iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6164

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Cardiovascular Disease Prediction Model using Machine Learning Algorithms

Siddhika Arunachalam
M.Tech., Dept. of Computer Engineering, Sardar Patel Institute of Technology, University of Mumbai, Mumbai, India

Abstract: ***A general term for conditions affecting the heart or blood vessels is called as Cardiovascular disease (CVD). It is commonly associated with an increased risk of blood clots and build-up of fatty deposits inside the arteries (atherosclerosis). Sometimes, it can also be associated with damage to arteries in organs such as the brain, kidneys, heart and eyes. CVD is the reason for the highest number of deaths globally and the major cause of death annually. Most cardiovascular diseases can often be prevented by leading a healthy lifestyle and addressing behavioural risk factors such as unhealthy diet and obesity, tobacco use, harmful use of alcohol and physical inactivity using population-wide strategies. Machine Learning can play an important role in predicting cardiovascular disease and such information, if predicted well in advance can provide significant insights to doctors who can then adapt their treatment and diagnosis for each patient accordingly. In the proposed research method, firstly the attributes are selected from the dataset, then data pre-processing takes place which uses techniques such as removal of noisy data, removal of missing data, filling default values if applicable, classification of attributes for prediction and decision making at different levels. Classification, accuracy, sensitivity and specificity analysis is done to obtain the performance of the diagnosis model. A prediction model which predicts whether a person has a heart disease or not and hence provide diagnosis or discussion on the results is proposed. This is accomplished by applying rules to the individual results of classification algorithms such as Gradient Boosting Classifier, Random Forest Classifier, Support Vector Machine, Extremely Randomized Trees Classifier (Extra Trees Classifier), Logistic Regression and Multi-Layer Perceptron (MLP) Classifier obtained on the dataset.***

Keywords: Cardiovascular Disease (CVD), Machine Learning, Gradient Boosting Classifier, Random Forest Classifier, Support Vector Machine, Extra Trees Classifier, Logistic Regression, Multi-Layer Perceptron (MLP) Classifier

I. INTRODUCTION

In today's era as cardiovascular disease is the primary reason for deaths, World Health Organization (WHO) has anticipated that 12 million people die every year worldwide because of this disease. Four main types of cardiovascular diseases are as follows:

1) *Coronary heart disease*- Blocked or reduced flow of oxygen-rich blood to the heart muscle causes coronary heart disease. This puts an increased stress on the heart, and can lead to heart attacks, angina or heart failure.
2) *Cerebrovascular disease* - It is a disease of the blood vessels supplying the brain.
3) *Peripheral arterial disease* - Blockage in the arteries to the limbs, usually the legs leads to Peripheral arterial disease. This can cause cramping or dull leg pain and weakness or numbness in the legs.
4) *Strokes and Transient Ischaemic Attack (TIA)* - When the blood supply to part of the brain is cut off, which could cause brain damage and possibly death, it is called as Stroke. A TIA is similar to Stroke, the only difference is blood flow to the brain is temporarily disrupted.
5) *Aortic disease* - It is a group of conditions affecting the Aorta, which is the largest blood vessel in the body. Blood is carried from the heart to the rest of the body.

A. Some Other Types Of Cardiovascular Diseases Are

1) *Rheumatic Heart Disease* – In this type of disease damage to the heart valves and heart muscles are caused due to rheumatic fever.
2) *Congenital Heart Disease* - It is a malformation of heart structure existing right from birth.
3) *Deep vein Thrombosis and Pulmonary Embolism* – It is a condition which causes blood clots in the leg veins, which can dislodge and move to the heart and lungs.

B. Some Of The Risk Factors That Can Increase The Chances Of Developing Cardiovascular Disease Are Outlined Below

1) *High Blood Pressure* – It is one of the most important risk factors for CVD. Having too high blood pressure can damage the blood vessels.
2) *Smoking* - Smoking and other tobacco use is also considered as a significant risk factor for CVD as the harmful substances in tobacco can narrow and damage the blood vessels.
3) *High Cholesterol* – Having high cholesterol can cause narrowing of blood vessels and thereby increasing the risk of developing a blood clot.
4) *Diabetes* - High blood sugar levels can cause damage to the blood vessels, which increases the possibility of them to become narrowed.
5) *Inactivity* – Not exercising regularly can generally increase the chances of having high cholesterol levels, high blood pressure and being overweight which are eventually the risk factors for CVD.
6) *Being Overweight or Obese* – It increases the risk of high blood pressure and diabetes which can ultimately lead to CVD.
7) *Family history of CVD* – Having a family history of CVD also increases the risk of developing CVD for the person. A family history of CVD is considered if either the father or brother are diagnosed with CVD before they are 55 or the mother or sister are diagnosed with CVD before they are 65.

Symptoms of cardiovascular disease generally varies depending on the specific condition but some common symptoms include pain or pressure in the chest, pain or discomfort in the left shoulder, arms, elbows, back or jaw, shortness of breath, nausea and fatigue, dizziness or light-headedness and cold sweats.

C. Some of the ways of Preventing or Reducing the risk of CVD are as Follows

1) *Stop Smoking* - Smoking is considered as a key risk factor for nearly all forms of CVD. Although quitting can be difficult, taking necessary steps to do so can significantly lower its damaging effects on the heart.
2) *Follow a heart-healthy Diet* - Eating foods that contain omega-3, such as oily fish and polyunsaturated fats, alongside vegetables and fruits can support heart health and decrease the risk of CVD. Also, reducing the consumption of salt, processed food, added sugar and saturated fat has a similar effect.
3) *Get Regular Exercise* - The American Heart Association (AHA) recommends doing 150 minutes of moderate-to-intense physical activity weekly.
4) *Manage Body Weight* - The National Institute of Diabetes and Digestive and Kidney Disorders (NIDDK) advise that if a person lose 5 to 10% of their body weight, they may lower their risk of developing CVD.

Numerous tests comprising of chest X-rays, angiography, echocardiography and stress test support the judgment and early prevention of heart disease complications. Nevertheless, these tests are costly and involve availability of precise medical equipment. The examination of the illness is a complex mechanism. It should be measured flawlessly and accurately. It would be enormously beneficial if the machine learning techniques are combined with the medical information system. This paper proposes different machine learning techniques to predict potential cardiovascular diseases in people based on the attributes present. In order to conduct this analysis, publicly available Cleveland dataset for cardiovascular disease is used.

II. PREVIOUS WORK

In [1], arising possibilities of heart disease are predicted using data mining techniques. The chances of occurring heart disease is provided in terms of percentage. The medical parameters classified in the datasets are evaluated using data mining classification technique. Python programming is used to process the datasets using two machine learning algorithms namely Naïve Bayes Algorithm and Decision Tree Algorithm. Among these two algorithms, the best algorithm is decided in terms of accuracy level of heart disease.

In [2], various tools and algorithms used for prediction of heart disease is discussed. After pre-processing and splitting, various classification algorithms are used such as Decision tree, K-nearest Neighbour and K-means Clustering. AdaBoost, a technique for increasing performance of decision trees on binary classification problems is also used.

A multifaceted and comprehensive review of all related studies that were published between 1992 and 2019 for Machine Learning-based Coronary Artery Disease (CAD) diagnosis is conducted in [3]. The impacts of various factors, such as dataset characteristics (sample size, geographical location, stenosis of each coronary artery and the features) and applied Machine Learning techniques (performance metrics, feature selection, and method) are examined in detail. Lastly, the important challenges and limitations of Machine Learning-based CAD diagnosis are discussed.

