Cardiovascular Disease (CVD) Prediction Using Machine Learning Techniques with XGBoost Feature Importance Analysis

Nurzahan Akter Joly[1], Zahrul Jannat Peya[2], Romjan Munchi[3]

[1] M. Sc Research Scholar, Computer Science and Engineering Discipline, Khulna University
[2] Ph.D. Research Scholar, Department of Computer Science and Engineering, Khulna University of Engineering and Technology
[3] Student, Department of Computer Science and Engineering, North Western University, Khulna

Abstract

Cardiovascular diseases (CVD) are a type of illnesses in the cardiovascular system including coronary, rheumatic heart and cerebrovascular. The leading causes of disease burden and mortality worldwide are CVDs. CVD can cause a wide range of consequences, which can lower standard of life and sometimes cause death. This emphasizes the requirement for the establishment of a technique that can ensure an exact and prompt prediction of the risk of CVD in patients. This study investigates effective CVD prediction system using several Machine Learning (ML) classification models. Rigorous data analysis through several preprocessing techniques as well as feature importance analysis has been performed through Spearman Correlation Analysis and XGboost feature importance technique. Finally, classification has been accomplished through Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR) using a standard benchmark dataset collected from IEEEDataPort. Highest accuracy of 95% has been achieved through Random Forest (RF). The findings of this study will assist professionals in the medical field in the early diagnosis of cardiovascular disease in patients.

Keywords: Cardiovascular Disease (CVD), Spearman Correlation Analysis, XGBoost Feature Importance, Random Forest, Support Vector Machine, Logistic Regression

1 Introduction

Cardiovascular diseases (CVD) are a type of illnesses in the cardiovascular system including coronary, rheumatic heart and cerebrovascular [1]. According to reports, over 17.9 million people globally pass away each year as a result of heart and vascular conditions. According to the Disease Report of Global Burden at the year 2019, there will be 523 million instances of CVD worldwide in 2019 and above to 18.6 million fatalities, or three percent of all fatalities. Experience both domestically and internationally demonstrates that quick identification and successful supervision of interventions in communities at risk is an obvious scientific path and economical early detection along with monitoring programs that can prolong life expectancy, improve living conditions, and decrease the impact of CVD. The leading cause of death in the globe is CVD that has an impact on people of all demographics, genders, races, and financial status [2]. Cardiovascular illnesses come in a wide variety, including but not limited to arrythmia, valve damage, cerebrovascular problem, failure of heart, coronary artery disease, congenital heart disease, pericardial disease etc. Depending on the exact type, cardiovascular disease can have a variety of causes.

If a person has risk factors such as hypertension, type 2 diabetes, cigarette use, family records of the disease, an absence of physical exercise, being overweight or obese, etc., they may be at an elevated risk for CVD. If CVDs are not appropriately managed, they may result in heart failure or fatal strokes. People can change their lifestyles or take medication to control CVDs. An early diagnosis can lead to more efficient treatment [3]. Machine learning (ML) has produced tremendous advancements in health services and clinical studies recently. Various ML feature extraction and classification methods are used in recent years for CVD. The selection of the essential variables that can operate as risk variables in prediction models is of highest relevance. Researchers must be very careful when selecting the features and machine learning algorithms to use in order to create reliable prediction models [4]. It is currently difficult for researchers to predict heart disease properly utilizing a combination of these factors and the right machine learning techniques. For machine learning algorithms to work at their best, relevant data must be used for training. Several feature processing algorithms such as PCA, KPCA, LDA [5], Minimum Redundancy Maximum Relevance Feature Selection (MRMR) [6] and XGboost[1] are used for CVD diagnosis in different papers. ML classification algorithms such as Decision Tree (DT) [6, 7], Naïve Bayes (NB) [6], Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM) [8] are used for classifying the objects.

This study investigates the analysis of feature impotance for CVD diagnosis and perform classification with improved accuracy. At first consolidated CVD dataset has been collected from IEEEDataPort. Preprocessing has been accomplished through python. Important features have been identified using XGboost feature importance technique. Finally, Classification on the dataset has been done using Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR) techniques.

This paper’s remaining sections are organized as comes afterwards. Section 2 of this paper reviews the existing research and Section 3 discusses the data and the various methods that are employed to develop the CVD prediction system. The performance and findings of this research are compared to other studies, and the details of the experiments are showed in Section 4. The paper's main implications are summed up in Section 5.

2 Literature Review

Over the past few years, a variety of ML-based techniques have been researched for CVD detection. A detailed review of the related existing studies is delineated below. S. Bashir et al. [6] used an opensource database called the UCI repository to carry out the prediction method. They applied the ensemble process on the data using the subprocesses MRMR. Applying several ML algorithms, they achieved highest accuracy of 84.85%. Due to its higher accuracy achievement compared to other approaches, the study suggests LR, SVM as the best method for heart illness prediction. Future real-time medical datasets can be used with the methods outlined in this paper, as well as ensembles, which are mixtures of different algorithms. This would result in increased accuracy and improved efficiency.

H. Meshref et al. [7] examined the Cleveland Heart Data Set, which is accessible at the UCI Machine Learning Repository. In the study, popular machine learning techniques such as Artificial Neural Networks, Naive Bayes, RF, SVM and Decision Trees have been investigated to improve the construction, comprehension, and interpretation of various heart disease diagnostic models. Rapid and simple attribute selection methods, such as attribute subset selection methods and single attribute evaluators with rankings, were investigated. The accuracy of the ANN model is the greatest at 84.25%. The size of the dataset is a drawback to the study. They used just 303 cases for their research. In their future work, they plan to combine the Hungarian and Cleveland data sets and conduct the appropriate experiment, which could increase accuracy and provide additional details on the opacity of every model that is generated.

In a study, the SVM, RF and LR approaches were used to create three separate models for categorizing coronary heart disease [8]. The IEEE Data Port database's dataset on cardiac disease was used in the study. The hyperparameter values were optimized using a 10-fold cross validation technique over three iterations. A classification accuracy LR of 0.861, SVM of 0.897, and RF of 0.929 were demonstrated. However, the paper does not sufficiently detail the data preprocessing methods. In addition, the optimal settings for each model's hyperparameters were determined using the time-consuming Grid Search method.

In order to improve performance, S.M. Saqlain et al. suggested method for choosing a subset of feature for a medical heart disease diagnosis system [9]. The three techniques provided in the suggested strategy for selecting feature subsets are the forward feature selection method, the reverse feature selection algorithm, and the Fisher score-based feature selection algorithm. The feature subset selection algorithm (FSSA) produced the best lower dimensional subset of feature. The accuracy was determined to be 82.7,81.19, 92.68, and 84.52 % for SPECTF, Cleveland, Switzerland and Hungarian respectively. In the study, the effectiveness of the Cleveland, Hungarian, Swiss, and SPECTF models in predicting heart disease was examined separately. It is important to assess the combined dataset's overall prediction capability.

Cleveland database from UCI's machine learning was used in [4]. The aim of the study was to identify the most significant predictors of cardiovascular disease and the data mining techniques that can improve the precision of the prediction. Different feature combinations were used to create prediction models using DT, KNN, LR, NB, Neural Networks (NN), SVM and Vote (a hybrid method with LR and NB). A 10-fold cross validation approach was used to assess the models' efficacy. According to the results of the studies, the prediction model developed using the acknowledged significant features and the most effective data mining technique (i.e. Vote) reaches an accuracy of 87.4%.

K. G. Dinesh et al [10] used a variety of machine learning techniques to forecast cardiac disease. They used LR, NB, GB, RF, and SVM in their system to find fascinating patterns in data taken from the UCI machine learning library. Out of these 76 attributes, only 14 were utilized in the authors' strategy. total accuracy from SVM was 79.77%, and total accuracy from RF was 80.89%. In this work, the tuning of the model's hyperparameters is not covered. The SVM and RF models' prediction performance can be enhanced by modifying the hyperparameters.

