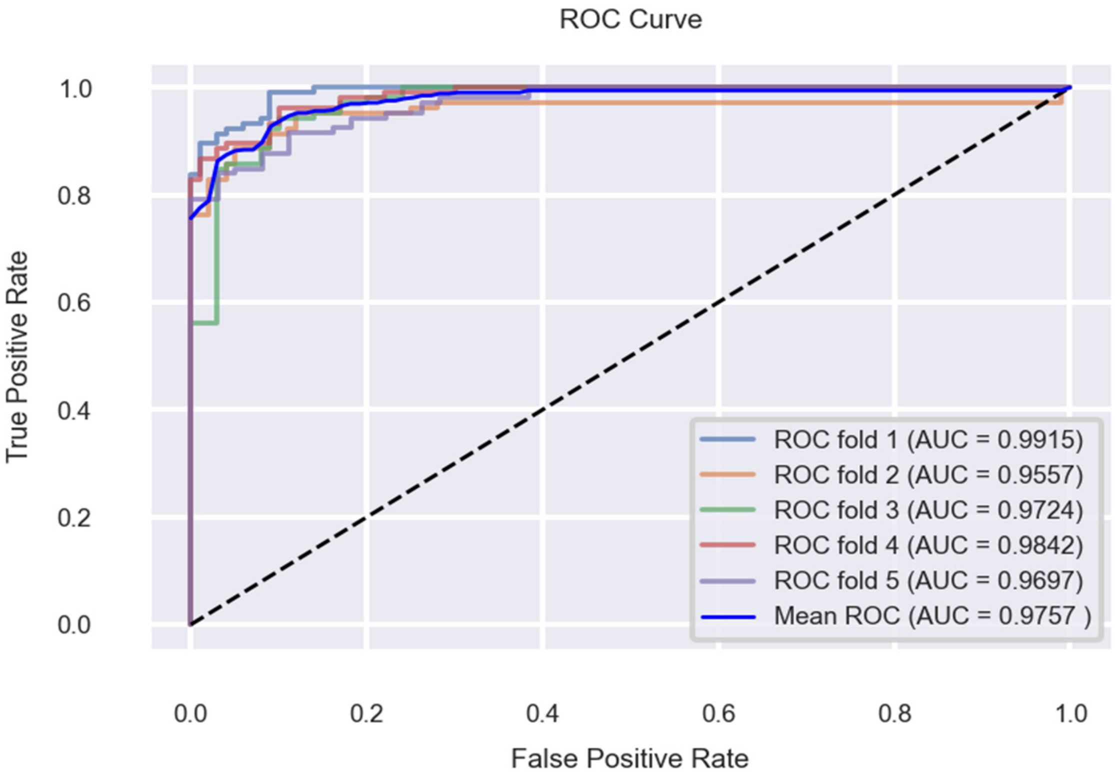

Figure 22. Stratified (K = 5)-Fold Cross-validation Results for Ada Boost Classifier (ROC-AUC).

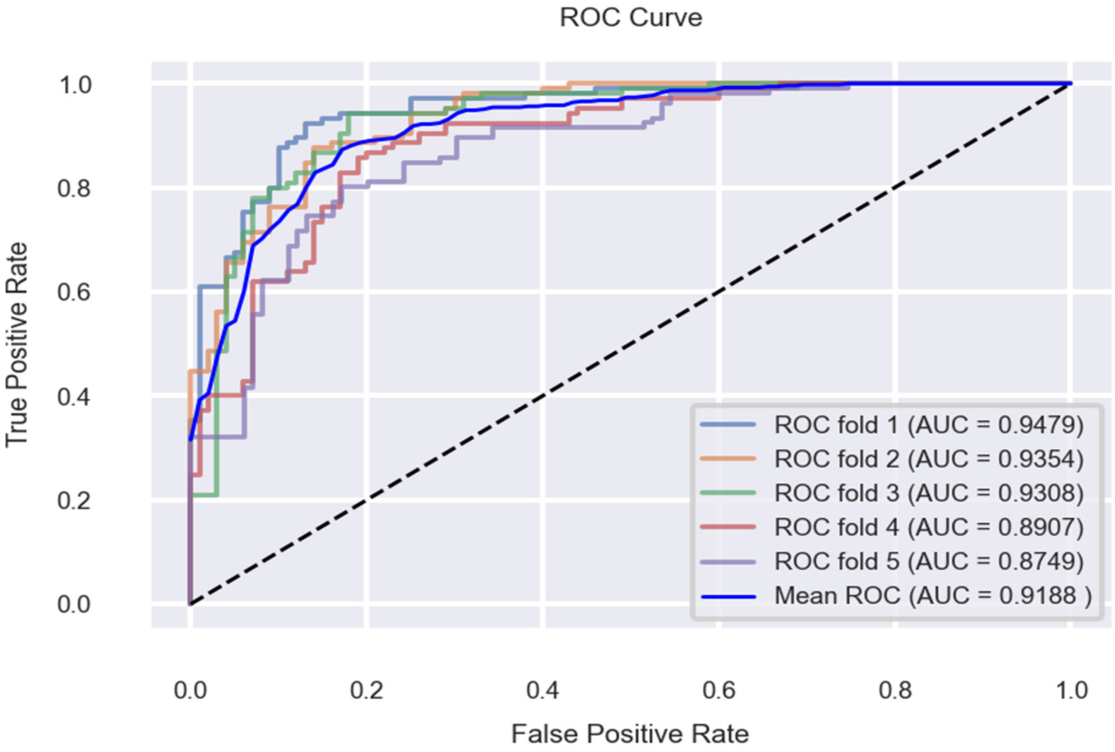

Figure 23. Stratified (K = 5)-Fold Cross-validation Results for Logistic Regression (ROC-AUC).