Different algorithms of Decision Trees classification are compared looking for better performance in diagnosis of heart disease using WEKA tool in [4]. The algorithms that are tested is J48 algorithm, Random Forest algorithm and Logistic model tree algorithm. The goal of this research was to extract hidden patterns by the application of data mining techniques, which are important to heart diseases and to predict the existence of heart disease in patients where this existence is valued from not present to likely present.

The main goal of paper [5] was to review some of the current research on predicting heart diseases using different data mining techniques such as Decision Tree, C4.5, K-means Algorithm, ID3 Algorithm, Support Vector Machine(SVM), Naive Bayes (NB), Artificial Neural Network (ANN), CART and Random Forest and also analyse the various combinations of these mining algorithms used. The uses of some data mining tools such as WEKA, RapidMiner, TANAGRA, Apache Mahout and MATLAB were also discussed and concluded with techniques are effective and efficient.

A detailed description of Decision tree classifier and Naïve Bayes algorithm is done in [6] for the prediction of heart disease. Applying Decision tree and Naïve Bayes with information gain calculations provide better results in the heart disease diagnosis and better accuracy as compared to other classifiers such as Neural network, KNN, SVM and binary discretization with Gain Ratio. Also, a greater number of attributes consideration results in improvement in accuracy. Some experiment has been conducted to compare the execution of data mining technique on the dataset, and the result reveals that Decision Tree outperforms the Bayesian classification algorithm.

In [7], the main aim was to provide accurate and immediate disease prediction to the users as they enter the symptoms and prediction of severity of disease. Different machine learning algorithms like Naïve Bayes Algorithm, Decision Tree Algorithm, K-Nearest Algorithm are used for getting accurate predictions. An application is developed where the severity of disease is predicted and also the doctor's consultation is provided along with providing drug consultation of disease predicted.

In [8], data mining is applied to the database to extract a hidden pattern from the clinical dataset. To achieve the highest accuracy, comparison of all available classification algorithms is done. The dataset is pre-processed using different supervised and unsupervised algorithms to increase the correctness of the solution. Different methods for enhancing the K-means clustering algorithm are also discussed. Accuracy, efficiency and performance are improved using these methods. Lastly, the classifiers are developed with the help of Logistic regression by using the data extracted from K-Means Clustering. This research also concludes that the techniques implemented in the design of classifier perform well in terms of classification results as compared to the clustering techniques.

III. PROPOSED METHODOLOGY

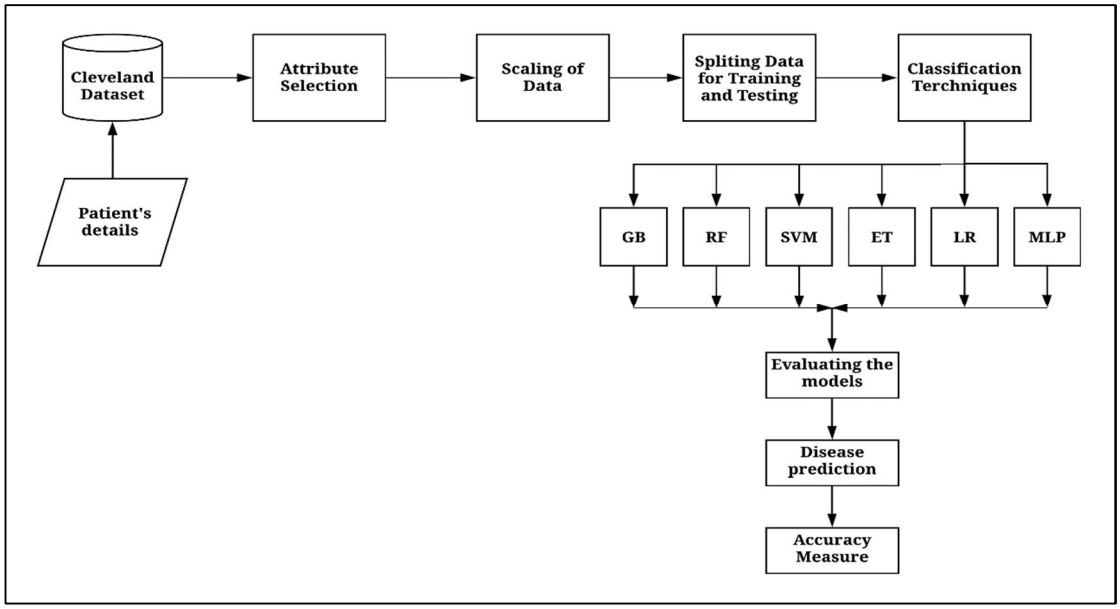


Fig. 1 Proposed System

Figure 1 shows the proposed methodology for predicting Cardiovascular disease and each step in detail are described as follows:

A. Data Source and Attribute Selection

Clinical databases have collected a substantial amount of information about the patient's medical conditions. Datasets were obtained from the Cleveland, Hungarian, Switzerland, Long Beach VA heart disease database present in the UCI machine Learning Repository (Center for Machine Learning and Intelligent Systems). It has a total of 300 unique instances with a total of 76 attributes. The following Table 1 shows the list of 14 attributes which are considered for this system and are commonly used for research till date. Datasets are used to extract the patterns related to the disease. The records are split into two datasets: training dataset and testing dataset.

TABLE I Cleveland dataset 14 features and descriptions

#	Feature	Description	Type	Units	Values
1	age	Age	continuous	years	
2	sex	Sex	categorical	#	
					1 = male
					0 = female
3	cp	Chest Pain Type	categorical	#	
					1 = typical angina
					2 = atypical angina
					3 = non-anginal pain
					4 = asymptomatic
4	trestbps	Resting Blood Pressure	continuous	mmHg	
5	chol	Serum Cholesterol	continuous	mg/dl	
6	fbs	Fasting Blood Sugar	categorical	#	
					1=True
					0=False
7	restecg	Resting Electrocardiographic	categorical	#	
					0=normal
					1=ST-T wave abnormal
					2=left ventricular hypertrophy Estes
8	thalach	Exercise Max Heart Rate Achieved	continuous	bpm	
9	exang	Exercise Induced Angina	categorical	#	
					1=yes
					0=no
10	oldpeak	ST depression induced by Exercise relative to Rest	categorical	#	
11	slope	Slope of Peak Exercise ST Segment	categorical	#	
					1=upsloping
					2=flat
					3=downsloping
12	ca	# of Major Vessels colored by Fluoroscopy	categorical	#	
					0= zero fluroscopy colored Major Vessels
					1= one fluroscopy colored Major Vessels
					2= two fluroscopy colored Major Vessels
					3= three fluroscopy colored Major Vessels
13	tha	Thalassemia	categorical	#	
					3=normal
					6=fixed defect
					7=reversable defect
14	num	diagnosis	categorical	#	
					0 = No Disease
					> 0 = Disease

B. Pre-processing of Data

1) *Cleaning:* The value of diagnosis column is changed to 0 or 1 for binary classification. After the inspection of information about the data to understand its types, two features were found that contained the 'object' values. They were: 1) # Major Vessels colored by Fluoroscopy and 2) Thalassemia. These two features contained 4 and 2 counts of missing values respectively. As these instances are not considered as a significant number, these instances were removed.
2) *Feature Engineering:* Some of the categorical values described have only a few unique values. Categorical Encoding is used which makes the Machine Learning algorithms to not overfit to unique values. Converting these into binary values allows the Machine Learning algorithms to process the data in a less biased manner without losing any of the information.