An enhanced sparse autoencoder-based ANN was suggested in [11] to help in predicting the risk of CVD. The sparse autoencoder was utilized to discover the most practical method to show the data, and the artificial neural network (ANN) was employed to generate forecasts depending on the learned records. The SAE was improved utilizing batch normalizing and the Adam method. The test accuracy was 90%.

Ozcan et al. [12] made early detection easier by developing several ML based classification applications in medicine. The Classification and Regression Tree (CART) algorithm, a supervised machine learning approach, has been used in this study to forecast heart disease and obtain decision rules in order to elucidate correlations underlying both input and output variables. The results of the study also assign a priority ranking to the characteristics that affect heart disease. The prediction's 87% accuracy when all performance factors are taken into account verifies the model's dependability. On the other hand, the study's disclosed extracted decision criteria make it easier to employ them for therapeutic reasons without the need for additional proficiency.

3 Methodology

A lot of experts have come up with different ways to predict cardiovascular disease. Some people only use one machine learning method, while others use a mix of methods. Figure 1 shows the suggested way of predicting cardiovascular disease. The methods are explained in the parts that follow.

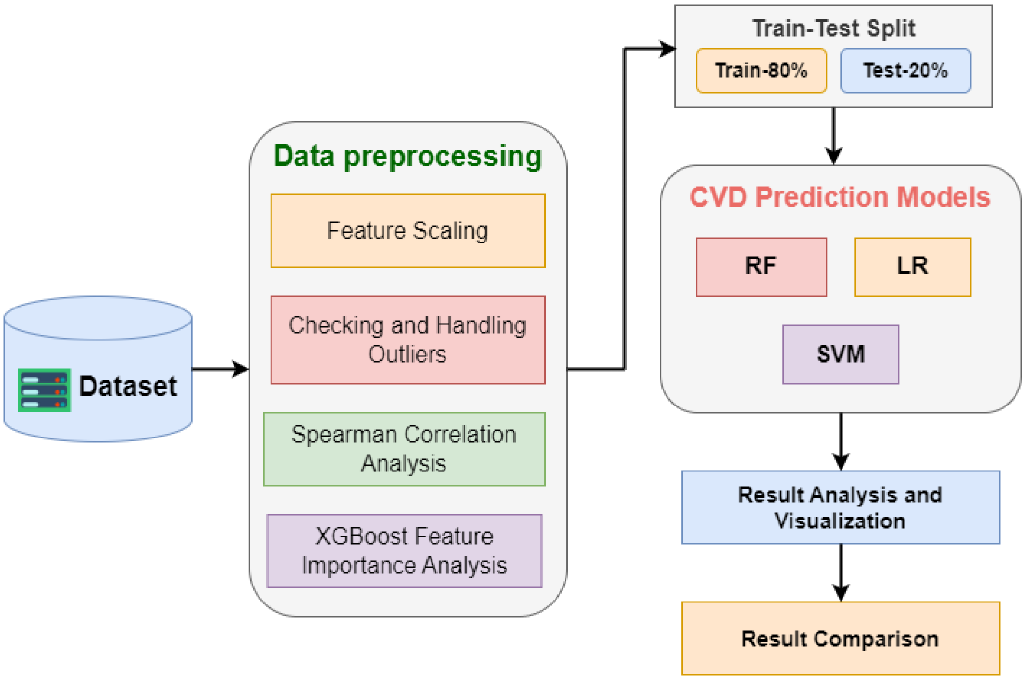

Dataset
Data preprocessing
Feature Scaling
Checking and Handling Outliers
Spearman Correlation Analysis
XGBoost Feature Importance Analysis
Train-Test Split
Train-80%
Test-20%
CVD Prediction Models
RF
LR
SVM
Result Analysis and Visualization
Result Comparison

Figure 1: System Architecture of the Proposed Model

3.1 Dataset

Having a reliable dataset for diagnosing cardiovascular illness is essential. An accurate result from applying machine learning techniques relies heavily on the quality of the data used in the process. This research utilized the already existing IEEEDataPort collection [8].

Table 1. Description of the Attributes for the CVD Dataset

Attribute	Code Given	Unit	Data Type
Age	Age	in years	Numeric
Sex	Sex	1,0	Binary
Chest pain type	chest pain type	1, 2, 3, 4	Nominal
Resting blood pressure	resting bp s	in mm Hg	Numeric
Serum cholesterol	Cholesterol	in mg/dl	Numeric
Fasting blood sugar	fasting blood sugar	1,0>120 mg/dl	Binary
Resting electrocardiogram results	resting ecg	0, 1, 2	Nominal
Maximum heart rate achieved	max heart rate	71-202	Numeric
Exercise induced angina	angina	0, 1	Binary
Oldpeak = ST	oldpeak	depression	Numeric
The slope of the peak exercise STsegment	ST slope	0, 1, 2	Nominal
Class	target	0, 1	Binary

Together, the Long BeachVA, Statlog, Cleveland, Hungarian, and Switzerland records made up the final dataset. This group of heart disease records was put together using eleven factors. This dataset has 1190 cases with 11 distinct features. Of these, 629 have CVD and 561 do not. Table 1 gives an entire summary of the dataset's properties.

3.2 Data Preprocessing

Data preprocessing refers to the steps used to transform raw data into a format usable by a ML model. The accuracy and efficiency of a machine learning model can be improved through data preprocessing. In order to get the data ready for analysis, the search was conducted to find missing values. There were no missing values for the twelve features that were taken from the dataset. In the case of feature scaling standardization is favored over normalization. Standardization is a method of rescaling data whereby a column's values are transformed into the form of a normal distribution with mean = 0 and variance = 1 [13]. The observations that behave in a very different way from the rest of the data are called outliers. An outlier or anomaly is a condition that is not typical, or a finding that was not expected [14].

To determine whether there are any outliers, a boxplot representation of the relevant features was used as in Figure 2. It is found that cholesterol and resting bp s features had some outliers. Following the identification of the outliers, the capping technique was utilized in order to get rid of the outliers.