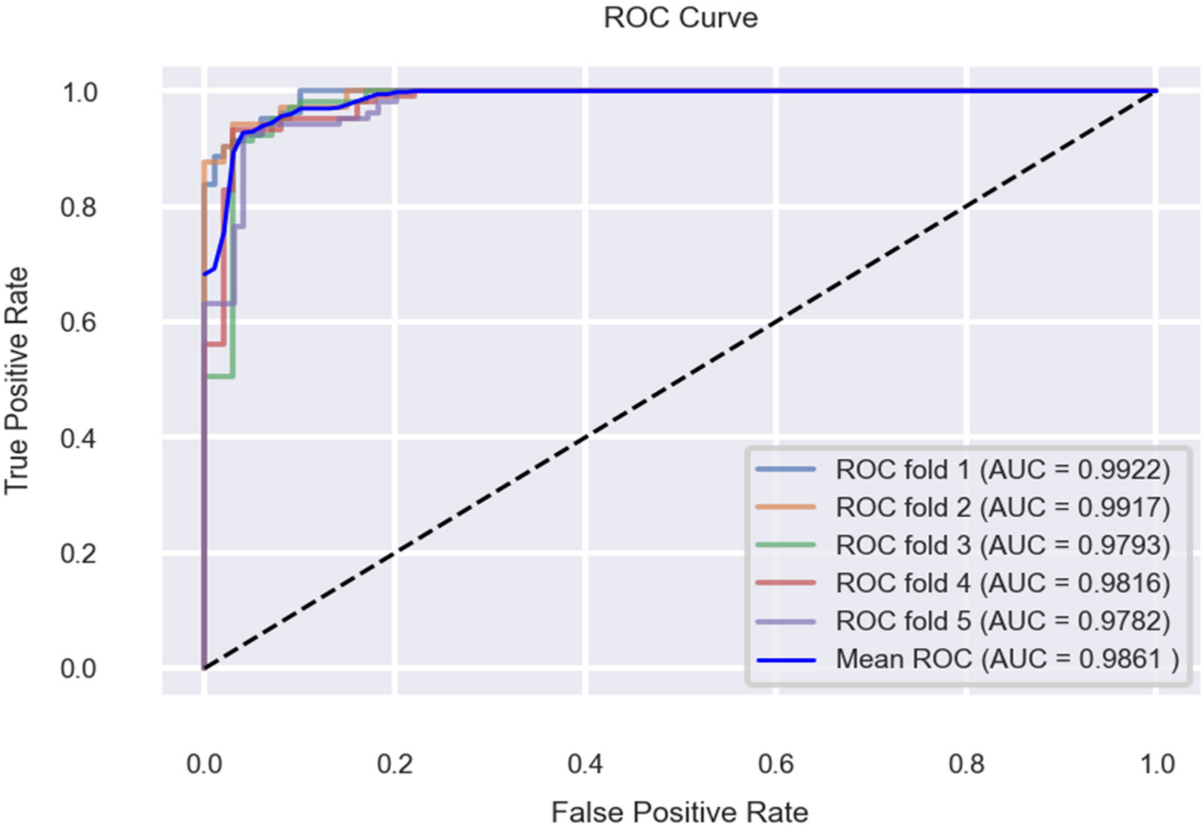

Figure 24. Stratified (K = 5)-Fold Cross-validation Results for Artificial neural network (ROC-AUC).