Figure 2 shows the histogram of Cleveland dataset after Cleaning and Feature Engineering. Some obvious relationships and outcome of Heart Disease can be observed in this figure. For example, looking at the subplot of Age in the top left, after around age 60, the number of people with Heart Disease nearly doubled as compared to No Disease at the same age. In addition, looking at the subplot of Sex, next to the subplot of Age, it can be seen that men (value = 1) have a higher number of Heart Disease than women (value=0). The Machine Learning algorithms can pick up on these relationships for making a diagnostic.

Fig. 2 Histogram of Cleveland dataset

C. Scale the Data

The scaling of data is important so that the Machine Learning algorithms does not overfit to the wrong features. Using the function of MinMaxScaler() , the values are scaled for each feature based on the minimum and maximum values between 0 and 1. This saves the information from being lost and also allows the Machine Learning algorithms to correctly train the data.

Figure 3 shows the Heatmap with Pearson Correlation Coefficient for Features after the scaling of data. A strong correlation value near 1 is indicated by a Pearson Correlation Coefficient. Therefore, from the Heatmap, values that correlate most with the first column, “diagnosis,” is important for training the Machine Learning algorithms. Values near 1 between other features are not ideal because the information for the Machine Learning to train is already present.

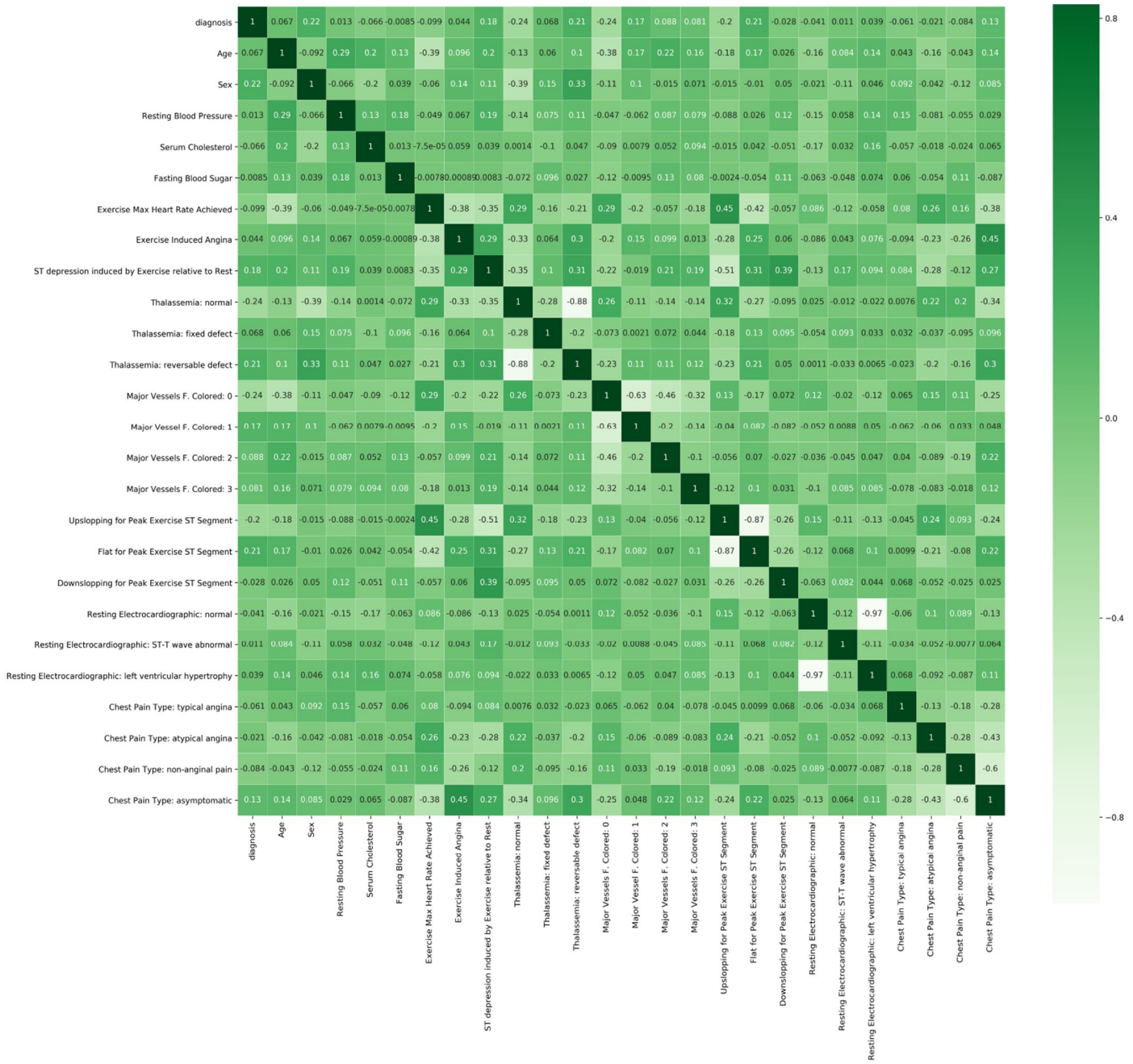


Fig. 3 Heatmap of Cleveland dataset

D. Split Data for Training

The Data was split into 80% training data and 20% testing data after dropping 6 instances with missing values.

E. Classification

The training data is trained by using five different machine learning algorithms such as Gradient Boosting Classifier, Random Forest Classifier, Support Vector Machine, Extra Trees Classifier, Logistic regression and Multi-layer Perceptron (MLP) Classifier. A sensitivity study using different Hyperparameters of the algorithms are iterated with GridSearchCV in order optimize each model [9]. The best model is the one that has the highest accuracy without overfitting by looking at both the training data and the validation data results. The model that has the highest accuracy without overfitting is the best one. This is decided by looking at both the training and the validation data results. These models are evaluated through k-fold Cross-Validation with k-fold = 10 using GridSearchCV, which iterates on different algorithm's hyperparameters. Table II provides the summary of algorithms and their corresponding hyperparameters used with GridSearchCV.

TABLE III Summary of algorithms and their hyperparameters

#	Algorithm	Parameters	New Parameters
1	Gradient Boosting Classifier	loss = deviance	
		learning_rate = 0.1	learning_rate = 0.01, 1
		n_estimators=500	
		max_depth=3	max_depth=1
		max_features='log2'	
2	Random Forest Classifier	n_estimators=500	n_estimators=1000
		max_features=0.25	
		criterion='entropy'	
3	SVC	C=0.01	C=1, 10
		gamma=0.1	
		kernel="poly"	kernel='linear','rbf'
		degree=3	
		coef0=10.0	
4	Extra Trees Classifier	n_estimators=1000	
		max_features='log2'	max_features=0.25
		criterion='entropy'	
5	Logistic Regression	C=1.5	C=0.001,0.01,0.1,1,10,100,1000
		penalty='l1'	
		fit_intercept=True	
6	MLP Classifier		hidden_layer_sizes=100
			activation='relu','tanh','logistic'
			learning_rate='constant','invscaling','adaptive'

1) *Gradient Boosting Classifier (GBC):* The main characteristics of Gradient Boosting is to optimize a loss function, make a weak learner to do predictions and adding weak learners to an additive model to minimize the loss function. The loss function depends on the type of the problem being solved and it must be differentiable. The Decision trees are considered as the weak learner in Gradient Boosting. Specifically, regression trees are used that gives real values as output for splits and whose output can be added together, allowing successive models outputs to be added and correct the residuals in the predictions. Trees are constructed in a greedy manner, to minimize the loss or to choose the best split points based on purity scores like Gini [10]. In additive model, Trees are added one at a time and existing trees in the model are not changed. To minimize the loss when adding trees, gradient descent procedure is used. To implement a gradient boosting classifier, the number of steps carried out are as follows:

a) Fit the model
b) Tune the model's parameters and Hyperparameters
c) Make predictions
d) Interpret the results

Here, using gradient boosting technique a variable importance of the attribute is provided that is related to predict the heart disease in this dataset.