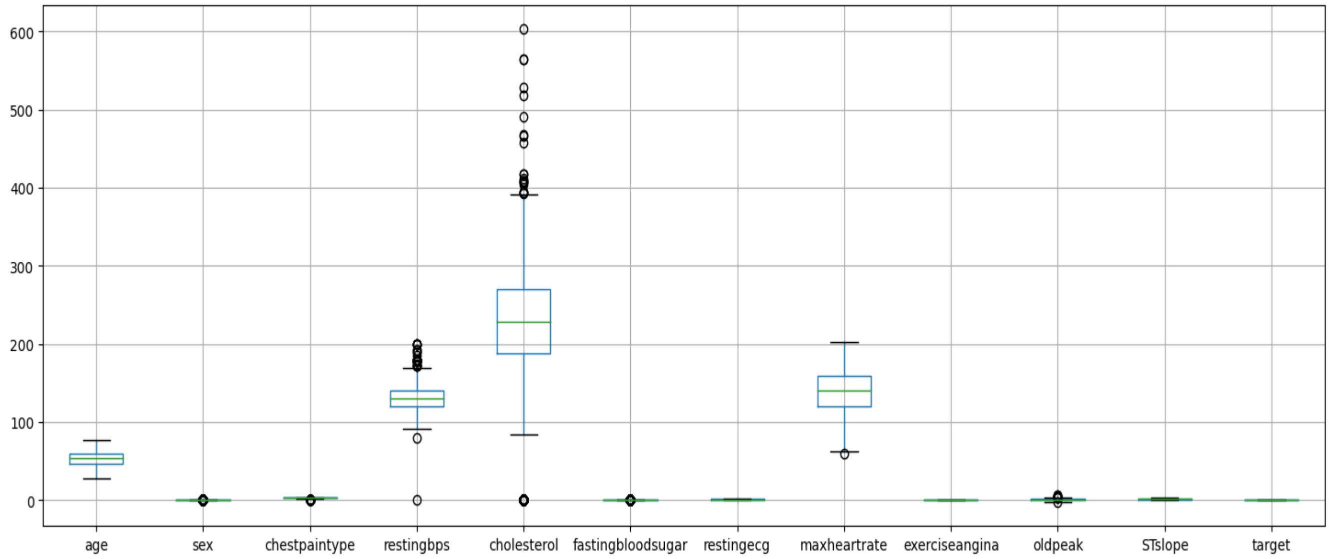


Figure 2: Boxplot Representation of the Twelve Features

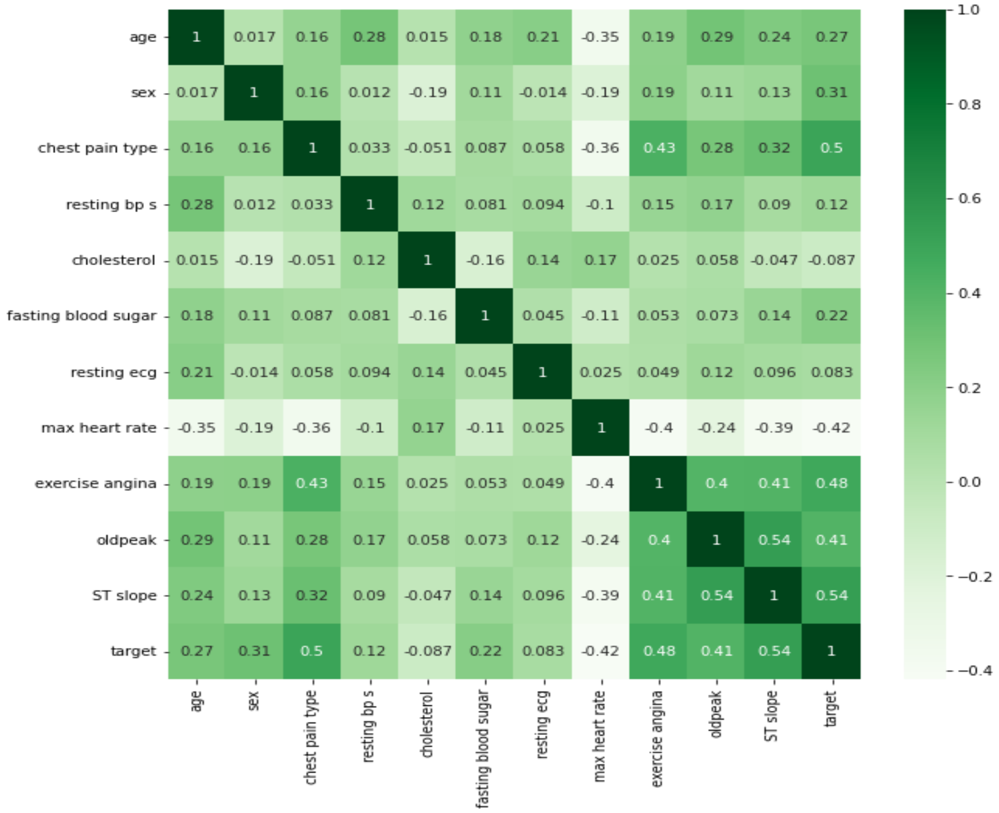
The relevance of features is a measure of how much insights is sent by each characteristic to the model's prediction. In essence, it evaluates how much a given variable contributes to the accuracy of the present model and predictions. When describing the significance of a feature as a whole, a numeric value is used that we refer to as the score; generally speaking, the higher the score value, the more significant the feature is. It is feasible to discover features that are not relevant and remove them from consideration by utilizing variable significance scoring. It's possible that the model's speed and performance could be enhanced by reducing the number of irrelevant variables. Many methods exist for determining which features are most significant. The approaches of spearman correlation analysis and XGBoost feature importance is utilized in this research to determine the most important attributes.

A common method for uncovering valuable connections in data, correlation analysis has seen the widespread application [15]. With numerical input and output variables, the spearman correlation coefficient outperforms other measures. spearman is a nonparametric coefficient, it can be used in situations where parametric assumptions have been broken (i.e., when data are not normally distributed), the sample size is small, or there is an outlier problem in the data set. The explained rank variability can be used to make sense of this correlation. Using the rank of the observations, it can also be used to evaluate monotonic relations. Because linear relationships are monotonic but not all monotonic relationships are linear, this is a significant concern [16]. Figure 3 displays the results of a spearman correlation test applied to the dataset.

As a decision tree-based classifier, XGBoost is one use of the gradient boosting (GB) method. It has seen widespread adoption because of its speed, efficiency, and scalability. The following is a simplified explanation of how GB and XGBoost work. There are n observations in the dataset D = [x, y], where x is the variable that is independent and y is the dependent variable. In GB, let's say there is k levels of boosting.

$$\hat{y}_i = \sum_{b=1}^{B} f_b(x_i) \qquad (1)$$

Then, B function can be used to predicts the outcome using $\hat{y}_i$ as the estimate for the i-th sample at the b-th level of boosting. Let f_b stand for a tree construction q, and leaf j have a weight score w_j. Then, for a given sample x_i, the end prediction can be found by adding up all the scores from all the leaves, as shown in Equation 1 [17].



	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target
age	1	0.017	0.16	0.28	0.015	0.18	0.21	-0.35	0.19	0.29	0.24	0.27
sex	0.017	1	0.16	0.012	-0.19	0.11	-0.014	-0.19	0.19	0.11	0.13	0.31
chest pain type	0.16	0.16	1	0.033	-0.051	0.087	0.058	-0.36	0.43	0.28	0.32	0.5
resting bp s	0.28	0.012	0.033	1	0.12	0.081	0.094	-0.1	0.15	0.17	0.09	0.12
cholesterol	0.015	-0.19	-0.051	0.12	1	-0.16	0.14	0.17	0.025	0.058	-0.047	-0.087
fasting blood sugar	0.18	0.11	0.087	0.081	-0.16	1	0.045	-0.11	0.053	0.073	0.14	0.22
resting ecg	0.21	-0.014	0.058	0.094	0.14	0.045	1	0.025	0.049	0.12	0.096	0.083
max heart rate	-0.35	-0.19	-0.36	-0.1	0.17	-0.11	0.025	1	-0.4	-0.24	-0.39	-0.42
exercise angina	0.19	0.19	0.43	0.15	0.025	0.053	0.049	-0.4	1	0.4	0.41	0.48
oldpeak	0.29	0.11	0.28	0.17	0.058	0.073	0.12	-0.24	0.4	1	0.54	0.41
ST slope	0.24	0.13	0.32	0.09	-0.047	0.14	0.096	-0.39	0.41	0.54	1	0.54
target	0.27	0.31	0.5	0.12	-0.087	0.22	0.083	-0.42	0.48	0.41	0.54	1

Figure 3: Spearman Correlation Analysis

The ST slope was in the top place for XGBoost importance, the exercise angina was in the second position, chest pain type was in third position and sex was in the fourth position. It was also found that age, resting bp s, and resting ecg has comparatively less significance score at identifying CVD. The significance of the XGBoost characteristic is seen in Figure 4.