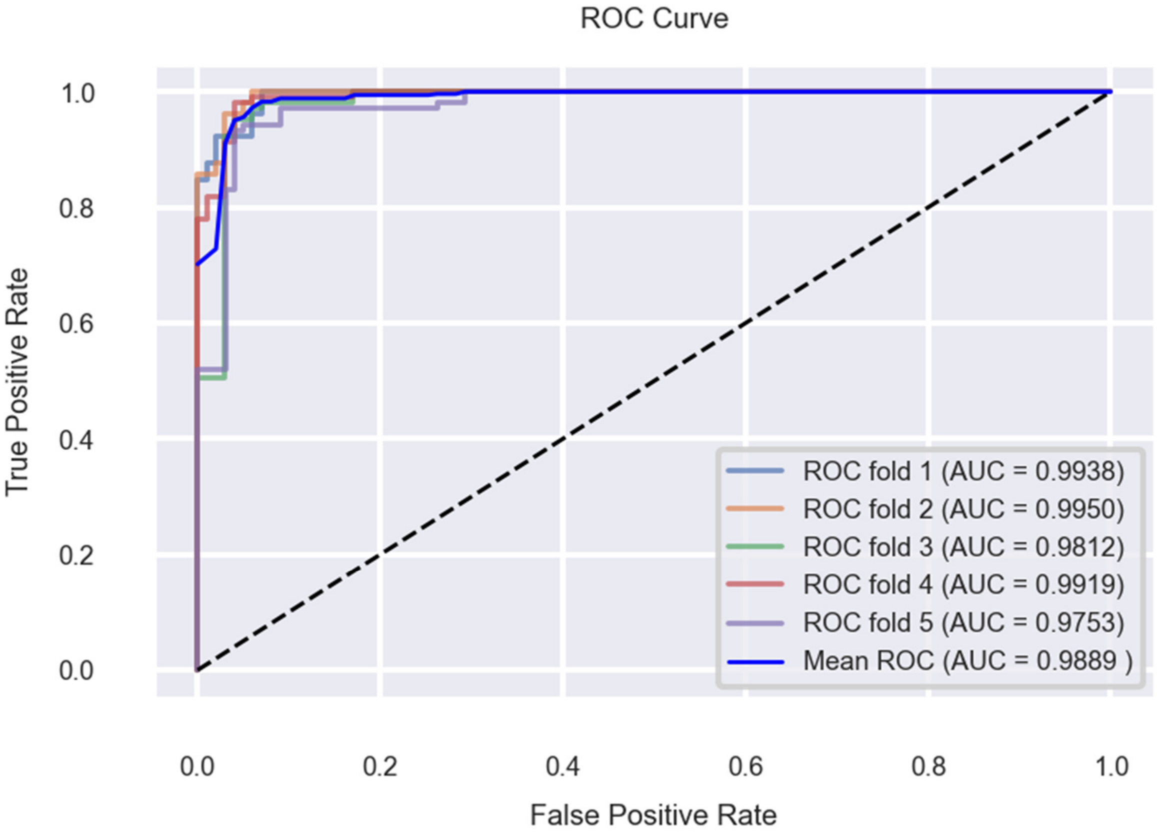

Figure 25. Stratified (K = 5)-Fold Cross-validation Results for Quadratic Discriminant Analysis (ROC-AUC).

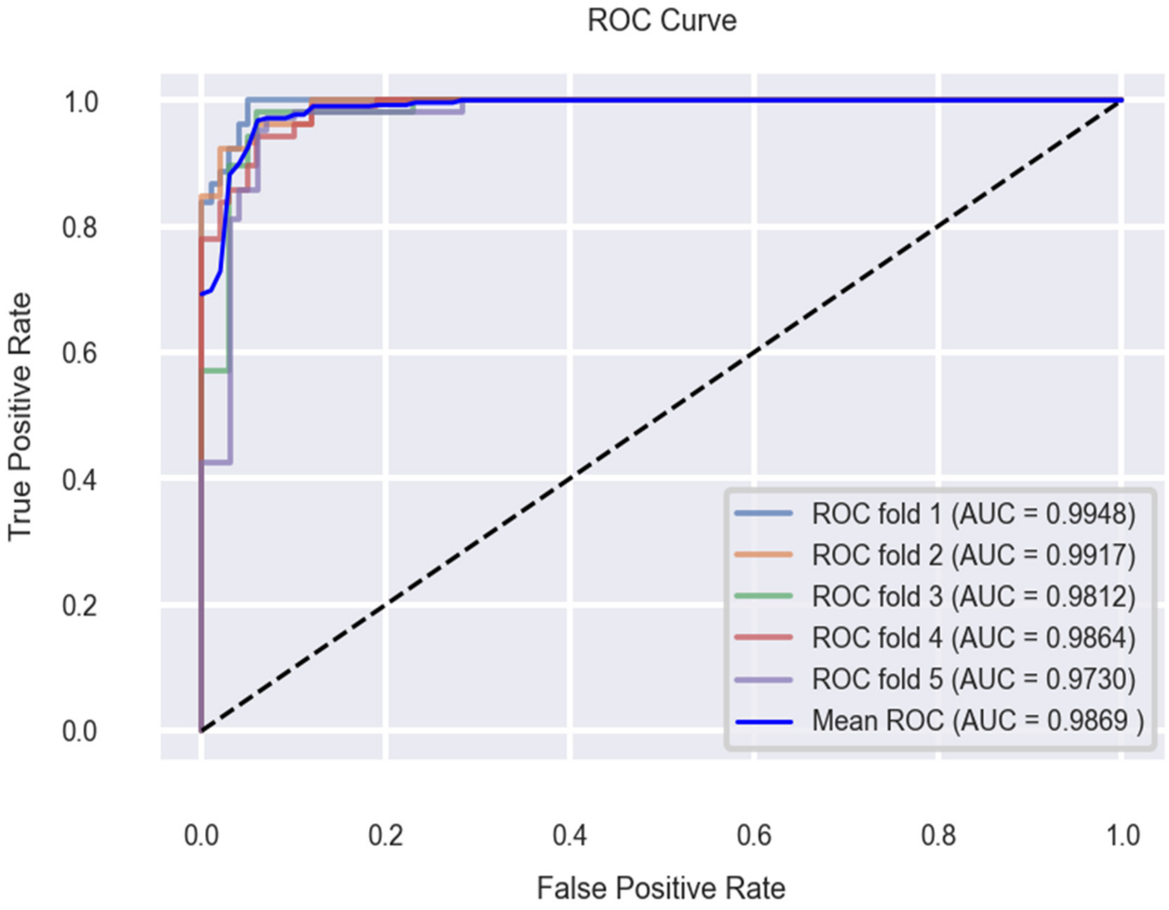

Figure 26. Stratified (K = 5)-Fold Cross-validation Results for NuSVC (ROC-AUC).