2) *Random Forest Classifier (RFC):* The Random Forest Classifier is an ensemble-tree based learning algorithm. It is a set of decision trees from randomly selected subset of training set [16]. The votes from different decision trees are aggregated to decide the final class of the test object. The algorithm works in four steps as follows:

a) Select random samples from a given dataset.
b) Construct a decision tree for each sample and get a prediction result from each decision tree.
c) Perform a vote for each predicted result.
d) Select the prediction result with the most votes as the final prediction.

As the Random Forest Classifier takes the average of all the predictions, the biases are cancelled which helps it to not suffer from overfitting problem. Another advantage is that it can also handle missing values. It is considered to be a highly accurate and robust method because of the number of decision trees participating in the process. Figure 4 shows the Random Forest Tree from the multiple decision trees derived from the dataset.

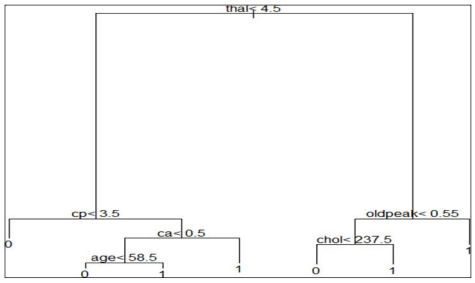


Fig. 4 Random forest tree

3) *Support Vector Machine (SVM):* Support Vector Machine is a machine learning algorithm which is used for both classification and regression tasks. It is widely used for classification objectives as it produces significant accuracy with reduced computation power [11]. This classifier aims at forming a hyperplane that can separate the classes as much as possible by adjusting the distance between the data points and the hyperplane. It plots a hyperplane for every attribute as a coordinate that is present in the dataset. Classification is performed by identifying the hyperplane that divides one class from the other class. Hyperplane can be decided based on several kernels. For this system four kernels are used namely, 'linear', 'rbf' and 'poly'.

4) *Extra Trees Classifier (ETC):* Extremely Randomized Trees Classifier (Extra Trees Classifier) is an ensemble machine learning algorithm that combines the predictions from several decision trees. It is quite similar to a Random Forest Classifier and only differs in the manner from which the construction of the decision trees in the forest is made. Each Decision Tree in the Extra Trees Forest is constructed from the original training dataset [12]. Then, at each test node, each tree is provided with a random sample of k features from the feature-set. From this, each decision tree must select the best feature to split the data based on the Gini Index. This random sample of features creates multiple de-correlated decision trees. During the construction of the forest, for each feature, the feature selection using the forest structure is performed by normalizing the total reduction in the Gini Index used in the decision of feature of split [15]. It uses averaging to improve the predictive accuracy and also control over-fitting.

5) *Logistic Regression (LR):* Logistic Regression is a machine-learning technique to classify records of a dataset based on the values of the input field. It forecasts a dependent variable based on one or more set of independent variables to predict outcomes. It uses the cost function defined as Sigmoid function. To map the predicted values to probabilities, Sigmoid function is used. This function maps a real value into another value between 0 and 1.

Formula of a sigmoid function is given as: $f(x) = \frac{1}{1+e^{-(x)}}$

Logistic regression is less prone to over-fitting but it can lead to over-fitting in high dimensional datasets. Regularization (L1 and L2) techniques can be considered to avoid over-fitting in this scenario. An advantage of using Logistic Regression is that it is extremely easy to implement and very efficient to train.

6) *Multi-layer Perceptron Classifier (MLPC):* Multi-layer Perceptron Classifier implements a multi-layer perceptron (MLP) algorithm that trains using Backpropagation. It is known as feed-forward neural network having one or more hidden layers. It maps sets of input data onto a set of appropriate outputs [13]. The neural network in this system accepts the clinical features as input and it is trained using back-propagation algorithm to predict the presence or absence of heart disease in a patient. Each node is a neuron that uses a nonlinear activation function, except for the input nodes. Figure 5 shows the architecture for Multi-layer Perceptron Neural Network that can have one or more hidden layers.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

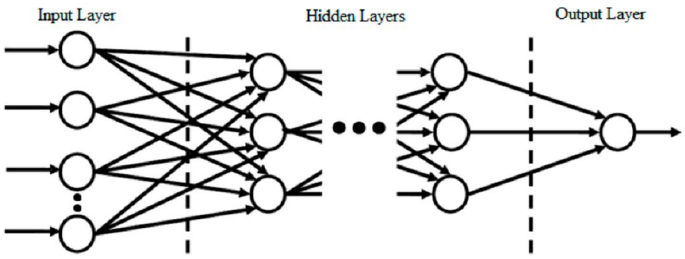


Fig. 5 Multi-layer Perceptron neural Network Architecture

F. Evaluating the Models

When evaluating the models with different machine learning algorithms applied, each algorithm provides different accuracy rate for the attributes considered which is the cause of the cardiovascular disease. Sensitivity and Specificity can be derived from the Confusion Matrix plots [14]. Sensitivity is the approach that identify the people with the cardiovascular disease (true positive rate) and specificity is the approach that identify the people without the cardiovascular disease (true negative rate). When diagnosing cardiovascular disease, it is taken care that there are not too many false-positives or false-negatives present. Hence, the highest overall accuracy model is chosen (accuracy is the sum of the diagonals on the confusion matrix divided by the total).

Positive Predictive Value (PPV) is the probability that following a positive test result, that individual will truly have that specific disease.

Negative Predictive Value (NPV) is the probability that following a negative test result, that individual will truly not have that specific disease.

	Disease Present	Disease Absent
Positive Test Result	True Positive (TP)	False Positive (FP)
Negative test Result	False Negative (FN)	True Negative (TN)

$$True\ Negative\ Rate\ (Specificity) = \frac{P}{P+Q}$$

$$False\ Positive\ Rate\ (1 - Specificity) = \frac{Q}{Q+P}$$

Sums to 1

$$True\ Positive\ Rate\ (Sensitivity) = \frac{R}{R+S}$$

$$False\ Negative\ Rate\ (Specificity) = \frac{S}{S+R}$$

Sums to 1

$$Positive\ Predictive\ Value = \frac{R}{R+Q}$$

$$Negative\ Predictive\ Value = \frac{P}{P+S}$$

Where, P = Number of true negatives, Q = Number of false positives and R = Number of true positives, S = Number of false negatives.

Figure 6 (a), (b), (c), (d), (e), (f) shows the Confusion Matrix plots for the various Classifier algorithms such as Random forest, Gradient Boosting, Support Vector machine, Extra Trees, Logistic Regression and Multi-layer Perceptron used in this system respectively.