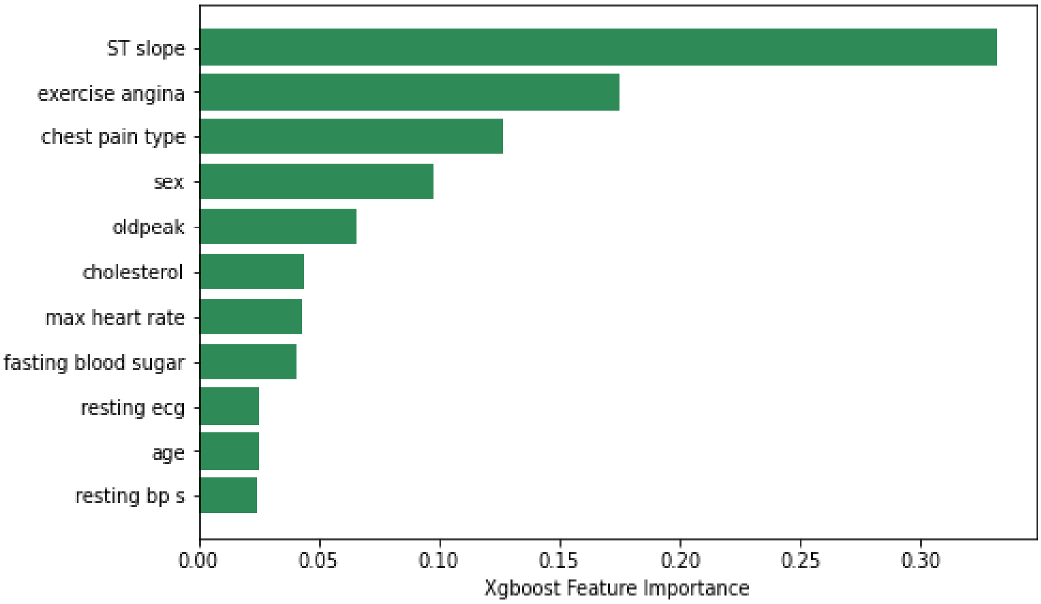


Figure 4: XGBoost Feature Importance

3.3 Train Test Split of the Dataset

A train-test split divides a dataset into a training set and a testing set. The training set was used to teach the model, and the testing set was used to check how accurate it was. The dataset used in this study was split into two portions, one for training (representing 80% of the dataset) and another for testing (20%). After being trained on 952, the model was put to the test on 238 different data. Figure 5 displays the graphical representation of the train-test split.

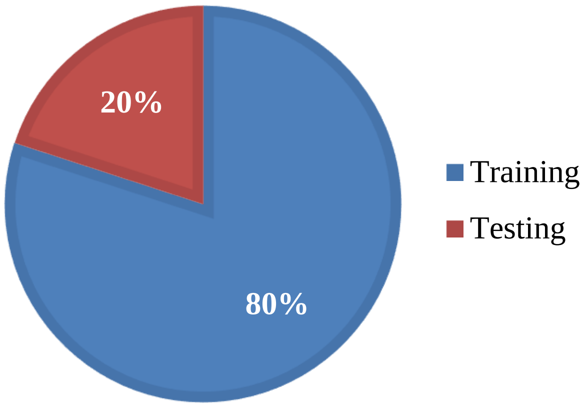


Figure 5: Train Test Split of Dataset

3.4 CVD Prediction with Machine Learning Techniques

Machine learning methods can be used in a lot of different medical situations to help with data analysis, modeling, and making sense of complicated clinical information. Random Forest, Support Vector Machine, and Logistic Regression are the three machine learning methods used in this investigation.

3.4.1 Random Forest

An effective ensemble learning technique, a Random Forest can be utilized for both classification and regression. It is a decision tree algorithm extension that is notable for its ability to handle complex data and reduce overfitting. The algorithm is frequently used for a range of tasks in machine learning and data science. Ensemble Learning, Decision Trees, Bootstrap Aggregating (Bagging), Random Feature Selection, Voting, and Averaging are the core concepts of RF. It samples the training dataset at random with replacement to generate various data subsets for training individual trees. A random subset of features is considered for splitting at each node of the tree [18]. This introduces randomness and diversity among the trees, reducing overfitting. Random Forest combines the predictions of individual trees for classification tasks using majority voting as in Figure 6.

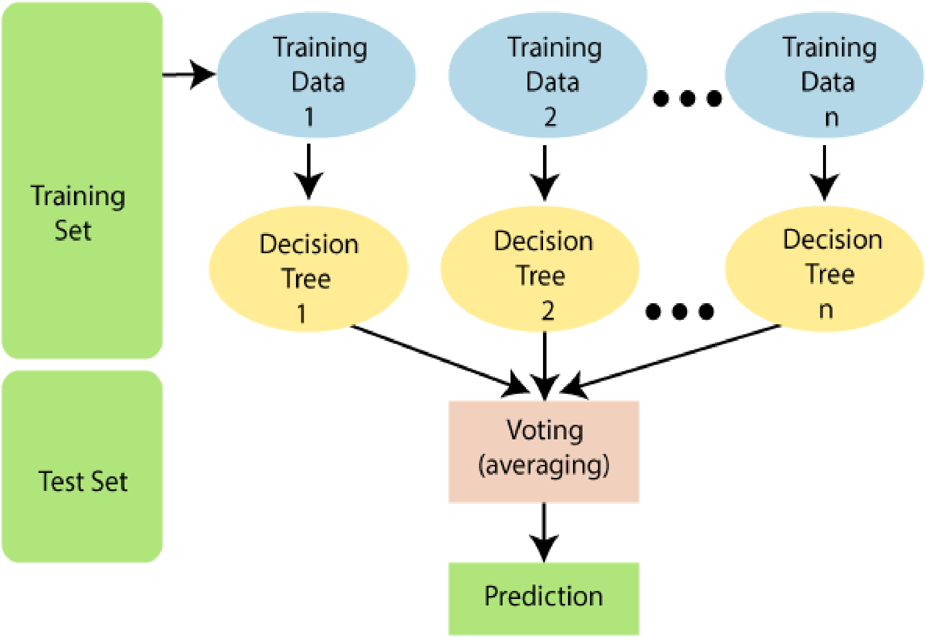


Figure 6: General Structure of Random Forest Classifier

The Random Forest algorithm encompasses various parameters, including the number of trees, the maximum depth of trees, the minimum number of samples per leaf, the number of features to consider at each split, the use of bootstrap samples (bootstrap), and the criterion for splitting (criterion). The Random Forest algorithm is readily accessible in widely-used machine learning frameworks such as scikit-learn in Python, facilitating its straightforward implementation.

3.4.2 Support Vector Machine

The support vector machine (SVM) is a powerful machine learning algorithm for classification and regression. It works by finding the best hyperplane for sorting data into groups. Data analysis and classification into two classes is the focus of Support Vector Machines, a type of supervised learning model. They are especially useful for binary classification jobs, which require categorizing data points into two groups as in Figure 7. Finding a hyperplane that makes the difference between two classes as notable as possible is what SVM is built on. How far away the hyperplane is from the closest data points in each class is what the margin is. Finding the weights (coefficients) and bias (intercept) that define the hyperplane is the mathematical formulation for this optimization problem. SVMs are classified into two types. That is Linear SVM and Nonlinear SVM. Linear SVM searches for a straight-line hyperplane using linearly separable data. When data can't be split up in a straight line, SVM uses kernel functions to move the data to a higher-dimensional area where it can be split up in a straight line. Some popular kernel functions are the Radial Basis Function (RBF) and the Polynomial kernel. SVM has a number of hyperparameters, including the regularization parameter (C), kernel type, and kernel parameters (for example, for an RBF kernel). Appropriate hyperparameter optimization is required for optimal model performance [19].

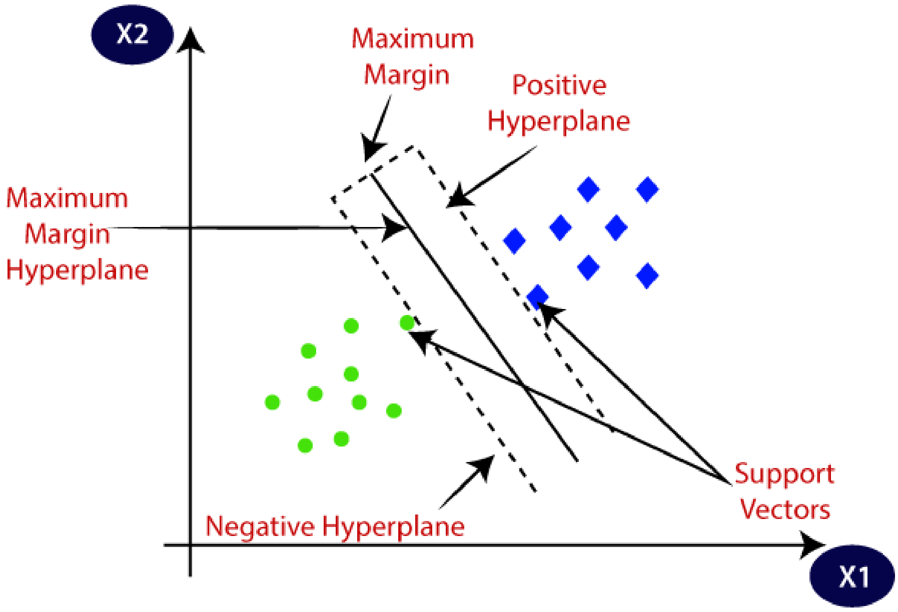


Figure 7: General Structure of Support Vector Distribution

3.4.3 Logistic Regression

Logistic regression is a supervised learning technique used in the fields of machine learning and statistics for the purpose of binary classification. Logistic Regression Analysis (LR) is a way to find out how the dependent variable and the independent factors are related to each other without having to assume a certain distribution. LR uses the maximum likelihood estimate method to find the unknown parameter values that give the best chance based on the data set. That is why the parameter estimates that make the probability function bigger are picked, along with the estimates that fit the data the best. Given its foundation in the logistic function, an S-shaped curve that converts every real number to a value between 0 and 1, the term "regression" seems fitting as in Figure 8. Logistic regression is a multi-stage procedure wherein one or more predictor factors and binary outcome variables are considered. The trained model can then be utilized for forecasting. The model, given a collection of predictor variables, determines the likelihood that the binary outcome is 1, and then, using a user-specified threshold, assigns the observation to one of two categories. Assuming a linear relationship between the predictor variables and the log-odds of the binary outcome, probability estimation is helpful for situations where there is a chance of an event occurring [20].

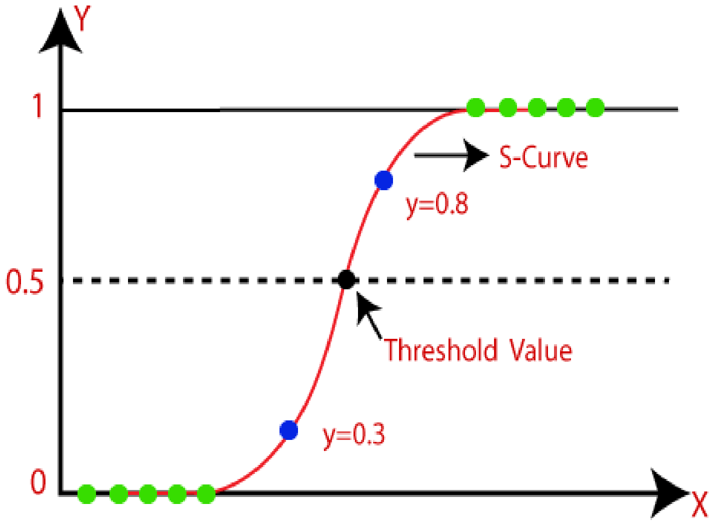


Figure 8: General Structure of Logistic Regression

3.5 Experimental Result Analysis

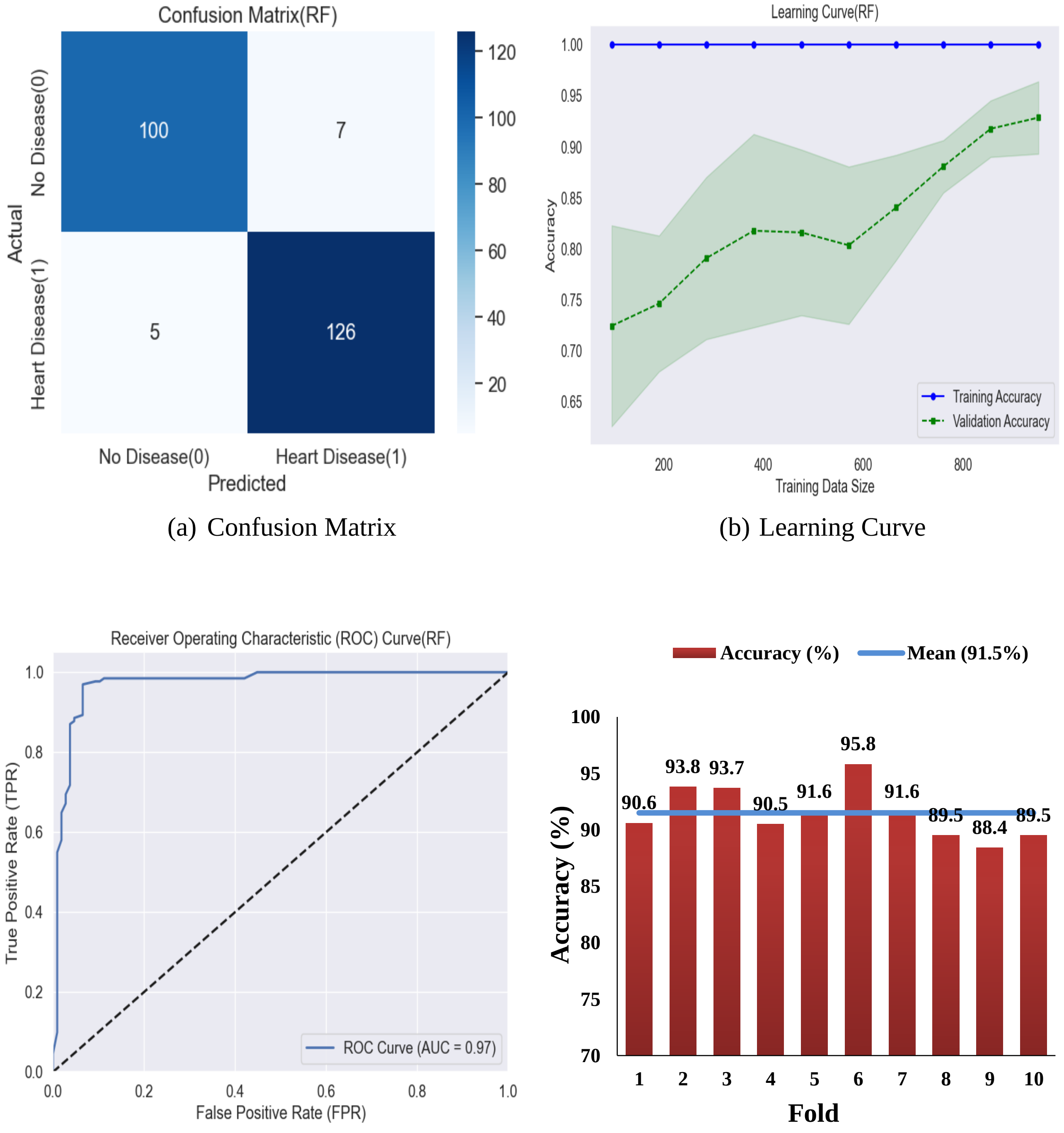
To ensure the highest level of accuracy, it is crucial to optimize the performance of a ML model before putting it into action. As part of the optimization process, specific factors known as hyperparameters are carefully adjusted to control how the model learns. Fine-tuning a model usually means fitting it to a training dataset several times with different sets of hyperparameters until the best setup for better performance is found. One effective way to find the best hyperparameter values is to use GridSearchCV, a method that involves making a full grid of possible hyperparameter values. The hyperparameter tuning values for the three machine learning models used in this research are shown in Table 2.

Table 2: Hyperparameter Tuning Values of the ML Techniques for the Dataset

Classifier	GridsearchCV Hyperparameter Tuning Values
RF	max_depth: 20, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 100
SVM	C: 10, gamma: 0.1, kernel: rbf
LR	Solver=liblinear, penalty=l1, C=1

In case of all three machine learning techniques 10-fold cross validation is used for the evaluation of the model's performance. The values of loss and accuracy for each fold offer an estimation of how well the model is performing on various subsets of the data. The confusion matrix of Random Forest model shows, 126 times patients are classified as cardiac patients correctly (TP), 5 times cardiac patients are classified as not cardiac patients (FN), 7 times patients having no cardiac problem are classified as cardiac patients (FP) and, 100 times patients are classified as not cardiac patients correctly (TN).

The RF classifier, demonstrates a perfect learning curve score of 100% for the dataset, which indicates that it efficiently learns from the training data. With the RF model, we were able to determine best Accuracy, Precision, Recall, F1-score and AUC value to be 0.95, 0.95, 0.95, 0.95 and 0.96, respectively. The average score for this model's cross-validation accuracy is 0.91. Plots of the RF model's measured performance on the dataset are shown in Figure 9.



(a) Confusion Matrix

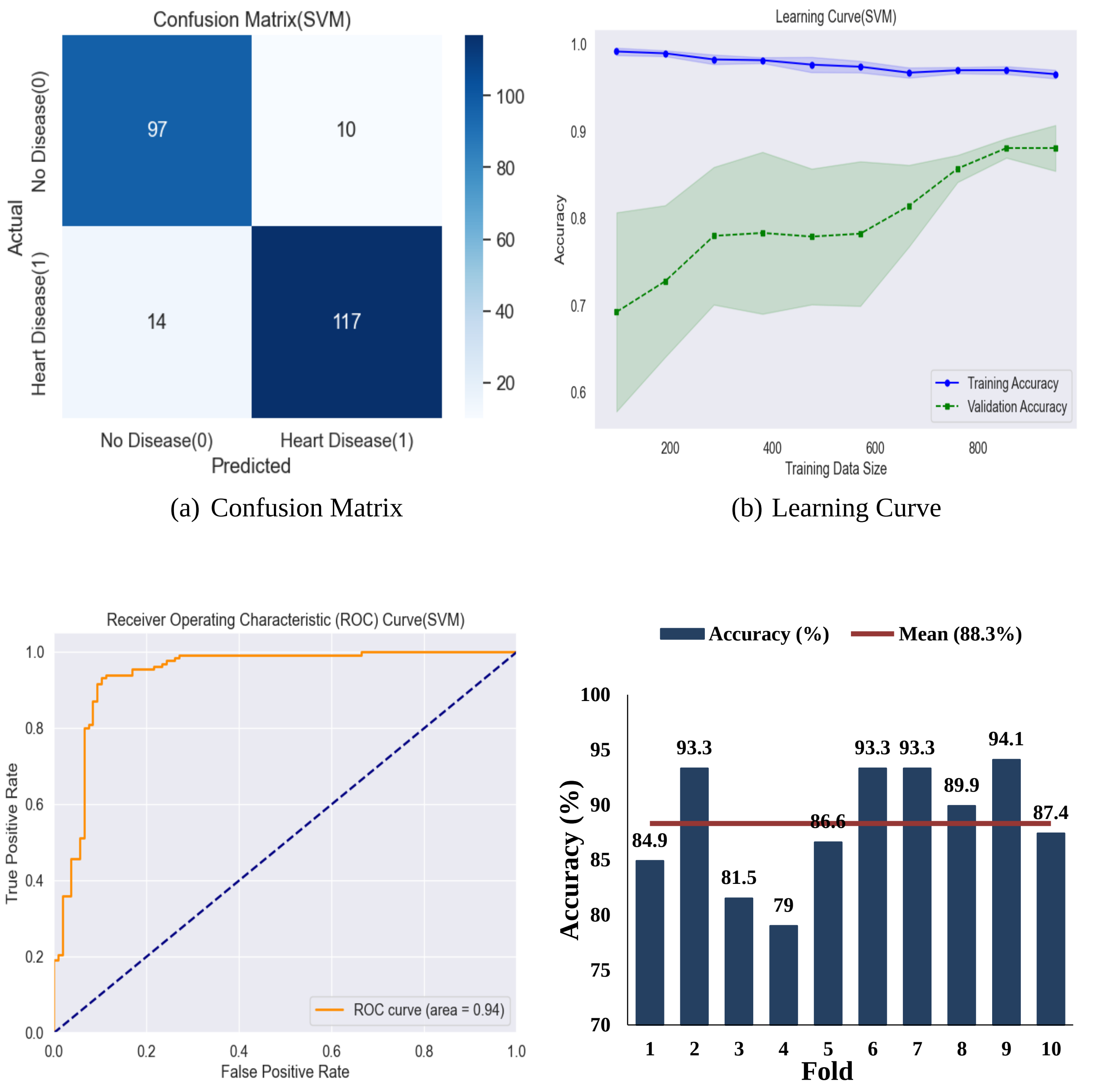
(b) Learning Curve

(c) ROC Curve

(d) Accuracy for Each Fold

Figure 9: Performance Measuring Curves of RF

There are 117 times when the SVM model correctly identifies patients as cardiac patients (TP), 14 times when cardiac patients are identifies as not cardiac patients (FN), 10 times when patients with no cardiac problem are correctly identifies as cardiac patients (FP), and 97 times when not cardiac patients are correctly identifies (TN). We were able to find that the Accuracy, Precision, Recall, F1-score, and AUC number for the SVM model were all 0.90, 0.90, 0.90, 0.90 and 0.94, in that order. This model has an average cross validation accuracy score of 0.883. On the dataset, Figure 10 shows the SVM model's performance measurement plots.



(a) Confusion Matrix

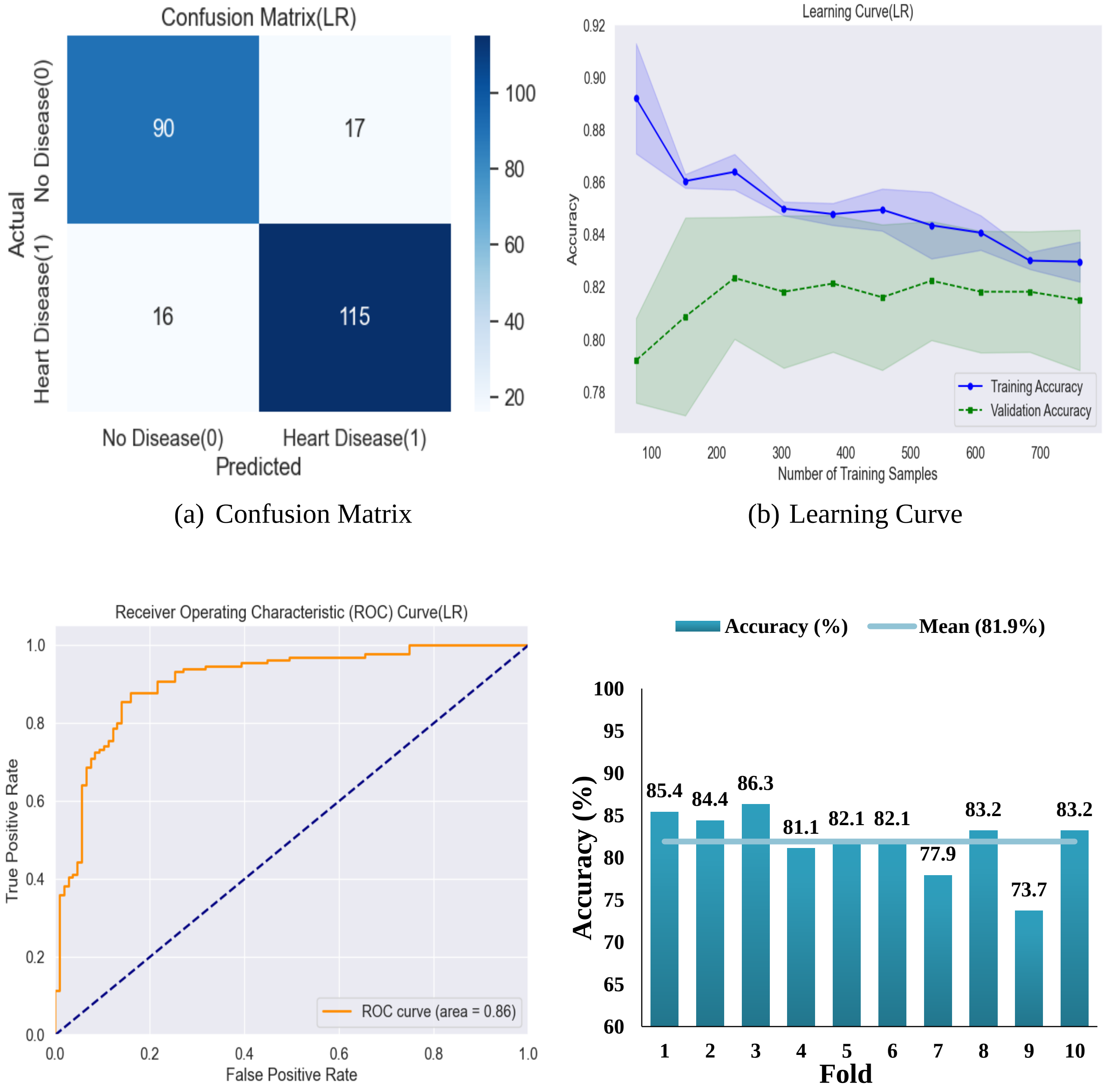
(b) Learning Curve

(c) ROC Curve

(d) Accuracy for Each Fold

Figure 10: Performance Measuring Curves of SVM

The LR model has a TP rate of 115, a FN rate of 16, an FP rate of 17, and a TP rate of 90 for identifying patients with cardiac problems versus those without. The LR model had an Accuracy score of 0.86, a Precision score of 0.85, a Recall score of 0.86, a F1-score of 0.86 and an AUC score of 0.86. In terms of cross-validation accuracy, this model scores an average of 0.819. The performance measurement plots for the LR model are shown in Figure 11 on the dataset.



(a) Confusion Matrix

(b) Learning Curve

(c) ROC Curve

(d) Accuracy for Each Fold

Figure 11: Performance Measuring Curves of LR

The findings for each classifier are outlined in Table 3, which provides a summary of the results obtained. Overall, the results show that the RF, LR, and SVM classifiers have higher ROC-AUC and precision-recall values than the other models. According to the findings, the RF model had the highest accuracy in

CVD prediction was 95%, and the highest cross validation mean accuracy was 91.5%. Figure 12, provides a more in-depth look at the results of this research.

Table 3: Different Performance Measure Values of the ML Models

Classifiers	Accuracy (%)	Mean Cross Validation Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC
RF	95	91.5	95	95	95	0.96
SVM	90	88.3	90	90	90	0.94
LR	86	81.9	85	86	86	0.86

Table 4 provides a comparison of the performances of a number of different classifiers and datasets, with the intention of finding which model delivers the best results under each specific set of conditions. As can be seen in Table 4, the proposed techniques improve upon the results of some of the most recent investigations. Using the IEEEdataport dataset, Yilmaz et al. [8]were able to get 86.1% accuracy for the LR algorithm, 89.7% accuracy for the SVM algorithm, and 92.9% accuracy for the RF algorithm. We have instead achieved 90% accuracy with SVM model, 86% with LR model, and 95% with RF model.

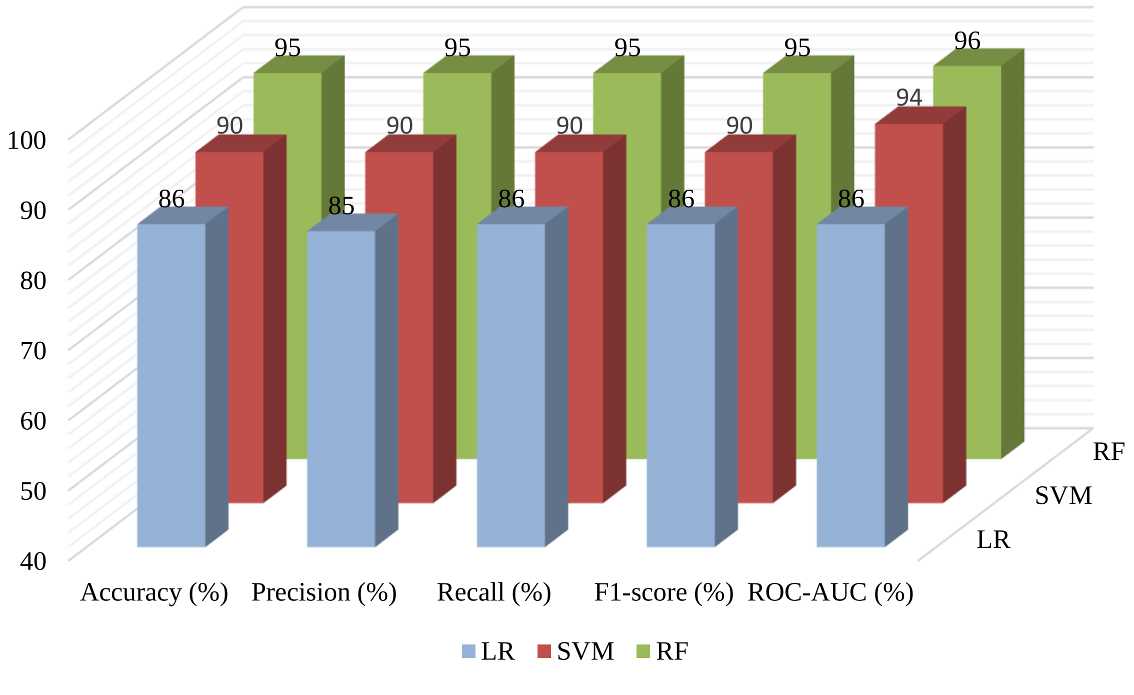


Figure 12: Classification Model Chart

M. Ozcan et al. [12] used the CART model to accurately predict cardiovascular disease on the IEEEdataport dataset, with an 87% success rate. With an accuracy of 89% on the IEEEdataport dataset and 85% on the Cleveland dataset, an RF model for cardiovascular disease prediction has been developed by N. Chandrasekhar [21]. Ghosh et al. [22] used the Cleveland, Long Beach VA, Switzerland, Hungarian, and Stat log datasets and got an 88.65% success rate with the RF method. While using the consolidated dataset from IEEEDataPort for the RF method, we were able to get 95% accuracy.

Table 4. Comparison of Efficiency of this Proposed Approach with Existing Methods

Sl.	Work Ref.	Dataset	Method	Accuracy (%)
1.	Bashir et al. [6]	Dataset from UCI	RF	84.17
2.	Yilmaz et al. [8]	Datasets from IEEEdataport	LR	86.1
			SVM	89.7
			RF	92.9
3.	M. Ozcan et al. [12]	Datasets from IEEEdataport	CART	87
4.	N. Chandrasekhar [21]	Datasets from IEEEdataport	RF	89
		Datasets from Cleveland	RF	85
5.	Ghosh et al. [22]	Cleveland, Long Beach VA, Switzerland, Hungarian, and Stat log datasets.	RF	88.65
6.	Proposed approach	Cleveland, Long Beach VA, Switzerland, Hungarian, and Stat log datasets from IEEEdataport.	RF	95
			SVM	90
			LR	86

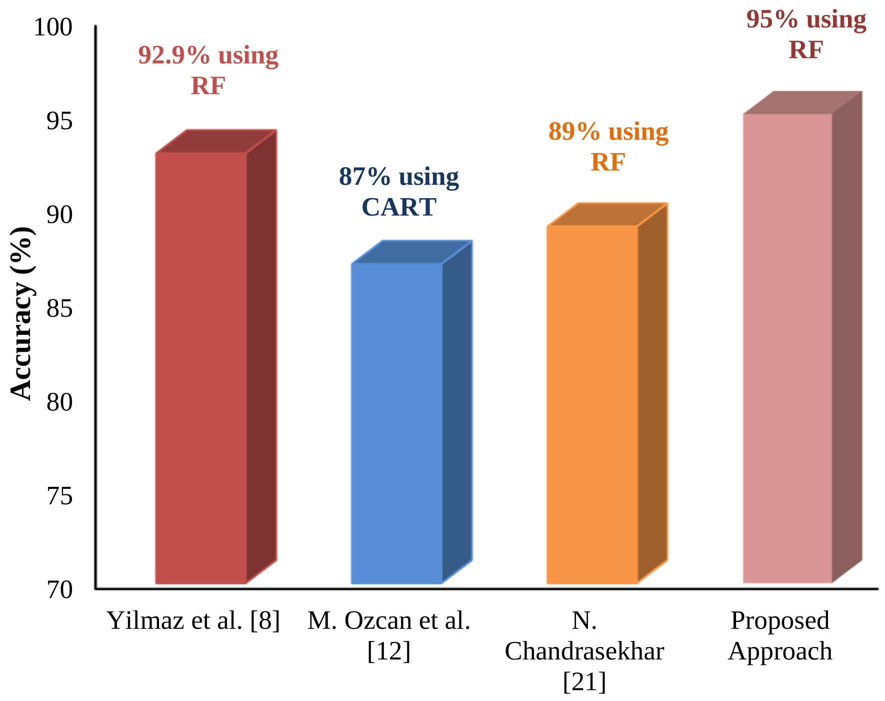


Figure 13: Accuracy Comparison of the Proposed and Related Methods on IEEEDataport Dataset

The comparison of the suggested method to various current works on the IEEEDataport dataset is depicted graphically in Figure 13. In the end, we discover that our suggested model works better than some other models used to predict CVD. Furthermore, we were able to improve our model's hyperparameters using a GridsearchCV method, which led to a better result than some previous work.

4 Conclusion

CVD, one of the leading causes of death, has become more prevalent among the population globally during the past several decades. All parties involved in the health sector must be aware of the potential of prospective algorithms based on machine learning to shape the doctor's perceptions since it could support medical professionals' efforts to create a more favorable environment for the treatment and diagnosis of patients. For this reason, this study proposed ML based predictive model for CVD diagnosis. On a well-known dataset collected from IEEDataport, feature importance analysis through Spearman correlation analysis and XGboost and interactive result visualization were performed through several techniques. For classifying CVD patients, Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR) achieved 95%, 90% and 86% respectively. In future, experiment on the optimal features will be performed and other machine learning as well as deep learning models will also be explored.

References

1. Peng, M., Hou, F., Cheng, Z., Shen, T., Liu, K., Zhao, C., Zheng, W.: Prediction of cardiovascular disease risk based on major contributing features. Sci Rep. 13, 4778 (2023). https://doi.org/10.1038/s41598-023-31870-8.
2. Bhatt, C.M., Patel, P., Ghetia, T., Mazzeo, P.L.: Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms. 16, 88 (2023). https://doi.org/10.3390/a16020088.
3. Cardiovascular Disease: Types, Causes & Symptoms, https://my.clevelandclinic.org/health/diseases/21493-cardiovascular-disease, last accessed 2023/10/12.
4. Amin, M.S., Chiam, Y.K., Varathan, K.D.: Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics. 36, 82–93 (2019). https://doi.org/10.1016/j.tele.2018.11.007.
5. Kausar, N., Palaniappan, S., Samir, B.B., Abdullah, A., Dey, N.: Systematic analysis of applied data mining-based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients. Intelligent Systems Reference Library. 96, 217–231 (2016). https://doi.org/10.1007/978-3-319-21212-8_9.
6. Bashir, S., Khan, Z.S., Hassan Khan, F., Anjum, A., Bashir, K.: Improving Heart Disease Prediction Using Feature Selection Approaches. In: 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST). pp. 619–623. IEEE (2019). https://doi.org/10.1109/IBCAST.2019.8667106.
7. Meshref, H.: Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach. International Journal of Advanced Computer Science and Applications. 10, (2019). https://doi.org/10.14569/IJACSA.2019.0101236.
8. YILMAZ, R., YAĞIN, F.H.: Early Detection of Coronary Heart Disease Based on Machine Learning Methods. Medical Records. 4, 1–6 (2022). https://doi.org/10.37990/medr.1011924.
9. Saqlain, S.M., Sher, M., Shah, F.A., Khan, I., Ashraf, M.U., Awais, M., Ghani, A.: Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. Knowl Inf Syst. 58, 139–167 (2019). https://doi.org/10.1007/s10115-018-1185-y.
10. Dinesh, K.G., Arumugaraj, K., Santhosh, K.D., Mareeswari, V.: Prediction of Cardiovascular Disease Using Machine Learning Algorithms. In: 2018 International Conference on Current Trends

towards Converging Technologies (ICCTCT). pp. 1–7. IEEE (2018). https://doi.org/10.1109/ICCTCT.2018.8550857.

11. Mienye, I.D., Sun, Y., Wang, Z.: Improved sparse autoencoder based artificial neural network approach for prediction of heart disease. Inform Med Unlocked. 18, (2020). https://doi.org/10.1016/j.imu.2020.100307.
12. Ozcan, M., Peker, S.: A classification and regression tree algorithm for heart disease modeling and prediction. Healthcare Analytics. 3, (2023). https://doi.org/10.1016/j.health.2022.100130.
13. Alshaher, H.: Studying the Effects of Feature Scaling in Machine Learning.
14. Zeeshan Ahmad Lodhia, Akhtar Rasool, Gaurav Hajela: A survey on machine learning and outlier detection techniques. IJCSNS International Journal of Computer Science and Network Security. 17, 271–276 (2017).
15. Kumar, S., Chong, I.: Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States. Int J Environ Res Public Health. 15, 2907 (2018). https://doi.org/10.3390/ijerph15122907.
16. Mirtagioglu, H., Mendes, M., Onsekiz, Ç., Üniversitesi, M., Temizhan, E.: Which Correlation Coefficient Should Be Used for Investigating Relations between Quantitative Variables?
17. Abdurrahman, G., Sintawati, M.: Implementation of xgboost for classification of parkinson's disease. In: Journal of Physics: Conference Series. Institute of Physics Publishing (2020). https://doi.org/10.1088/1742-6596/1538/1/012024.
18. Breiman, L.: Random Forests. Mach Learn. 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324.
19. Alty, S.R., Millasseau, S.C., Chowienczyk, P.J., Jakobsson, A.: Cardiovascular disease prediction using support vector machines. In: 2003 46th Midwest Symposium on Circuits and Systems. pp. 376–379. IEEE. https://doi.org/10.1109/MWSCAS.2003.1562297.
20. G, A., Ganesh, B., Ganesh, A., Srinivas, C., Dhanraj, Mensinkal, K.: Logistic regression technique for prediction of cardiovascular disease. Global Transitions Proceedings. 3, 127–130 (2022). https://doi.org/10.1016/j.gltp.2022.04.008.
21. Chandrasekhar, N., Peddakrishna, S.: Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. Processes. 11, 1210 (2023). https://doi.org/10.3390/pr11041210.
22. Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F.M.J.M., Ignatious, E., Shultana, S., Beeravolu, A.R., De Boer, F.: Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques. IEEE Access. 9, 19304–19326 (2021). https://doi.org/10.1109/ACCESS.2021.3053759.