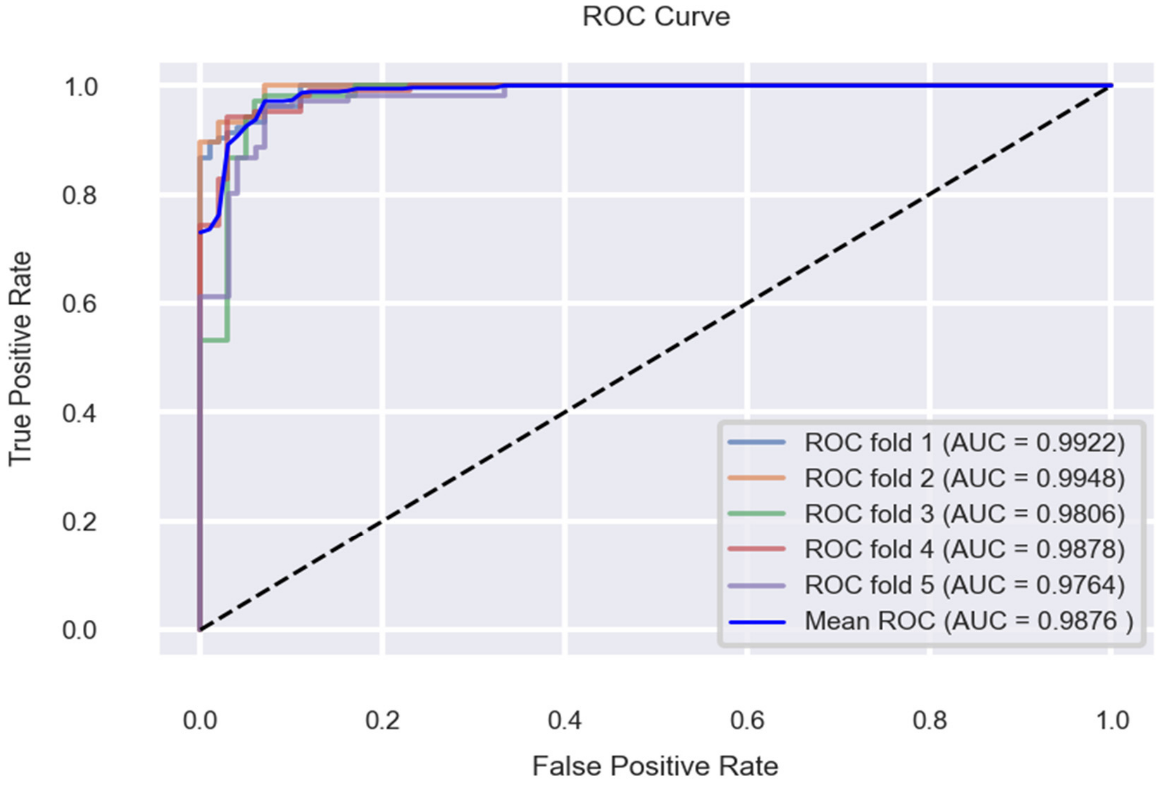

Figure 27. Stratified (K = 5)-Fold Cross-validation Results for Linear Discriminant Analysis (ROC-AUC).

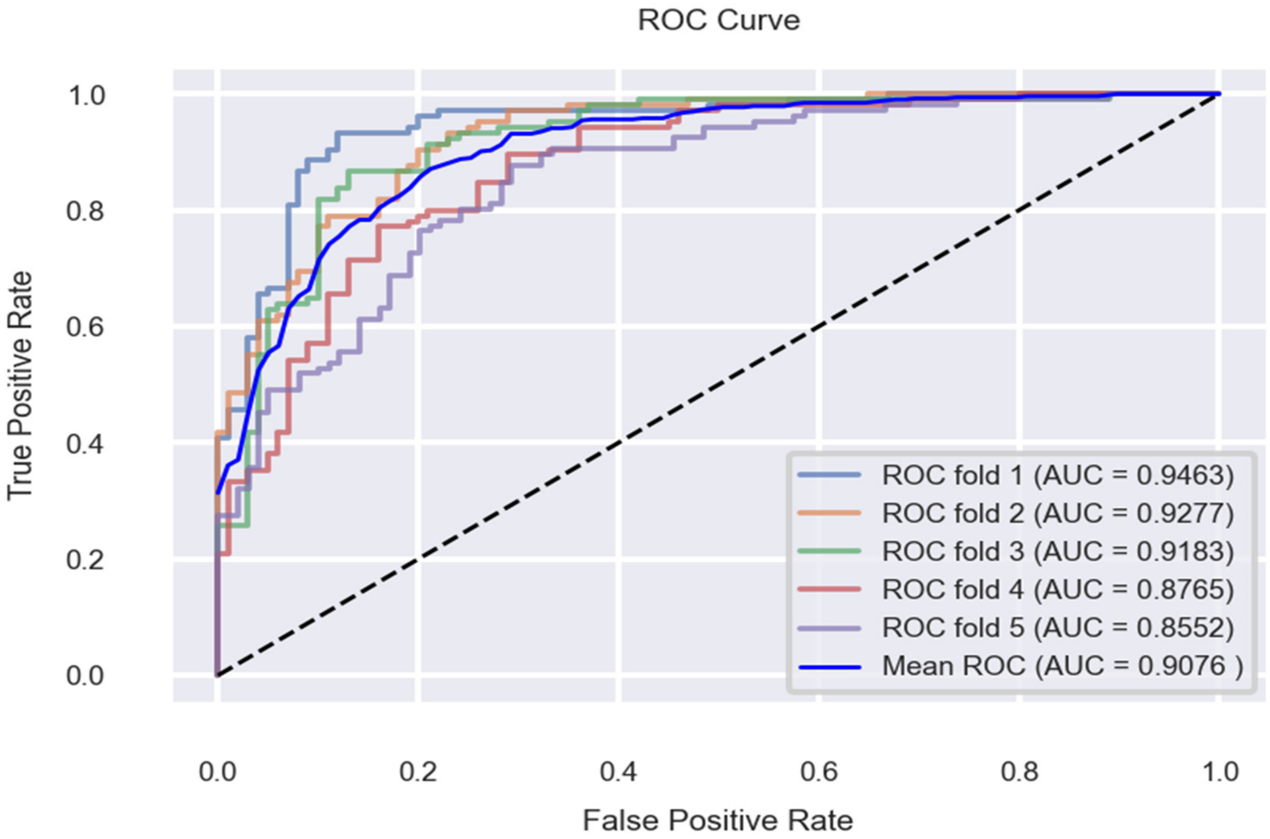

Figure 28. Stratified (K = 5)-Fold Cross-validation Results for Naive Bayes (ROC-AUC).