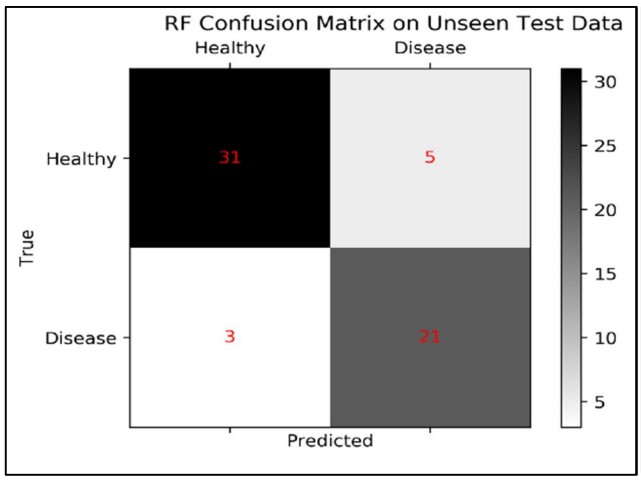


Fig. 6 (a) Random Forest Classifier Confusion Matrix

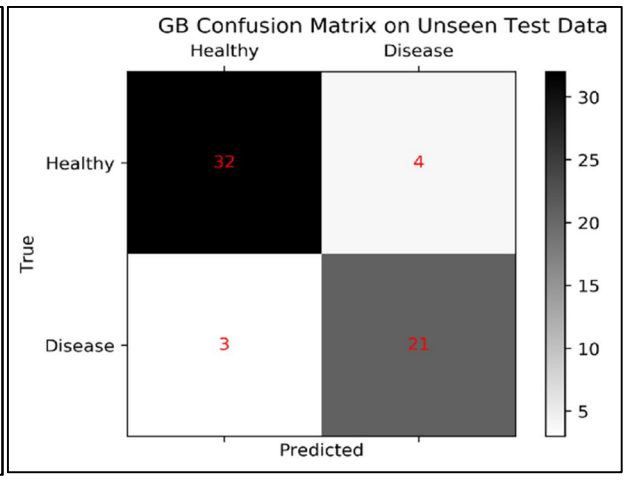


Fig. 6 (b) Gradient Boosting Classifier Confusion Matrix

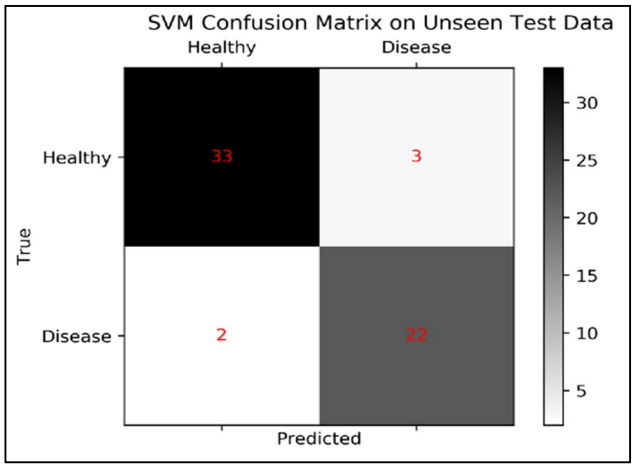


Fig. 6 (c) Support Vector Machine Confusion Matrix

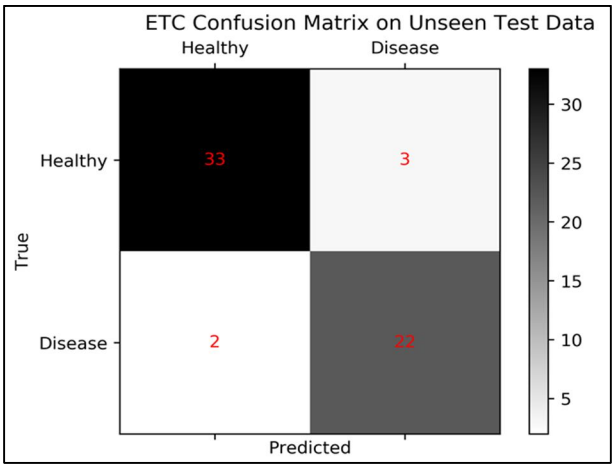


Fig. 6 (d) Extra Trees Classifier Confusion Matrix

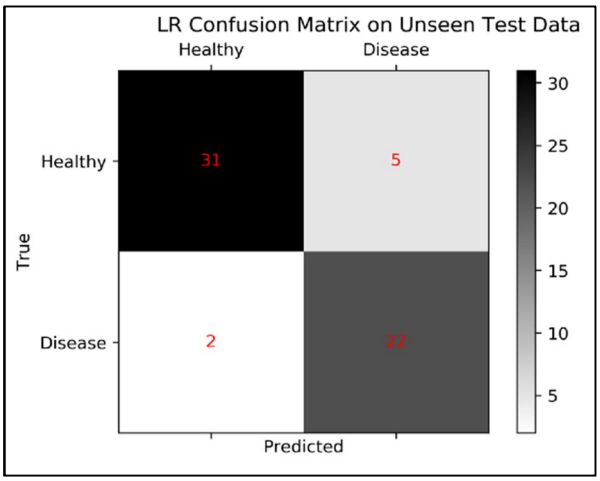


Fig. 6 (e) Logistic Regression Confusion Matrix

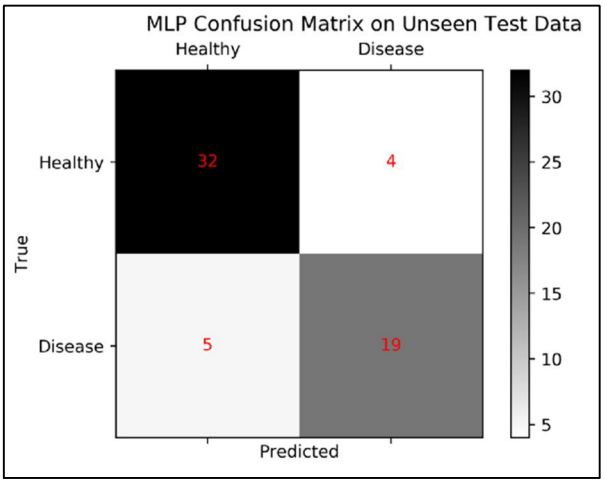


Fig. 6 (d) Multi-layer Perceptron Classifier Confusion Matrix

Figure 7 (a), (b), (c) shows the Variable Importance Plots. Most of the ensemble models have a parameter called 'predict_proba', which gives the most significant features of the model as the output. It is based on their probability through majority vote via either the Gini Index or Entropy.

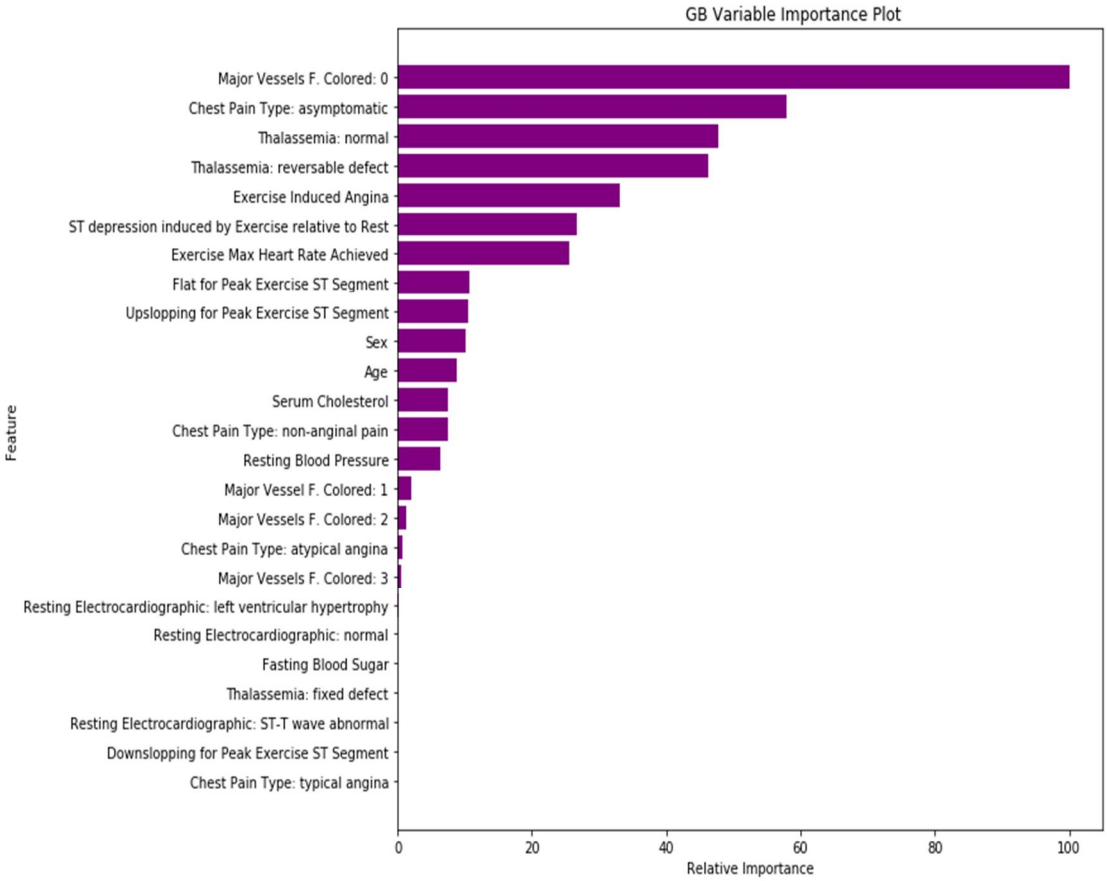


Fig. 7 (a) Gradient Booster Classifier Variable Importance Plot

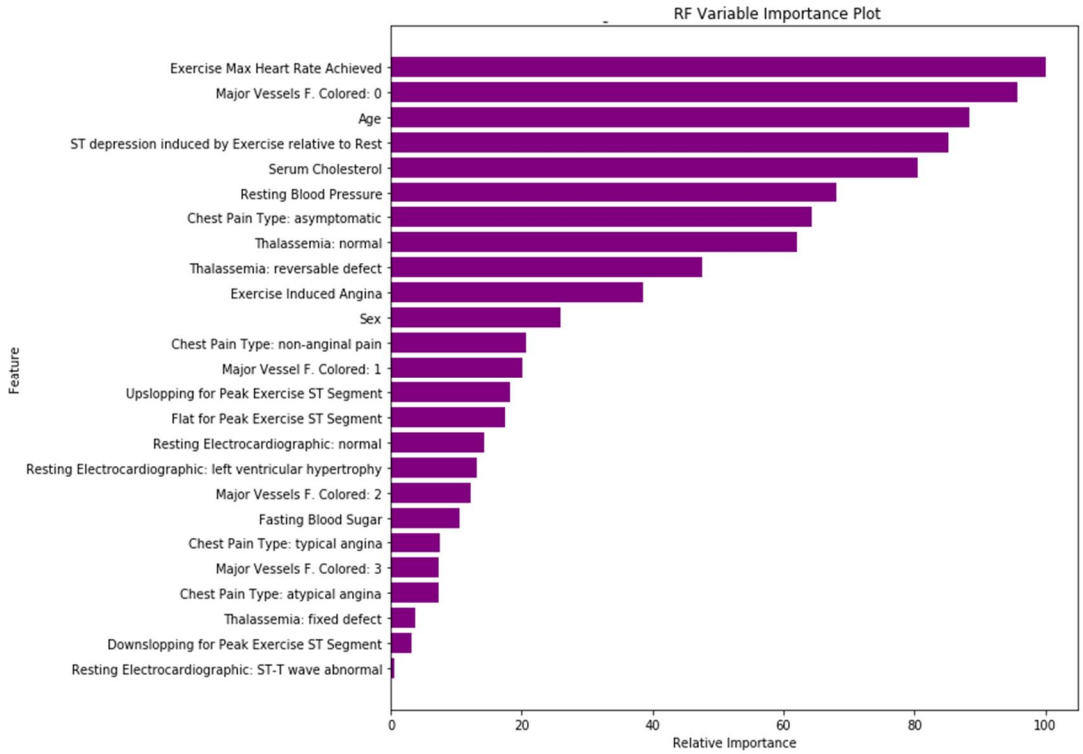


Fig. 7 (a) Random Forest Classifier Variable Importance Plot

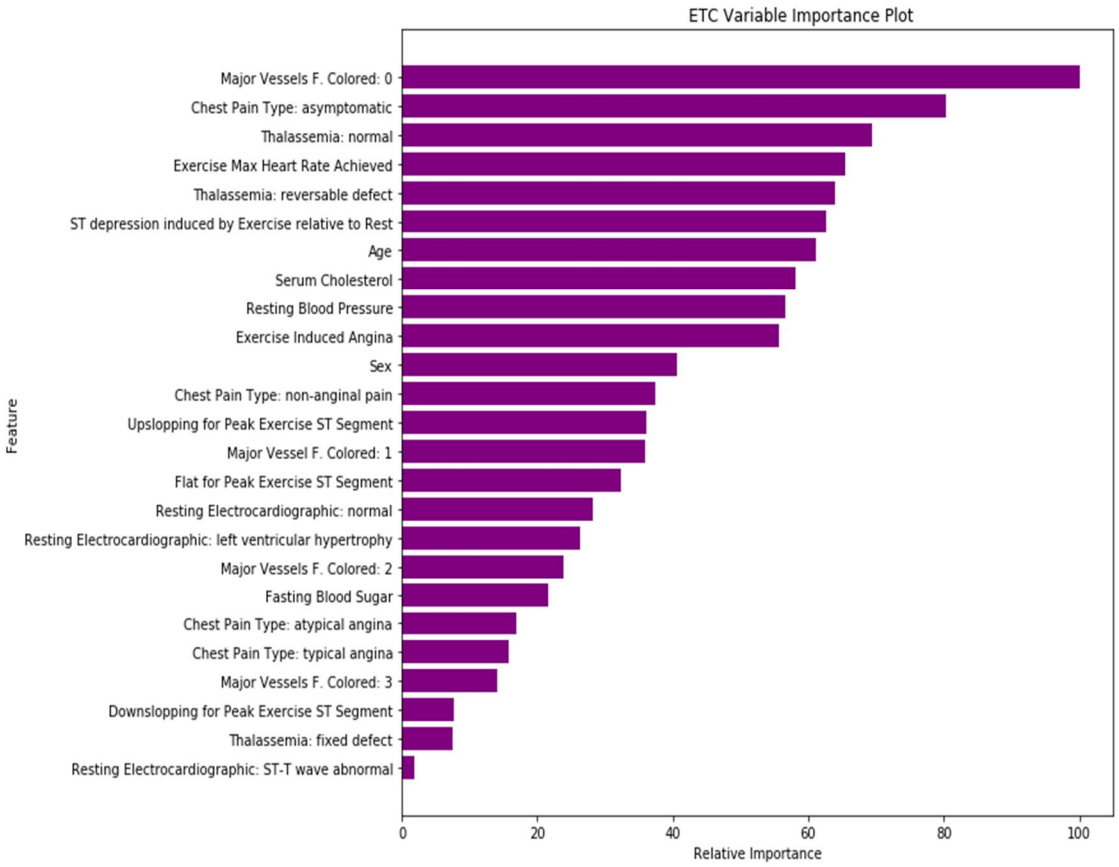


Fig. 7 (c) Extra Trees Classifier Variable Importance Plot

IV. RESULTS AND DISCUSSION

Here, the outputs and the accuracies generated by the Classifiers are reviewed and the results are displayed. All of the models performed well after fine tuning of their hyperparameters, but the best model is considered as the one with highest accuracy. In this analysis, Support Vector Machine (SVM) and Multilayer Perceptron Classifier (MLP) tied for the highest accuracy, 91.7%. The following Table III shows the comparison of the accuracies of the various classifier algorithms used and also some of the other statistical results such as False Positive, False Negative, true positive, True Negative, Sensitivity and Specificity.

TABLE IIIII Comparison of Accuracy and other Statistical Results

Sr. No.	Name of Classifier Algorithm	Accuracy	False Positive [Did not actually have Heart Disease]	False Negative [Actually has Heart Disease]	True Positive	True Negative	Sensitivity	Specificity
1	Gradient Boosting	0.867	5	3	21	31	0.875	0.861
2	Random Forest	0.867	5	3	21	31	0.875	0.861
3	Support Vector Machine	0.917	3	2	22	33	0.917	0.917
4	Extra Trees Classifier	0.867	5	3	21	31	0.875	0.861
5	Logistic Regression	0.883	5	2	22	31	0.917	0.861
6	Multilayer Perceptron	0.917	4	1	23	32	0.958	0.889

From Figure 8, it is seen that the model's Receiver Operating Characteristic (ROC) curves have good Area Under the Curves (AUC) because their values are greater than 90% (close to 95% for all), which means that all would serve as excellent diagnostics. This is also shown by their high Specificity and Sensitivity values. The SVM has a higher AUC, 96.8%, over MLP, which has an AUC of 93.6%. This means SVM is the better diagnostic model in this case.

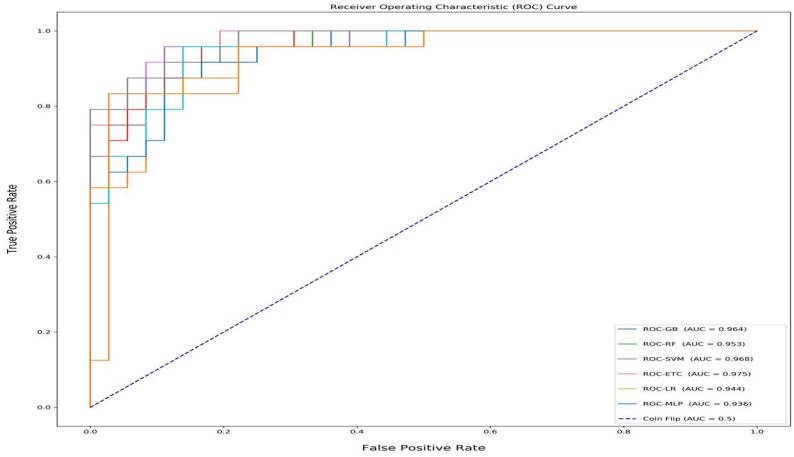