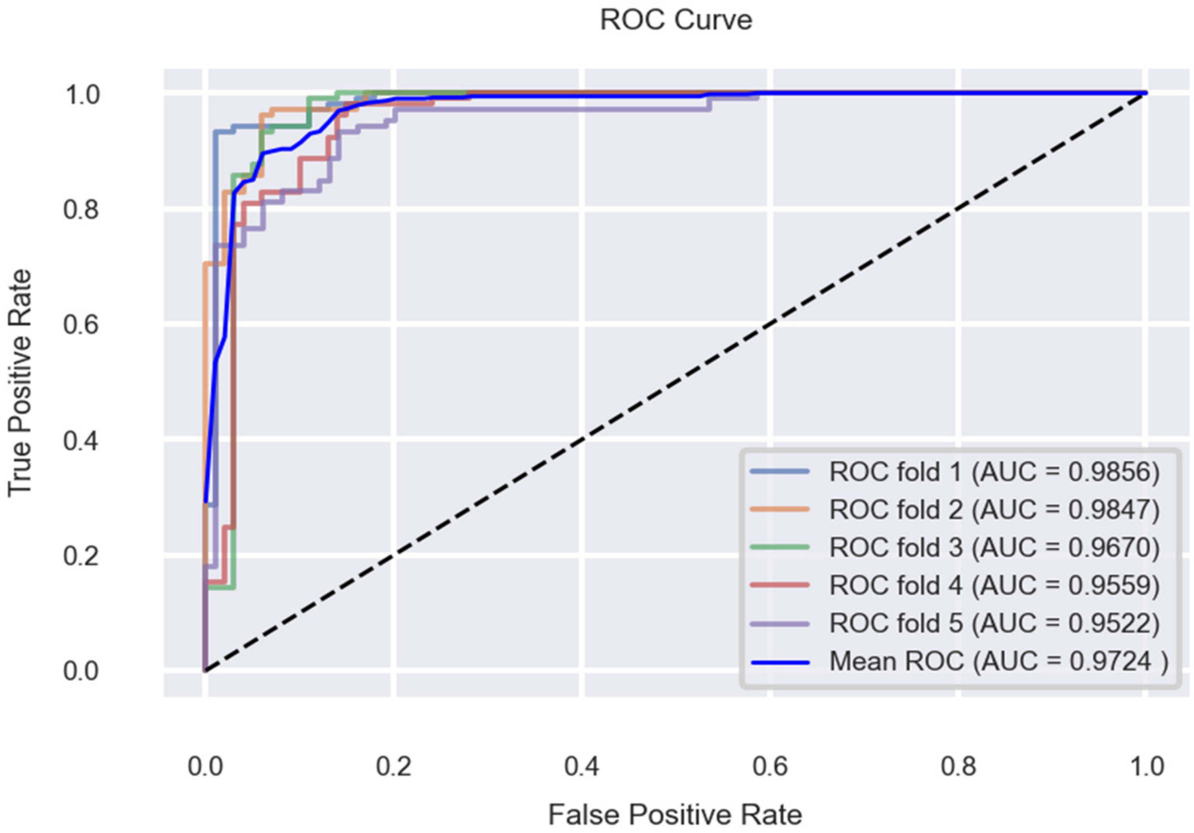

Figure 29. Stratified (K = 5)-Fold Cross-validation Results for Support Vector Machine (ROC-AUC).

As shown in Figure 15 and Table 4 the random forest classifier and decision tree classifier with the staking CV Classifier achieved the highest accuracy of 100%, precision of 100%, recall of 100%, and F1-score of 100%, and of top three combined, the staking CV classifier achieved 100% ROC-AUC. LR has the lowest accuracy of 86%, precision of 87%, and F1-score of 86%. Performance ROC-AUC is shown in Figure 30. It can be seen from the graphics that the ensemble model outperformed all of the other models.

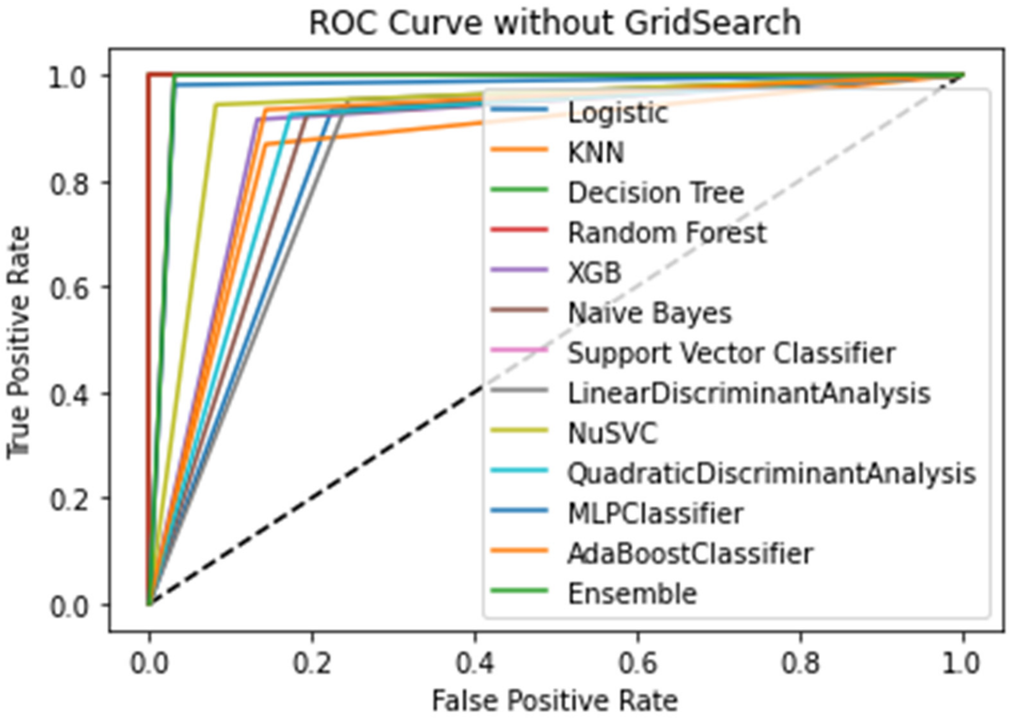

Figure 30. The performance of ML classifier using Staking CV Classifier.

Table 4. The performance of ML classifier using Staking CV classifier.

S. No.	Classifier	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
1	RF	100	100	100	100
2	DT	100	100	100	100
3	GB	89.26	89	89	89
4	KNN	86.34	86.19	86.25	86.04
5	AB	90.00	90.00	90.00	90.00
6	LR	86	87	86	86
7	NN	97	97	97	97
8	QDA	88	88	88	88
9	Nu SVC	93.17	93	93	93
10	LDA	86	87	86	86
11	NB	87	87	87	87
12	SVM	99	98	99	99
13	RFC+NN+SVM	100	100	100	100
14	RFC+NN+QDA	100	100	100	100
15	RFC+NN	100	100	100	100
16	LDA+QDA	89	89	89	89

We also compared our strategy to a number of other researchers' previously published methodologies. Ram Prakash et al. [41], for example, used a combination of the PCA feature extraction technique and DNN to achieve a high accuracy classification, but the recall was only 97%. The findings of the specific comparison are shown in Table 5.

Table 5. Comparative study of the proposed method with existing method.

Authors	Methods	Accuracy (%)	Recall (%)	Precision (%)	F1-Measure (%)	ROC-AUC (%)
Ramprakash et al. [41]	χ^2-DNN	94	93	-	-	-
Gao et al. [42]	Decision tree Plus PCA	99.00	97.00	98.00	-	-
Ali et al. [43]	MLP	97.95	98.00	98.00	-	-
D. Zhang et al. [44]	Linear SVC + DNN	98.56	99.35	97.84	-	-
Proposed **Base model results**	**RF**	**100**	**100**	**100**	**100**	**100**
	DT	**98.80**	**99**	**98**	**100**	**99**
	GB	**97.60**	**98**	**98**	**98**	**100**
	KNN	**90.80**	**84**	**97**	**90**	**97**
	AB	**88.00**	**91**	**86**	**89**	**94**
	LR	**84.80**	**89**	**82**	**86**	**91**
	NN	**84.80**	**92**	**81**	**86**	**91**
	QDA	**83.60**	**85**	**83**	**84**	**91**
	Nu SVC	**83.20**	**91**	**79**	**85**	**90**
	LDA	**82.80**	**87**	**80**	**84**	**91**
	NB	**82.40**	**86**	**81**	**83**	**88**
	SVM	**68.80**	**69**	**69**	**69**	**73**
Proposed Stratified (K = 5)-Fold Cross-validation Results	**RF**	99.41	100	98.89	99.40	99.97
	DT	99.60	99.23	100	99.61	99.62
	GB	99.41	99.50	99.45	99.53	100
	KNN	86.06	86.19	86.25	86.04	94.86
	AB	90.93	89.99	92.24	91.05	97.57
	LR	84.59	89.98	82.31	85.59	91.88
	NN	95.03	95.64	94.76	95.19	98.54
	QDA	94.16	94.49	94.27	94.33	98.74
	Nu SVC	94.83	95.45	94.60	94.97	98.54
	LDA	93.56	93.54	93.93	93.71	98.60
	NB	82.15	85.94	80.63	83.12	90.76
	SVM	92.19	93.36	91.68	92.49	97.24

Table 5. *Cont.*

Authors	Methods	Accuracy (%)	Recall (%)	Precision (%)	F1-Measure (%)	ROC-AUC (%)
Proposed stacking CV Classifier	RF	100	100	100	100	-
	DT	100	100	100	100	-
	GB	89.26	89	89	89	-
	KNN	86.34	86.19	86.25	86.04	-
	AB	90.00	90.00	90.00	90.00	-
	LR	86	87	86	86	-
	NN	97	97	97	97	-
	QDA	88	88	88	88	-
	Nu SVC	93.17	93	93	93	-
	LDA	86	87	86	86	-
	NB	87	87	87	87	-
	SVM	99	98	99	99	-
	RFC+NN+SVM	100	100	100	100	-
	RFC+NN+QDA	100	100	100	100	-
	RFC+NN	100	100	100	100	-
	LDA+QDA	89	89	89	89	-

7. Conclusions

Machine learning algorithms (MLAs) are important in healthcare because they analyze medical data to detect human cardiac disease. CVDs are a serious medical condition that healthcare providers and researchers must address. In this paper, twelve strategies are considered for doing comparative analysis and obtaining positive findings. In this investigation, it is concluded that machine learning algorithms beat human learning algorithms. Some of the comparison approaches used are the confusion matrix, precision, recall, F1-score, and ROC-AUC. For the 14 characteristics dataset, the random forest classifier performed better in the ML approach when data pre-processing was utilized. To address this issue, MLAs are employed to assess a dataset of CVD clinical data to identify the critical risk factors that influence CVD progression. To improve the training and testing of the algorithms, the random forest and decision tree are used. Then, using the train-test split approach, the prediction models' classification performance on combined datasets is examined. Finally, comparisons of the findings are given. Accuracy, precision, recall, F1-score, and ROC-AUC are used as performance indicators, for the Cleveland heart disease dataset, whereas Switzerland, Hungarian, and Long Beach VA datasets were used to predict CVD. The following are the results from the combined heart disease dataset: F1-score is 100%, accuracy is 100%, precision is 100%, and recall is 100%. The following are the analytical results of the decision tree classifier for the pooled heart disease dataset: F1-score of 100%, accuracy of 98.80%, precision of 98%, recall of 99%, ROC-AUC of 99%, and the performance of ROC-AUC of 100% for top models' random forest and gradient boosting classifier, respectively. The performances of the machine learning algorithms are improved by using five-fold cross validation. Again, the Stacking CV Classifier is also used to improve the performances of the individual machine learning algorithms by combining two and three techniques together. Several reduction methods may also be used to improve the random forest classification algorithm's accuracy.