Fig. 8 Receiver Operating Characteristic (ROC) Curve

The user interface is designed in a way that makes it effective for the use by the user. The design of the user interface for the Cardiovascular disease prediction system is shown in Figure 9 and Figure 10. Figure 9 shows the probability of heart disease for a patient based on the algorithms used. Here, considering the 14 attributes of the patients that are discussed earlier, appropriate result that is predicted with the help of algorithms. Figure 10 shows the explanation of the medical range value for the probabilistic value based on the results displayed.

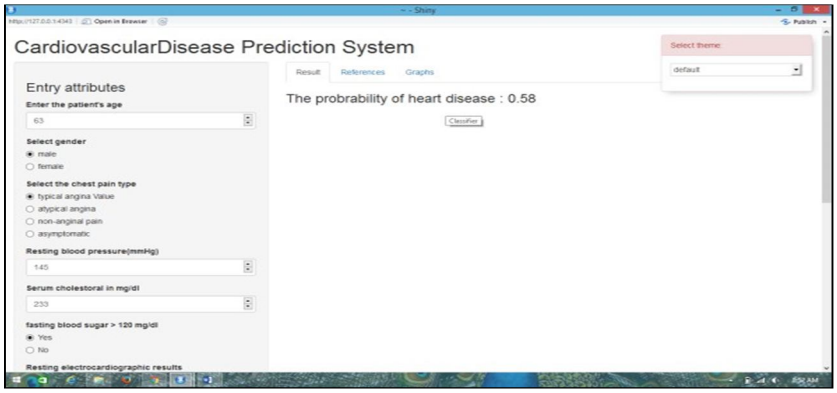


Fig. 9 Front-end view of the system

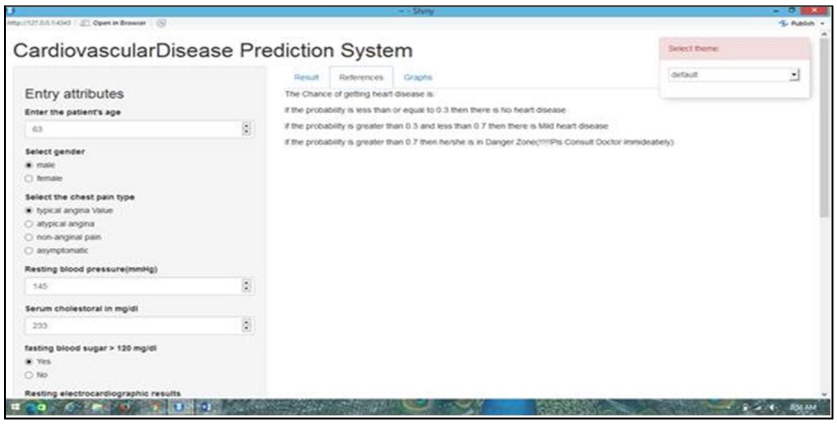


Fig. 10 Explanation of the results displayed

V. CONCLUSION AND FUTURE WORK

In this paper, six classification algorithms are used such as Gradient Boosting Classifier, Random Forest Classifier, Support Vector Machine, Extremely Randomized Trees Classifier (Extra Trees Classifier), Logistic Regression and Multi-Layer Perceptron Classifier for the classification of heart disease after analyzing the 14 attributes of the Cleveland dataset. Also, the accuracies of these algorithms are obtained with the help of Sensitivity and Specificity and it can be concluded from the observations that both SVM and MLP had the highest accuracy of 91.7%. The high Area Under the Curves (AUC) for Receiver Operating Characteristic (ROC) which had values greater than 90% makes it as an exceptional diagnostic for the Heart disease. Furthermore, a front-end system was also developed which displayed the probability of the heart disease after entering the details based on the 14 attributes. The results are formulated based on the algorithms used and also interpreted the meaning of the ranges of the probabilities. Different ensemble methods of these algorithms used can be included in the future work which can advance to better performance with more parameter settings for these algorithms.