Author Contributions: Conceptualization, G.N.A. and S.; methodology, G.N.A. and S.; software, G.N.A.; validation, H.F. and G.N.A.; formal analysis, G.N.A., S. and H.F.; investigation, G.N.A.; resources, I.; writing—original draft preparation, G.N.A. and S.; writing—review and editing, I. and M.A.; visualization, O.R. and M.S.A.; supervision, H.F. and S.; project administration, M.A. and M.S.A.; funding acquisition, M.A., O.R. and M.S.A. All authors have read and agreed to the published version of the manuscript.

Funding: The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University (KKU) for funding this research project Number (RGP.2/248/43).

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

References

1. Ornish, D. Can lifestyle changes reverse coronary heart disease? The Lifestyle Heart Trial. *Lancet* **1990**, *336*, 129–133. [CrossRef]
2. Ambrosy, A.P.; Fonarow, G.C.; Butler, J. The global health and economic burden of hospitalizations for heart failure: Lessons learned from hospitalized heart failure registries. *J. Am. Coll. Cardiol.* **2014**, *63*, 1123–1133. [CrossRef] [PubMed]
3. Bui, A.L.; Horwich, T.B.; Fonarow, G.C. Epidemiology and risk profifile of heart failure. *Nat. Rev. Cardiol.* **2011**, *8*, 30–41. [CrossRef] [PubMed]
4. Bathrellou, E.; Kontogianni, M.D.; Chrysanthopoulou, E. Adherence to a dash-style diet and cardiovascular disease risk: The 10-year follow-up of the Attica study. *Nutr. Health* **2019**, *25*, 225–230. [CrossRef] [PubMed]
5. Das, R.; Turkoglu, I.; Sengur, A. Effffective diagnosis of heart disease through neural networks ensembles. *Expert Syst. Appl.* **2009**, *36*, 7675–7680. [CrossRef]
6. Sinha, R.K.; Aggarwal, Y.; Das, B.N. Backpropagation artificial neural network classifier to detect changes in heart sound due to mitral valve regurgitation. *J. Med. Syst.* **2007**, *31*, 205–209. [CrossRef]
7. Dangare, C.S.; Apte, S.S. Improved study of heart disease prediction system using data mining classification techniques. *Int. J. Comput. Appl.* **2012**, *47*, 44–48.
8. Spencer, K.T.; Kimura, B.J.; Korcarz, C.E.; Pellikka, P.A.; Rahko, P.S.; Siegel, R.J. Focused cardiac ultrasound: Recommendations from the american society of echocardiography. *J. Am. Soc. Echocardiogr.* **2013**, *26*, 567–581. [CrossRef]
9. Beymer, D.; Syeda-Mahmood, T. Cardiac disease recognition in echocardiograms using spatio-temporal statistical models. In Proceedings of the 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, BC, Canada, 20–25 August 2008; pp. 4784–4788.
10. Wu, H.; Huynh, T.T.; Souvenir, R. Motion factorization for echocardiogram classification. In Proceedings of the 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), Beijing, China, 29 April 2014–2 May 2014; pp. 445–448.
11. Buman, M.P. Reallocating time to sleep, sedentary behaviors, or active behaviors: Associations with cardiovascular disease risk biomarkers, NHANES 2005–2006. *Am. J. Epidemiol.* **2014**, *179*, 323–334. [CrossRef]
12. Javeed, A. Machine learning-based automated diagnostic systems developed for heart failure prediction using different types of data modalities: A systematic review and future directions. *Comput. Math. Methods Med.* **2022**, *2022*, 9288452. [CrossRef]
13. Virani, S.S. Heart Disease and Stroke Statistics—2020 Update: A Report from the American Heart Association. *Lippincott Williams Wilkins* **2020**, *141*, e139–e596. [CrossRef]
14. American Heart Association. *Heart Disease and Stroke Statistics Update Fact Sheet American Heart Association Research Heart Disease, Stroke and Other Cardiovascular Diseases, Coronary Heart Disease (CHD)*; American Heart Association: Dallas, TX, USA, 2021.
15. Sturgeo, K.M. A population-based study of cardiovascular disease mortality risk in US cancer patients. *Eur. Heart J.* **2019**, *40*, 3889–3897. [CrossRef] [PubMed]
16. Kaptoge, S.; Pennells, L.; De Bacquer, D.; Cooney, M.T.; Kavousi, M.; Stevens, G.; Di Angelantonio, E. World Health Organization cardiovascular disease risk charts: Revised models to estimate risk in 21 global regions. *Lancet Glob. Health* **2019**, *7*, e1332–e1345. [CrossRef]
17. Khan, M.A.; Algarni, F. A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS. *IEEE Access* **2020**, *8*, 122259–122269. [CrossRef]
18. Ferdousi, R.; Hossain, M.A.; Saddik, A.E. Early-stage risk prediction of non-communicable disease using ML in health CPS. *IEEE Access* **2021**, *9*, 96823–96837. [CrossRef]
19. Shalev-Shwartz, S.; Ben-David, S. Understanding machine learning. In *Theory to Algorithms*; Cambridge University Press: Cambridge, UK, 2020.
20. Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning. In *Data Mining, Inference, and Prediction*; Springer: Cham, Switzerland, 2020.
21. Marsland, S. Machine learning. In *An Algorithmic Perspective*; CRC Press: Boca Raton, FL, USA, 2020.
22. Amin, M.S.; Chiam, Y.K.; Varathan, K.D. Identification of significant features and data mining techniques in predicting heart disease. *Telemat. Inform.* **2019**, *36*, 82–93. [CrossRef]
23. Spencer, R.; Thabtah, F.; Abdelhamid, N. Exploring feature selection and classification methods for predicting heart disease. *Digit. Health* **2020**, *6*, 2055207620914777. [CrossRef]
24. Khan, M.A. An IoT framework for heart disease prediction based on MDCNN classifier. *IEEE Access* **2020**, *8*, 34717. [CrossRef]

25. Mehmood, A.; Iqbal, M.; Mehmood, Z. Prediction of heart disease using deep convolutional neural networks. *Arab. J. Sci. Eng.* **2021**, *46*, 3409–3422. [CrossRef]
26. Budholiya, K.; Shrivastava, S.K.; Sharma, V. An optimized XGBoost based diagnostic system for effective prediction of heart disease. *J. King Saud Univ. Comput. Inf. Sci.* **2020**, *34*, 4514–4523. [CrossRef]
27. Martins, B.; Ferreira, D.; Neto, C.; Abelha, A.; Machado, J. Data mining for cardiovascular disease prediction. *J. Med. Syst.* **2021**, *45*, 6. [CrossRef]
28. Miranda, E.; Irwansyah, E.; Amelga, A.Y.; Maribondang, M.M.; Salim, M. Detection of cardiovascular disease risk's level for adults using naive bayes classifier. *Healthc. Inform. Res.* **2016**, *22*, 196–205. [CrossRef] [PubMed]
29. Pandey, A.; Pandey, P.; Jaiswal, K.L.; Sen, A.K. A Heart Disease Prediction Model Using Decision Tree. Available online: www.iosrjournals.orgwww.iosrjournals.org (accessed on 27 May 2021).
30. Mienye, I.D.; Sun, Y.; Wang, Z. Improved sparse autoencoder based artificial neural network approach for prediction of heart disease. *Inform. Med. Unlocked* **2020**, *18*, 100307. [CrossRef]
31. Siontis, K.C.; Noseworthy, P.A.; Attia, Z.I.; Paul, A. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat. Rev. Cardiol.* **2021**, *18*, 465–478. [CrossRef] [PubMed]
32. Anitha, S.; Sridevi, N. Heart disease prediction using data mining techniques. *J. Anal. Comput.* **2019**, *8*, 48–55.
33. Kumar, N.K.; Sindhu, G.S.; Prashanthi, D.K.; Sulthana, A.S. Analysis and prediction of cardio vascular disease using ML classifiers. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 March 2020; pp. 15–21.
34. Chowdhury, M.E.; Khandakar, A.; Alzoubi, K.; Mansoor, S.; Tahir, A.M.; Reaz, M.B.I.; Al-Emadi, N. Real-Time Smart-Digital Stethoscope System for Heart Diseases Monitoring. *Sensors* **2019**, *19*, 2781. [CrossRef]
35. Negi, S.; Kumar, Y.; Mishra, V.M. Feature extraction and classification for EMG signals using linear discriminant analysis. In Proceedings of the 2016 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA), Bareilly, India, 30 September 2016–1 October 2016.
36. Linda, P.S.; Yin, W.; Gregory, P.A.; Amanda, Z.; Margaux, G. Development of a novel clinical decision support system for exercise prescription among patients with multiple cardiovascular disease risk factors. *Mayo Clin. Proc. Innov. Qual. Outcomes* **2021**, *5*, 193–203.
37. Ahmad, G.N.; Ullah, S.; Algethami, A.; Fatima, H.; Akhter, S.M.H. Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using ML Technique with and without Sequential Feature Selection. *IEEE Access* **2022**, *10*, 23808–23828. [CrossRef]
38. Heart Disease Dataset. Available online: https://www.kaggle.com/johnsmith88/heart-disease-dataset (accessed on 20 July 2022).
39. Heart Disease Data Set. Available online: https://archive.ics.uci.edu/ml/datasets/Heart+Disease (accessed on 20 July 2022).
40. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
41. Ramprakash, P.; Sarumathi, R.; Mowriya, R. Heart disease prediction using deep neural network. In Proceedings of the International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 26–28 February 2020; pp. 666–670.
42. Gao, X.Y.; Ali, A.A.; Hassan, H.S.; Amwar, E.M. Improving the accuracy for analyzing heart diseases prediction based on the ensemble method. *Complexity* **2021**, *2021*, 6663455. [CrossRef]
43. Ali, M.M.; Kumar, B.P.; Ahmad, K.; Francis, M.B.; Julian, M.W.Q.; Moni, M.A. Heart disease prediction using supervised ML algorithms: Performance analysis and comparison. *Comput. Biol. Med.* **2021**, *136*, 104672. [CrossRef] [PubMed]
44. Zhang, D.; Chen, Y.; Chen, Y.; Ye, S.; Cai, W.; Jiang, J.; Chen, M. Heart Disease Prediction Based on the Embedded Feature Selection Method and Deep Neural Network. *J. Healthc. Eng.* **2021**, *2021*, 6260022. [CrossRef] [PubMed]