REFERENCES

[1] Shadman Nashif, Md. Rakib Raihan, Md. Rasedul Islam, Mohammad Hasan Imam, "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System", Scientific Research Publishing, World Journal of Engineering and Technology, ISSN Online: 2331-4249 ISSN Print: 2331-4222 November 2018, Vol. 6, pp. 854-873.

[2] Rudra A. Godse, Smita S. Gunjal , Karan A. Jagtap , Neha S. Mahamuni , Suchita Wankhade, "Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively" International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE) ISSN (Online) 2278-1021 ISSN (Print) 2319-5940 Vol. 8, Issue 12, December 2019, pp. 50-52.

[3] Dinesh Kumar G, Arumugaraj K, Santhosh Kumar D, Mareeswari V "Prediction of Cardiovascular Disease Using Machine Learning Algorithms", Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India, 978-1-5386-3702-9, IEEE, pp. 01-07.

[4] R. Alizadehsani, et al. "Machine learning-based coronary artery disease diagnosis: A comprehensive review", Computers in Biology and Medicine, Elsevier Ltd. June 2019, pp. 01-14.

[5] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019, pp. 944 - 950.

[6] Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee, "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Advances in Computational Sciences and Technology ISSN: 0973-6107, Volume 10, Number 7 (2017), pp. 2137-2159.

[7] Sonam Nikhar, A.M. Karandikar, "Prediction of Heart Disease Using Machine Learning Algorithms", International Journal of Advanced Engineering, Management and Science (IJAEMS), Vol-2, Issue-6, June- 2016, ISSN: 2454-1311, pp 617-621.

[8] Amin Ul Haq , Jian Ping Li , Muhammad Hammad Memon , Shah Nazir and Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", Hindawi Mobile Information Systems, 2018, pp. 01-21.

[9] Reetu Singh, E. Rajesh, "Prediction of Heart Disease by Clustering and Classification Techniques", International Journal of Computer Sciences and Engineering, Vol.-7, Issue-5, May 2019, E-ISSN: 2347-2693, pp. 861-866.

[10] Jaymin Patel, Tejal Upadhyay, Samir Patel, "Heart Disease Prediction Using Machine learning and Data Mining Techniques", International Journal of Computer Sciences and Communications (IJCSC), Volume 7, Issue 1, March 2016, ISSN 0973-7391, pp. 129-137

[11] Randal S. Olsony_, William La Cavay_, Zairah Mustahsan, Akshay Varik, and Jason H. Moore, "Data-driven advice for applying machine learning to bioinformatics problems", Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution, January 2018, pp. 01-12.

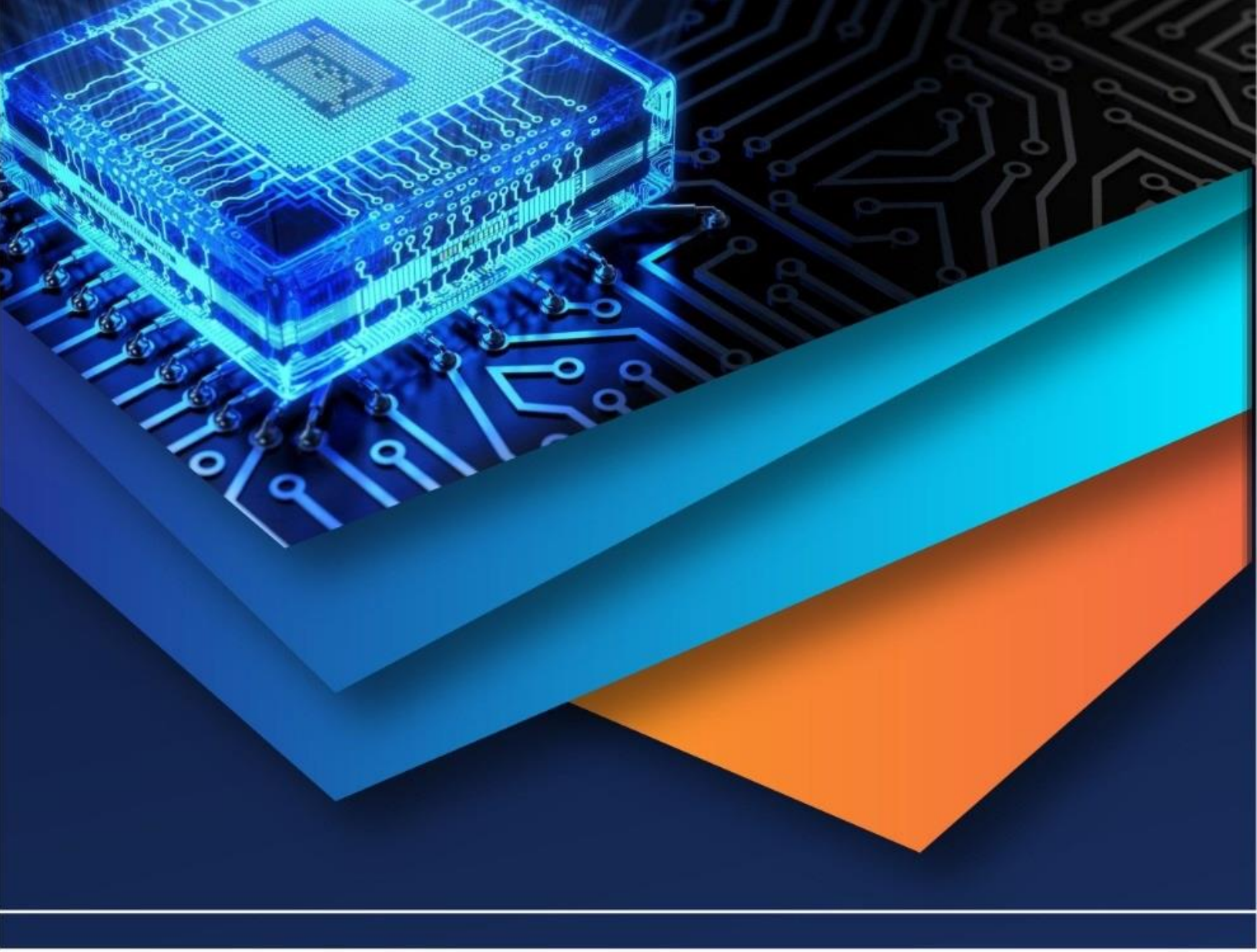
[12] Shashikant U. Ghumbre and Ashok A. Ghato, "Heart Disease Diagnosis Using Machine Learning Algorithm", Proceedings of the InConINDIA, Springer-Verlag Berlin Heidelberg 2012, AISC 132, pp. 217–225.

[13] Soni, J., Ansari, U., Sharma, D. and Soni, S. (2011) "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications, 17, pp. 43-48.

[14] Dangare, C.S. and Apte, S.S. (2012), "Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques", International Journal of Computer Applications, 47, pp. 44-48.

[15] Masethe, H. and Masethe, M. (2014), "Prediction of Heart Disease Using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science, San Francisco, pp. 809-812.

[16] Singh, M., Martins, L.M., Joanis, P. and Mago, V.K. (2016), "Building a Cardiovascular Disease Predictive Model Using Structural Equation Model and Fuzzy Cognitive Map", IEEE International Conference on Fuzzy Systems (FUZZ), Vancouver, 24-29 July 2016, pp. 1377-1382.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY