Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 4, 10454–10469

Prediction of Heart Disease using Hybrid Feature Selection

Veena S T, Jeevetha R, Abirami N

Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi

Abstract

Heart disease is serious disorder which threatens many people's lives and illness in the world's population. Predicting heart disease helps physicians and doctors to make effective decisions with respect to the health of the patients. Hence the development of machine learning (ML) leads the major part in predicting presence or absence of various serious health disorders. This study seeks to predict heart disease by using various ML models that employ hybrid feature selection. The hybrid selection involves selecting the predictive featuresby applying the fusion of filter-based feature selection and wrapper-based feature. Grid Search approach is then utilised to tune hyperparameters of classification algorithms. Finally, the comprehensive investigation of five ML classifiers such as Decision tree (DT), Logistic Regression (LR), SVM, Random Forest (RF) and Ada boost algorithms are accompanied by using metrics such as confusion matrix, accuracy score, precision, recall, kappa score, F1-score and Receiver operating curve (ROC). In the Kaggle heart disease dataset, this study discovered that the RF technique obtains an accuracy of 91% and recall of 95% compared with other classifiers with reduced feature set.

Keywords – Heart disease prediction,feature selection, classification, hyperparameters, machine learning, health.

I. INTRODUCTION

Heart disease is known to be serious disorder of human health. In accordance with the World Health Organization (WHO) released in 2022, 32% of people dying from serious disorder namely Cardiovascular Disease (CVD). CVD is known to be some familiar circumstances that infect the heart and the blood vessels in human body.

There major risk factors in CVD are high blood pressure, intake of alcohol, high cholesterol, diabetes, fatty deposits in body, genetic nature, age, etc. It is necessary to predict heart disease,therefore the medical physicians can start treatment.

When it comes to recognising heart disease risk in a short amount of time, feature selection is crucial for enhancing classification performance and heart disease prediction in large datasets. The major goal here is to identify the subset of attributes that both identifies the behaviour and generates efficient solutions. In this paper, feature selection is performed by MI- RFE for the extracted statistical and higher order statistical features. Then a model is built using the Random Forest Classifier by optimizing its hyper parameters using the GridSearchCV. Finally, the model is trained and thus a heart disease prediction is made.

The overview of this paper is obtained as, Section II is followed by literature work, our proposed works for our heart disease prediction is undertaken in Section III and in Section IVdiscusses comparison and analysis of experimental results. Finally, Section V summarizes the study's conclusions.

II. RELATED WORKS

For a long time, heart disease has been the focus of research. Over the years, researchers have experimented with a variety of technologies for doing heart disease prediction analysis, including neural networks and machine learning classifiers.

Jayshril S. Sonawane et al. [1] suggested a prediction system for major health disorder such as heart disease. This is controlled by applying a multilayer perceptron neural network. Cleveland heart disease dataset is used for this system. It consists of 303 records having 13 attributes such as age, sex, chest

pain type etc. Multilayer Perceptron is a neural network which are used to valuate any continuous function and can solve problems and classify the functions which are not linearly separable. The neurons present in this network are used with the back propagation learning algorithm. This algorithm is used to predict the heart disease in the patient with the usage of 13 attributes as input. The accuracy attained by using this neural network is 97.5%. The demerit of this study includes that the system consumes more time-to-time process, since it uses boolean features for training.

Ketut Agung Enrico et.al. [2] proposed a predictive model for leading disorder such as heart disease. K-Nearest Neighbor classifier is the supervised machine learning classifiers is used to determining the distances by selecting the average label in the classification process. From UCI repository, Hungarian dataset are utilized for this system. The accuracy obtained here is 81.85%. The limitation for this study is by applying KNN classifiers, the boundary values are increased then the functionality of the system get decreases. This system is high expensive when compared with another system.

AH Chen et al. [3] suggested a model for major killer disorder such as heart disease. The dataset used is the ML UCI repository. Applying Artificial neural networks (ANN) the model get computationally construct the result. The system was developed by using C and C# programming languages. Artificial Neural Network is the efficient model which contains various processing units that receive inputs and deliver the outputs. The proposed method achieves an accuracy of 80%. This paper does not talk about the amount of time required for analysis since it uses black box technique, and amount of data required is computationally expensive.

Sibo Prasad Patro [4] suggested ML classifiers such as K-Nearest Neighbors (KNN), Salp Swarm Optimized Neural Network (SSA-NN), Naïve Bayes, (NB), Bayesian Optimized Support Vector Machine (BO-SVM), for heart disease prediction. BO-SVM is a bayesian network which provides a probabilistic evaluation for SVM and allows direct uncertainty quantification. In Bayesian Optimization, the hyperparameters are tuned in which the values are obtained in the dataset. SSA-NN is the novel nature-inspired optimization algorithm to reduce the weight attributes for this system using dataset. The performance obtained by BO-SVM of accuracy = 93.3% and SSA-NN having performance of accuracy with 86.7%. It reveals the novel optimization algorithm which provides an effective healthcare monitoring system.

KarenGárate-Escamila et al. [5] developed a model for serious health disease namely heart disease. This is overcome by reducing the hybrid values present in the dataset blending two analysis such as Chi-square and principal component analysis (CHI-PCA). The study is performed by using three different datasets. The performance of the model was compared with five ML classifiers such as random forests, gradient-boosted tree, decision tree, multilayer perceptron, and logistic regression. CHI-PCA using random forests classifer achieves the accuracy of 99.4%.

Jae Kwon Kim et al. [6] implemented a neural network for harmful disorder in coronary arteries. It is performed by using feature correlation analysis. Feature Correlation analysis is used to determine statistical evaluation and determine the strength of a relationship between two, numerically measured, continuous functions. This research used the dataset KNHANE S-VI (The Korea National Health and Nutrition Examination Survey). The proposed method achieves an accuracy of 81.163%. The demerit of this paper is that the computational model analysis with high function values. The dependency values obtained in the model failed to explain the cause-and-effect realationship.

Manpreet Singh et.al [7] implemented the model for the prediction for major killer disease such as a CVD prediction. This is done by using structural equation modeling (SEM) and Fuzzy cognitive map (FCM). Both SEM and FCM shows 74% of accuracy. The structural relationships between measured variables and latent variables are done by using SEM. FCM shows the formulation of learning process and captures the theoretical knowledge. The dataset used in this paper is Canadian community health survey (CCHS). This paper exhibits the method of categorical

data analysis applying SEM and FCM to identify the cause of Cardio- vascular Diseases. The limitation of the paper shows the by applying large data, the system loses its functionality nature, and the accuracy obtained is comparatively less with other systems.

Kathleen j. Miao et al [8] proposed a better model for treating the coronary heart disease. Deep Neural Network (DNN) is the unsupervised learning approach which uses one or more layers of neural networks to perform in processing data and information. The proposed DNN model includes classification and prediction models built on a deep multilayer perceptron with linear and nonlinear transfer variables. Cleveland heart disease dataset has been utilized in this research. The accuracy of this system is about 83.67%.The limitation of this research includes that it is hard to include performance based on understanding, which is included in the use of classifiers.

The next section deals with the proposed approach for the heart disease prediction system which enhance overall time consumption with reduced features.

III. PROPOSED WORK

The development of complicated learning-based models for automated early diagnosis of heart issues has been facilitated by the availability of massive volumes of data for medical diagnostics. By using machine learning algorithm, this model generalizes to new data sets which are not observed in the training set. As a result, the trained model has a lower prediction accuracy. This model uses feature selection to select the predictive features using fusion of filter and wrapper method. The study thus aims to find the classifier by tuning the hyper parameters that predicts heart disease efficiently and perform classification between healthy and unhealthy people with balanced dataset.

Fig 1 depicts the architecture for heart disease prediction system. As an initial step, from the dataset of heart disease, insights of the features like higher order statistical and statistical features are obtained. Following that the training and testing dataset is obtained by splitting the dataset into 2 sets. Applying the training dataset,selection of predictive featuresof the dataset is performed. RFC model is built while optimizing its hyperparameters. Finally, a model to predict heart disease is trained.

A. Dataset Description:

In this paper, the dataset for heart disease was handled to create our required model. Kaggle was used to get the data.This dataset has 14 characteristics. Table 1 shows the detailed description of all characteristics. The collection contains 1025 patient records, with 713 males and 312 females of varied ages. In those 499 people are healthy, whereas 526 have cardiac disease. Among the 526 persons with heart illness, there are 300 men and 226 women. From the dataset we can see that 48.68% patients are normal and 51.32% have heart disease. Table 1 describes the dataset used in this model.

Table 1:
Dataset explanation.

S.no	Attribute Name	Description
1	sex	Female = 0; Male = 1
2	chol	serum cholesterol and its unit are mg/dl
3	restecg	resting electrocardiographic results
4	Cp	Types of chest pain (has 4 types)
5	age	Age in years
6	exang	exercise induced angina
7	trestbps	resting blood pressure (it is denoted as mm Hg while admitting in hospital)
8	Oldpeak	ST depression
9	thalach	High heart rate obtained
10	slope	peak exercise ST segment’s slope
11	fbs	fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
12	ca	number of major vessels (0–3) colored by fluoroscopy

13	Target (class)	0 = no disease and 1 = disease
14	thal	Normal is denoted as '1'; fixed defect is denoted as '2'; reversible defect is denoted as '3'.

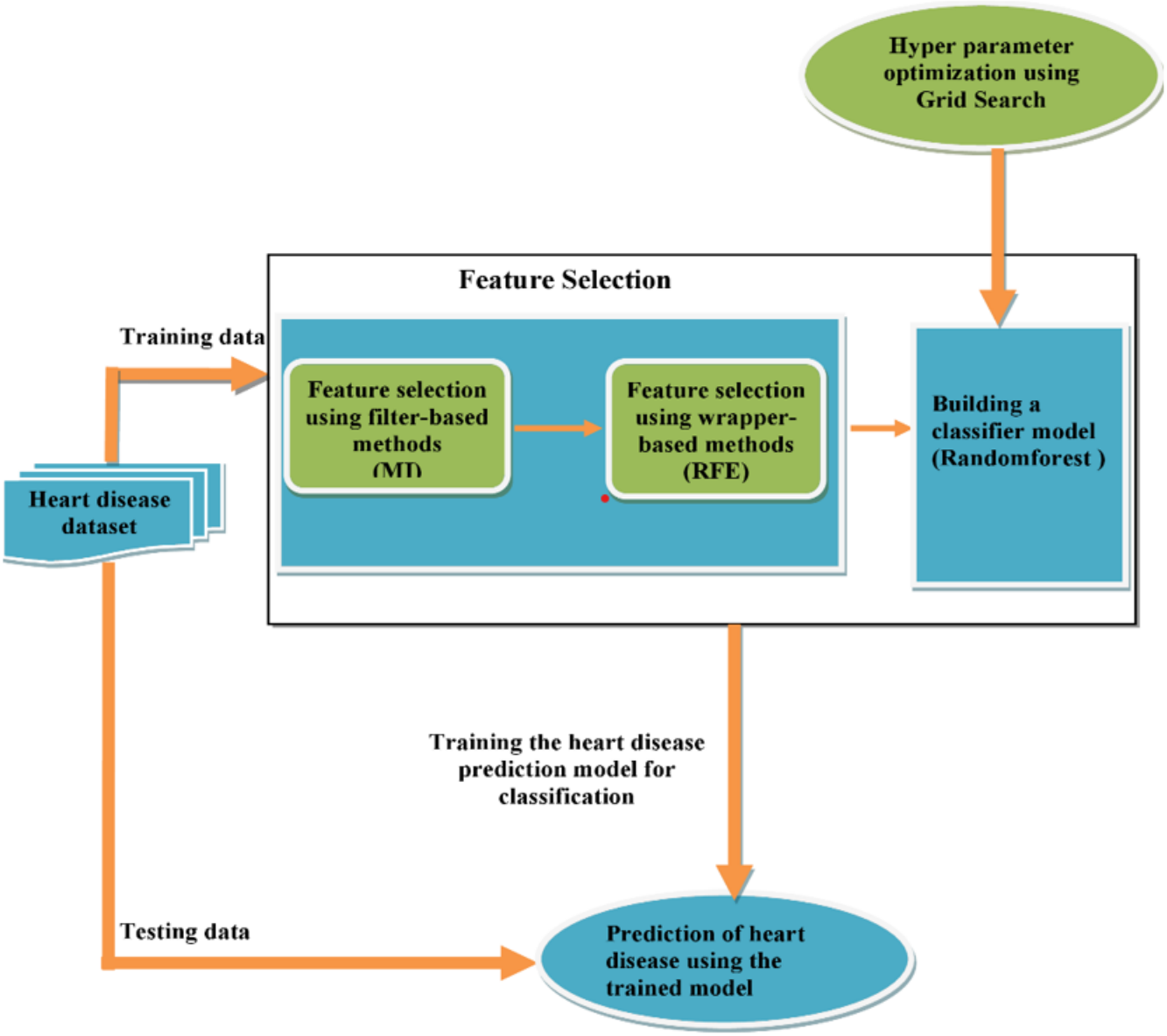

Fig 1: System Design

B. Feature Selection:

The selection of features is also called as attribute choosing for model construction which denotes the process of choosing a portion of predictive features. The feature set may be vast because to the large amount of data available and processing them in short time with more effectiveness is a key task. As a result, a specific technique must be used to reduce the available feature set.

In this proposed work, firstly, the dataset's higher-order statistical and statistical features are extracted. The utilized dataset includes 13 features, where statistical metrics like minimum, mean, maximum,standard deviation and median are calculated. Likewise, we calculate the entropy of the available 13 features by initially obtaining the probability of the 13 features. Further, higher-order statistical features the probability of features occurrence is used to determine features. Hence from the statistical measures we obtained 5 features, from the entropy calculation we obtained 13 features and finally

from the higher-order statistical features we obtained 5 features. Therefore, The selection of features technique yielded around 36 features. Fig 2 shows the procedure forfilter and wrapper-based (MI-RFE) fusion feature selection method.

The model then obtains the predictive features by using fusion of filter-based (MI) and wrapper-based (RFE) feature selection methods.

B.1Filter based feature selection:

This method finds the irrelevant attributes by ranking the features based on a univariate metric. After that the features are ranked according to score obtained in a decreasing order. Now from the obtained order the highest k-ranking features are chosen as the predictive feature. The different metrics available in filter methods are Chi-Square, Pearson's Correlation, Mutual Information (MI), Anova etc. Here the univariate metric used is MI which gives better result than other metrics.

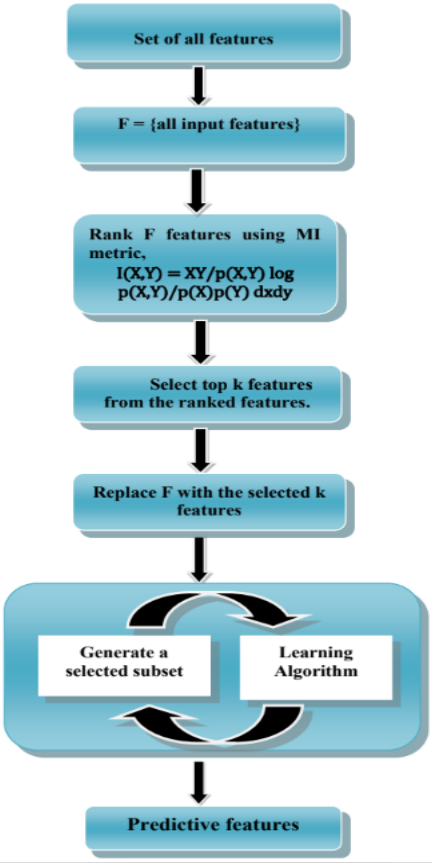

Fig 2: Fusion of filter and wrapper method (MI-RFE)

B.1.1Mutual Information:

Mutual information is a measure that estimates the amount of knowledge about one random variable received via the use of another random variable, such as X and Y. It is given by

$$I(M,N) = \frac{MN}{p(M,N)} \log \frac{p(M,N)}{p(M)p(N)}\, dm\, dn$$

where

the joint probability density function of *M* is *p(M,N)*, while the marginal density functions are *p(M)* and *p(N)*.

The MI finds how similar the joint distributions are and if *M* and *N* are totally not related then the integral value becomes zero. Also, the MI value must be higher between the features selected as subset *M*s and the target variable *n*, it is given by

S = argmax SIXs; s.t.S = k, where k is the total number of characteristics to choose from.

B.2Wrapper based feature selection:

In this it uses a collection of features and uses them to train a model. It evaluates by figuring out all possible combinations of features against the performance measure which depends on the type of problem, that is for classification type of problem the performance measure can be like accuracy, f1 score, precision, etc. The model trains until it gives a subset of features which gives the optimal result. Forward selection, backward elimination, and recursive feature elimination (RFE) are some of the wrapper-based feature selection approaches. Here the recursive feature elimination feature selection is used after performing the filter-based feature selection method. RFE is performed after performing MI where the results of MI are sent to RFE to obtain the predictive features.

B.2.1Recursive Feature Elimination:

It is essentially a selection of predictors by adapting the backward selection where this method starts by creating a model based on whole list of predictors and assigning an importance value to each one. The design is then rebuilt, and significance scores are recalculated after the least important predictors are eliminated. The predictors are chosen on the basis of importance rankings using the

collection size that optimizes the performance requirements. The final model is then trained using the best subset.

If the number of predictors exceeds the number of samples, some models cannot be employed due to RFE as it mandates that the initial model use the entire predictor set. Logistic regression, linear discriminant and multiple linear regressionanalysis are examples of these models.If a model wishes to use one of these RFE techniques, it must first filter down the predictors. Furthermore, the adoption of RFE benefits some models more than others. One of these models is random forest, which is employed with RFE because it has a better internal approach for determining feature importance. The algorithm of fusion of MI- RFE method is given below.

Input: Training dataset
Output: The predictive features of the dataset
Begin
F = {all input features}
Rank F features using MI metric,

$$I(M,N) = \frac{MN}{p(M,N)} \log \frac{p(M,N)}{p(M)p(N)} da\, db$$

From the ranking features, choose the top k features.
Replace F with the selected k features
While (F is not empty) do
Train F by a given model
Compute the vector weight
Rank the features in F by
Find the bottom ranked feature
F = F – {bottom ranked feature}
Return the ranked feature dataset
End

Thus, from the extracted higher order statistical and statistical features 25 features are obtained using the filter-based (MI) feature selection method. Those features are then passed to the wrapper based (RFE) feature selection method which gives 12 features.

C. Building a Classifier Model

C.1 Hyper parameter tuning:

Many machine-learning algorithms rely heavily on the hyper-parameters that are employed, especially for complicated models. Experiments are used to set hyper-parameters including kernel size, learning rate, number of trees and number of estimators among others.

Prior to the learning phase, which establishes the model's optimum parameters, they are defined and estimated. Many hyper-parameters must be tweaked in machine-learning models. With adequate management of hyper-parameters, overfitting of the model may be prevented, which occurs when the model utilised is too flexible.

The tweaking of hyper parameters is the focus of this research, and it is applied in multiple classifiers such as Adaboost, Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM) and Logistic Regression (LR). Selecting the best collection of hyper-parameters is critical for optimising HDP task performance.The hyper-parameter values are chosen to get the best-maximized F-score. The hyper parameters are obtained using the GridSearchCV approach in this modelbecause it aids in looping over predetermined hyperparameters and fitting the model training set As a result, the best hyperparameters from the list are chosen in the end. Here hyper parameter optimization of different classifiers is studied.To begin the prediction process, we utilise the HPO approach to find the optimal hyperparameter values for our RFC-based classifier on the testing dataset.

C.1.1. Random Forest:

The Random Forest Technique is a supervised learning approach that is used in regression and classification to handle data sets with both continuous and categorical variables. It incorporates bagging to increase the Decision Tree's performance. Instead of dividing nodes based on variables, it combines tree predictors and separates nodes based on the best predictor subset selected at random from the node itself. The worst case of learning with Random Forests has a temporal complexity of O(M(dnlogn)), where the number of developing trees is M, the number of instances is n, and the data dimension is d.

Algorithm Steps:

- Obtain a random sample of properties from a dataset.
- For each expected outcome, cast a vote.

- As the final forecast, pick the prediction that has received the most votes.

Hyperparameters taken for tuning RFC is given in Table 2.

Table 2:
Hyperparameters of RFC

Hyper parameters	Values
max_features	[2, 3, 4]
max_depth	[8, 10, 12]
min_samples_split	[8, 10, 12]
min_samples_leaf	[2, 4]
n_estimators	[100, 200, 300]
bootstrap	[True, False]
max_leaf_nodes	[6, 8, 10]

C.1.2. Decision Tree:

A Decision Tree's internal nodes represent dataset properties, the branches represent decision rules, and each leaf node represents the outcome. Two nodesavailable in a Decision tree in which decision nodes make decisions and have several branches, whereas Leaf nodes represent the decisions' outputs. The results of the features in the given dataset are used to make the decisions.

Algorithm Steps:

- Begin with S as it is the root nodewhich also has the entire dataset.
- Find the best attribute in the dataset using the Attribute Selection Measure (ASM) method.
- Split the root node S into subgroups with the best attribute results.
- Make a Decision Tree node that contains the better attribute from the dataset.
- Generate new decision trees recursively adopting subsets of the dataset you've constructed and keep doing so until all the nodes have been classified as final nodes. a node in the form of a leaf.

The hyperparameters taken for tuning DT classifier is given in Table 3.

Table 3:
Hyperparameters od DT classifier.

Hyper parameters	Values
min_samples_split	range (1,10)
max_depth	[2, 3, 5, 10, 20]
min_samples_leaf	[5, 20, 10, 50, 100]
max_features	[2, 3, 4]
criterion	["gini", "entropy"]

C.1.*3. Logistic Regression:*

The supervised machine learning method logistic regression predicts a categorical dependent variable's output. As a result, the outcome must be either discrete or categorical.This algorithm's result is a set of probabilistic values ranging from 0 to 1. It is used to categorise values using various sorts of data and can quickly determine which variables are the most successful for classification.

Algorithm Steps:

- S suggests beginning with the root node, which holds all the data.
- Identify the best attribute in the dataset with the Attribute Selection Measure (ASM).
- Split the root node S into subgroups with the best attribute results.
- Generate a Decision Tree node that includes the dataset's better property.
- Recursively adopt parts of the dataset you've created to create new decision trees and keep doing so until all nodes have been classed as final nodes. a node shaped like a leaf. The hyperparameters taken for tuning LR is given in Table 4.

Table 4:
Hyperparameters of LR.

Hyper parameters	Values
solver	['liblinear']
C	np.logspace(-4, 4, 20)
penalty	['l1', 'l2']

C.1.4. Support Vector Machine:

A supervised learning tool for categorising, forecasting, and detecting outliers is the support vector machine.A basic linear SVM classifier networks with two classes by connecting them with a straight line.That is, data points on one side of the line will be attributed to one category, while data points on the other will be assigned to a different category.

Algorithm:

- Assemble the important libraries.
- Import dataset and extract the X variables and Y variables separately.

- Chop the dataset into test and train.
- Initializing and fitting the SVM classifier model
- Coming up with predictions
- Evaluating the model's performance

The hyperparameters taken for tuning SVM is given in Table 5.

Table 5:
Hyperparameters of SVM.

Hyper parameters	Values
gamma	[1, 0.1, 0.01, 0.001, 0.0001]
C	[0.1, 1, 10, 100, 1000]
verbose	[True, False]

C.1.5. Ada Boost:

AdaBoost is a binary classification boosting technique that has proven to be successful. It's used to improve the performance of decision trees and predict categorization values. Weak models are initially added one by one and trained using weighted training data. This practise is repeated until just a few weak learners remain.

Algorithm Steps:

- Primarily, at random the algorithm selects a training subset.
- The model is trained by selecting a training set depending on the accuracy of the previous training's prediction.
- In each repetition, the trained classifier is assigned with the weight based on the result of the classifier's accuracy. The better accurate result will be given more weight.
- Continue the process in expectation where all the training data gets perfectly fitted or in expectation of reaching the higher number of estimators.

The hyperparameters taken for tuning Adaboost is given in Table 6.

Table 6:
Hyperparameters of Adaboost.

Hyper parameters	Values
learning_rate	[0.0001, 0.001, 0.01, 0.1]
n_estimators	[10, 100, 50, 500]

IV. EXPERIMENTAL RESULTS

D.1Performance evaluation metrics:

Performance metrics applied in this model are precision,F1 score, Kappa score, recall, accuracy and finally ROC and AUC score. Predictions for classification problems consists of four types of namely false-positives**,** true-positives,false-negativesand true-negatives.

Confusion matrix

The confusion matrix serves as the foundation for all other categorization metrics. It's a matrix that thoroughly represents the model's performance. A confusion matrix breaks down the correct and wrong classifications of each class in detail. Positive prediction that turns out to be true positives are known as true positives. The classifier labels the negative tuples correctly are referred to as true negatives. Predictions that appear to be positive but are negative are known as false positives. Positive tuples that were mistakenly labelled by the classifier are referred to as false negatives.

The measures used are defined as follows.

Accuracy:

It is given by the ratio of number of absolute predictions to total sample count.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Recall:

The fraction of correctly obtained positive observations to all estimation in the actual class is referred to as recall.

$$Recall = \frac{TP}{FN + TP}$$

Precision:

The division of exactly predicted positive information to total expected positive observations is known as precision.

$$Precision = \frac{TP}{FP + TP}$$

F1 score:

It is given by the harmonic mean bean precision and recall.

$$F1 - Score = \frac{2(Precison * Recall)}{Precision + Recall}$$

ROC curve:

The Receiver Operating Characteristic (ROC) for various threshold outcomes, the curve is displayed against TPR and FPR. When TPR rises, FPR rises with it. When comparing

different predictors on a particular dataset, the area under the curve (ROC AUC) has a higher numerical value than the others.

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$False\ Postive\ Rate = \frac{FP}{TN + FP}$$

Result analysis:

The model utilizes the heart disease dataset take from the Kaggle. It has 14 attributes namely slope, Oldpeak,restecg, Cp, target, age, sex, trestbps, fbs, thalach, exang,nca, thal andchol. The statistics and higher-level statistical aspects of the dataset are extracted first. The Kaggle dataset for heart disease is then chopped into testing and training data in the ratio 80:20, with 80% of the data being used for training and 20% for testing. Multiple fusion of filter and wrapper-based feature selection approach is applied like MI-RFE, MI-FS, CS-RFE and CS-FS using RFC for training set. Table 7 shows the accuracy, f1-score, recall and precision of different feature selection methods mentioned. MI-RFE filter and wrapper method gives higher accuracy of 90.73%. Now MI-RFE method is used for evaluating different classifiers to determine the better classifier.

Table 7:
Recall, precision, f1-score and accuracy of feature selection methods using RFC

Feature Selection	Recall	Precision	F1-score	Accuracy
MI-RFE	95.33	87.93	91.48	90.73
MI-FS	92.52	88.39	90.41	89.76
CS-RFE	97.2	85.25	90.83	89.76
CS-FS	93.28	86.81	89.93	89.11
No feature selection	93.46	88.5	90.91	90.24

The model is trained using the features obtained after applying MI-RFE to the training dataset. In this model different classifiers are taken like RF, DT, LR, SVM and Adaboost and their hyper parameters are tuned and then the model is trained. Table 8 gives the hyper parameters obtained for RFC. Table 9 gives the hyper parameters obtained for DT. Table 10 gives the hyper parameters obtained for LR. Table 11 gives the hyper parameters obtained for SVM. Table 12 gives the hyper parameters obtained for Adaboost. The hyper parameters obtained are utilized for calculating the performance values and Table 13 gives the precision, f1-score, accuracyand recall of different classification algorithms.

Table 8:
Hyperparameters of RFC

Hyper parameters	Values
max_features	4
min_samples_split	8
n_estimators	100
min_samples_leaf	2
bootstrap	False
max_depth	8
max_leaf_nodes	10

Table 9:
Hyperparameters od DT classifier.

Hyper parameters	Values
min_samples_leaf	5
max_features	4
criterion	entropy
min_samples_split	2
max_depth	10

Table 10:
Hyperparameters of LR.

Hyper parameters	Values
penalty	l1
C	0.615848211066026
solver	liblinear

Table 11:
Hyperparameters of SVM.

Hyper parameters	Values
C	1000
gamma	1
verbose	True

Table 12:
Hyperparameters of Adaboost.

Hyper parameters	Values
n_estimators	500
learning_rate	0.1

Table 13:

Various classification algorithms produced different categorization results.

Classification Algorithms	Accuracy	Recall	Precision	F1-score
RF	90.73	95.33	87.93	91.48
DT	89.11	88.81	90.15	89.47
LR	88.29	92.52	86.09	89.19
SVM	89.76	93.46	87.72	90.5
Adaboost	85.37	88.79	84.07	86.36

It can be interpreted that RF classifier gives higher accuracy of 90.73% with better recall rate of 95.33% among all the other classifiers utilized. Table 13 gives the kappa score and ROC AUC score of various algorithms for classification is used.

RF outperforms LR, SVM, Adaboost, and DT, according to the results presented in the table. Fig 3displays RFC's confusion matrix. Fig 4 shows DT's confusion matrix. Fig 5 LR's confusion matrix. Fig 6 shows SVM's confusion matrix. Fig 7 shows the confusion matrix of Adaboost

Table 13:
Kappa score and ROC AUC score of different classification algorithm.

Classification Algorithms	Kappa Score	ROC AUC score
RF	81.36	90.52
DT	78.18	89.12
LR	76.46	88.1
SVM	79.41	89.59
Adaboost	70.6	85.21

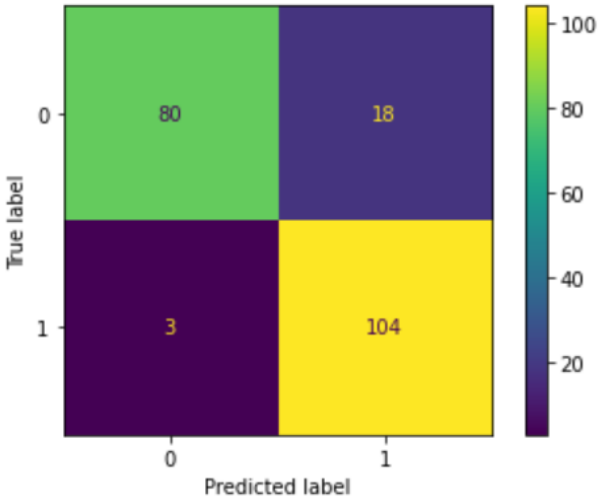

Fig 3: RFC's Confusion matrix

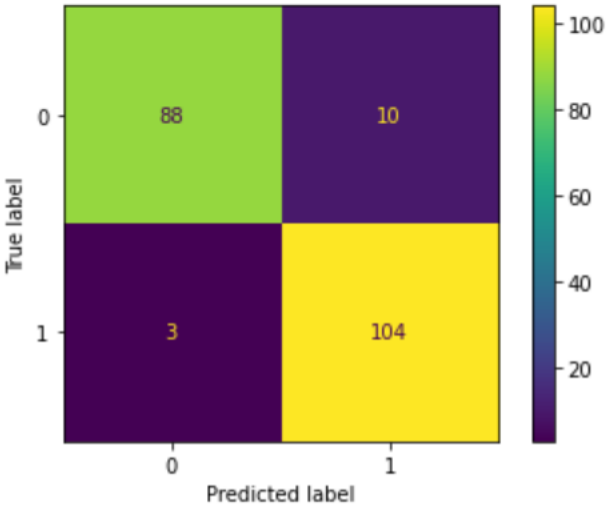

Fig 4: DT's Confusion matrix

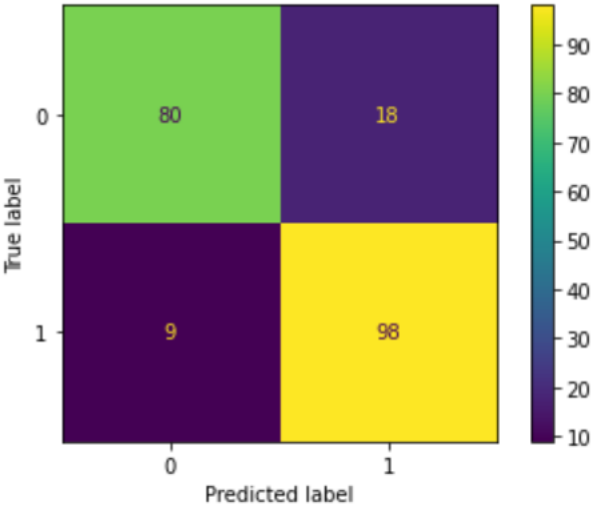

Fig 5: LR's Confusion matrix

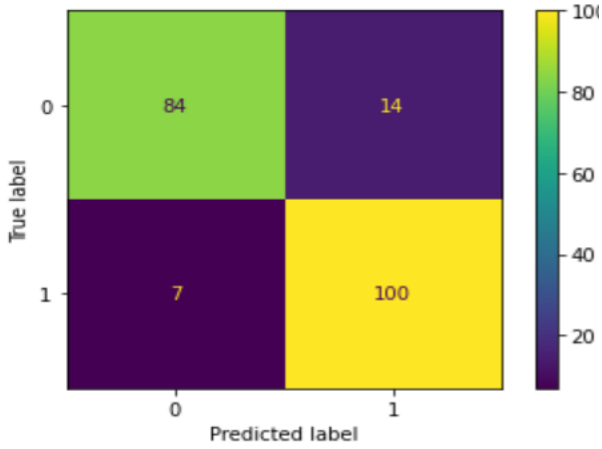

Fig 6: SV's Confusion matrix M

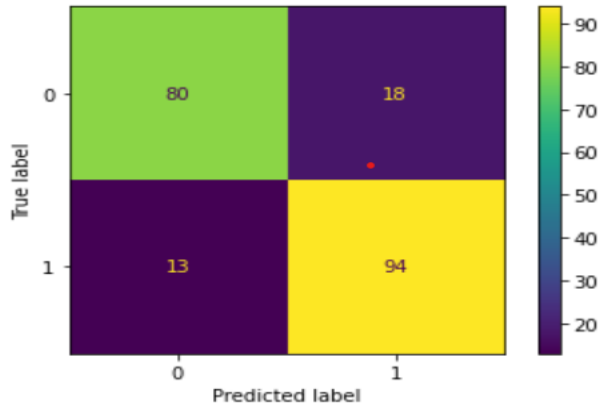

Fig 7: Adaboost's Confusion matrix

The ROC curve for RFC is shown in Figure 8, which is constructed using the true positive rate and false positive rate values. Fig. 9 represents DT's ROC curve. Fig. 10 represents LR's ROC curve. Fig. 11 represents SVM's ROC curve. Fig. 12 representsAdaboost'sROC curve.

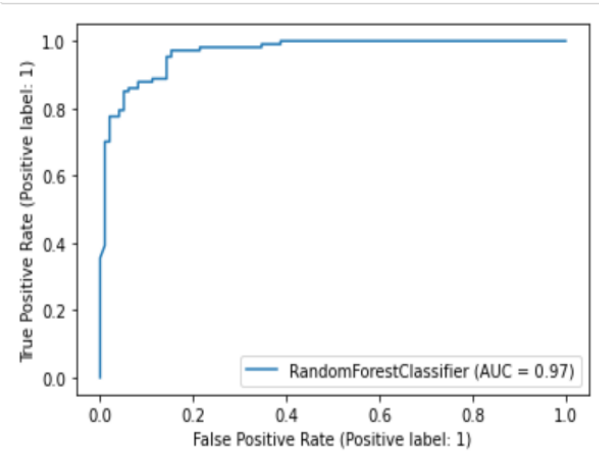

Fig 8: RFC's ROC curve.

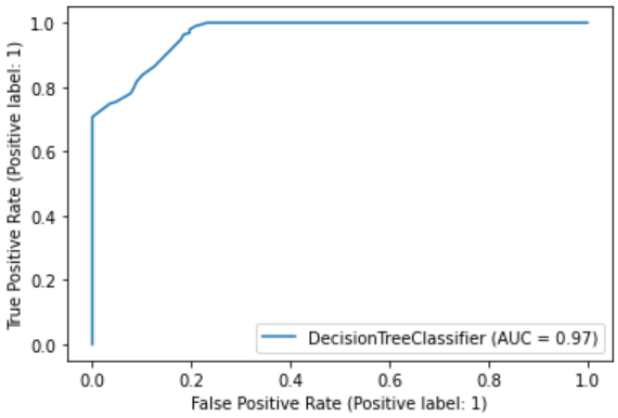

Fig 9: DT's ROC curve.

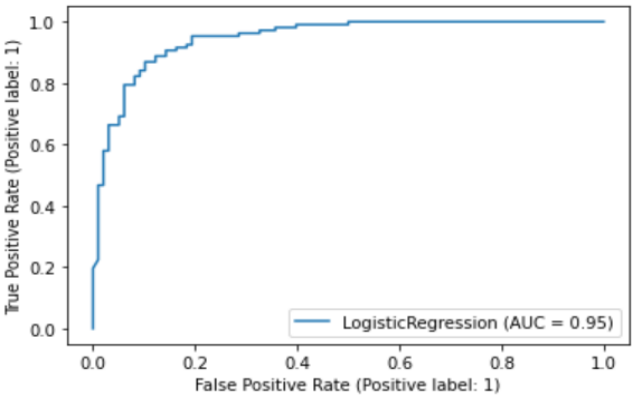

Fig 10:LR'sROC curve.

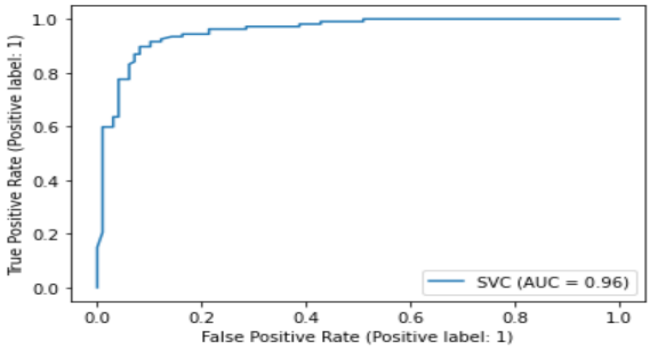

Fig 11: SVM's ROC curve.

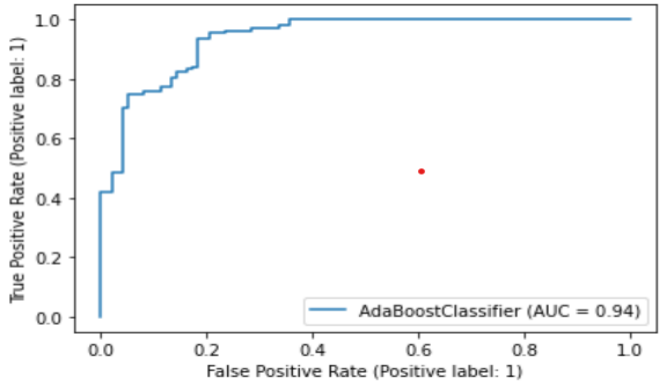

Fig 12: Adaboost'sROC curve.

It is understood that, from the ROC curves of different classifiers RFC gives good result. The area under the precision-recall curve (AUPRC) of various RF classifier is shown in Fig 13.AUPRC of DT is shown in Fig 14. AUPRC of LR is shown in Fig 15. AUPRC of SVMs is shown in Fig 16. AUPRC of Adaboost is shown in Fig 17. Fig. 18 represents the learning curve of RFC. Fig. 19 represents the learning curve of DT. Fig. 20 represents the learning curve of LR. Fig. 21 represents SVM's learning curve. Fig. 22 represents Adaboost's learning curve.

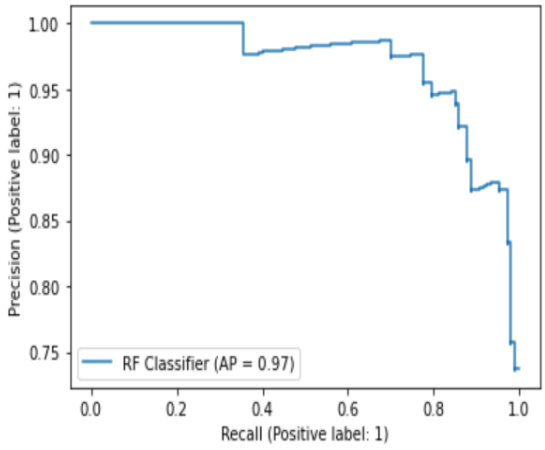

Fig 13: Precision recall curve of RFC.

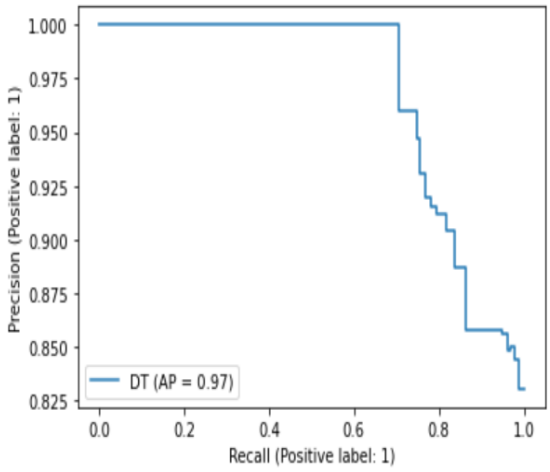

Fig 14: Precision recall curve of DT.

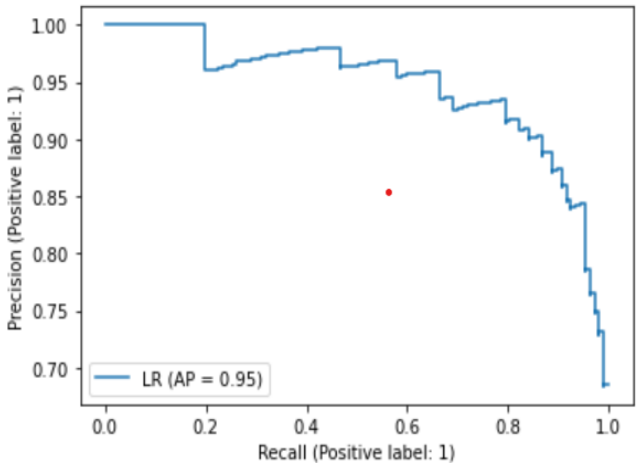

Fig 15: Precision recall curve of LR.

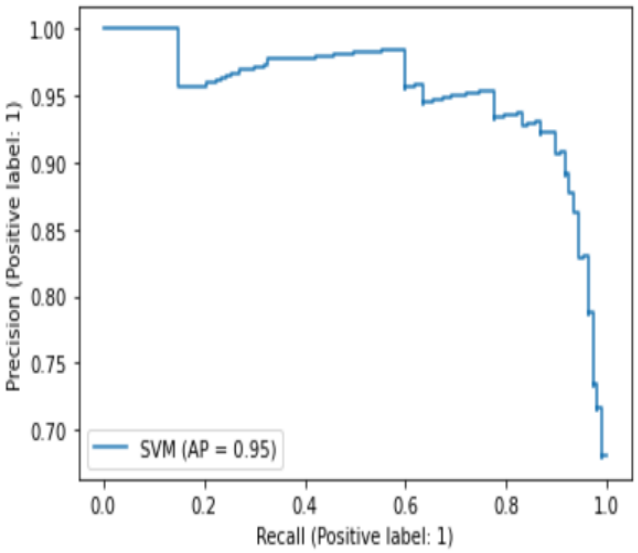

Fig 16: Precision recall curve of SVM.

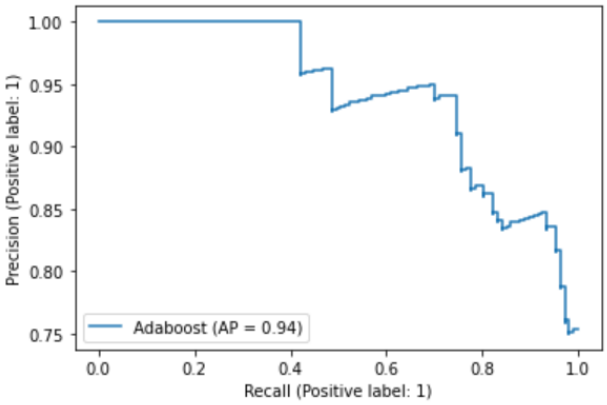

Fig 17: Precision recall curve of Adaboost.

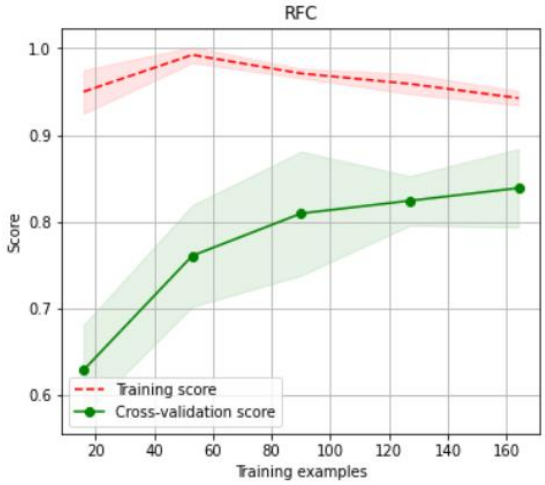

Fig 18: RFC's learning curve

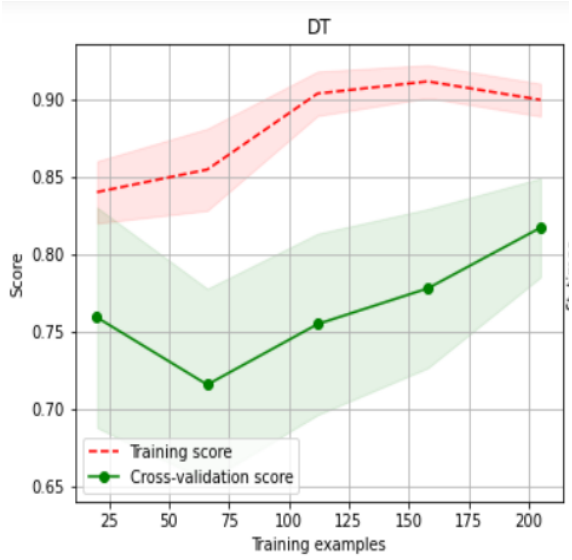

Fig 19: DT's learning curve

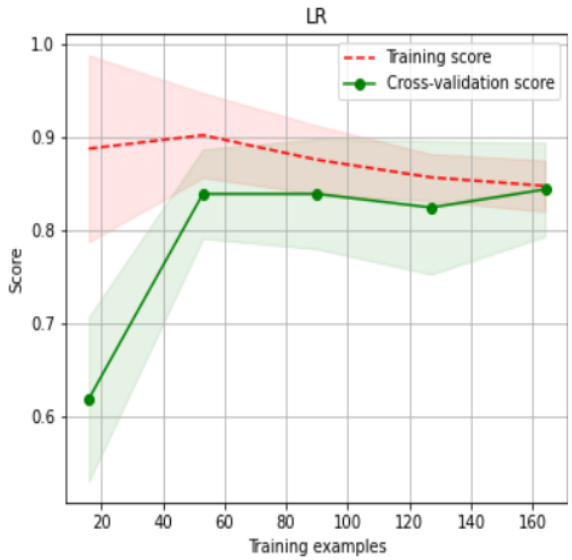

Fig 20: LR's learning curve

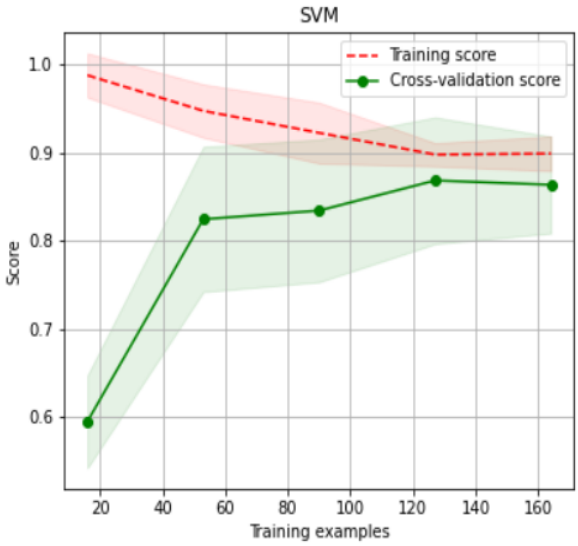

Fig 21: SVM's learning curve

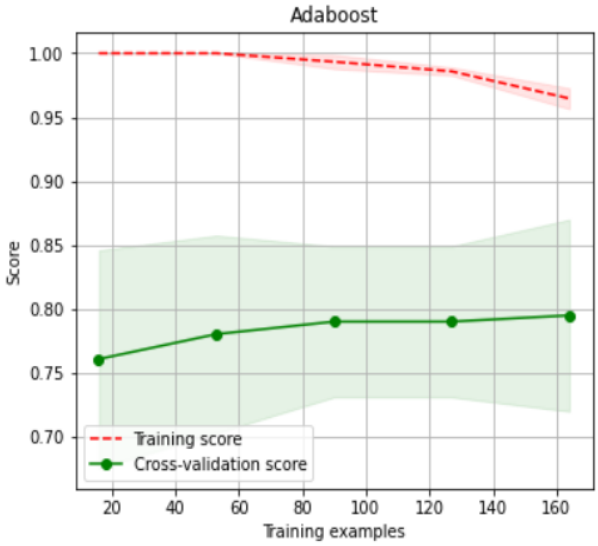

Fig 22: Adaboost'slearning curve

Also, from the precision recall curve RFC gives better results than other classifiers utilized. Hence, RFC classifier holds an important part in the procedure of predicting heart disease, since it works with simple parameters and is adaptable to many sorts of data properties and features.The RFC can easily manage huge quantities of data that is used for training by splitting the dataset into some samples before commencing the process of learning. Also, many factors influenced our decision to choose RFC as the best classifier such as, it is very quick to predict along with ensuring good performance and efficiency when dealing with large databases. The RFC method has been shown to be a very effective classifier for predicting heart disease.

The Kaggle dataset of heart disease is used to train the modelandit attains better results

during the heart disease prediction process. On proceeding this model gives good results in case of recall, f1-score, accuracy and precision in real time. This study reveals that simple machine learning technique is more than enough for predicting heart disease than going for more complex algorithms. This proposed model is compared with few current works available, and it is depicted in the table 14.

Table 14:
Comparison of RFC using MI-RFE with current work

Current work	Methodology/Result in current work	Methodology/Result in proposed work
[1]	Takes more time comparatively	Prediction time is comparatively lesser in real time.
[2]	Using only KNN gives an accuracy of 81.85%	Uses RFC (MI-RFE, HPO) gives accuracy 90.73%
[3]	Uses Artificial Neural Network (ANN) and gives accuracy of 80%	Uses simple machine learning algorithm with feature selection and performing hyper parameter optimization gives accuracy 90.73%
[6]	Accuracy obtained by using structural equation modeling (SEM) and Fuzzy cognitive map (FCM) is 74%	Accuracy achieved for using RFC (MI-RFE, HPO) is 90.73%

IV.CONCLUSION

Since heart disease is a leading cause of death in India and around the world, applying promising technology such as machine learning to the early detection of heart disease would have a greater influence on society. Prediction of heart diseases at earlier stages may prevent possible deaths due to serious risk factors. Each year, the number of people suffering from heart disease grows. This necessitates early detection and treatment. The use of appropriate technology support in this regard can be extremely advantageous to both the medical industry and the patients. A good classification algorithm may help the physician predict the presence of cardiovascular disease before its occurrence. This paper focuses on predicting a possible heart disease by incorporating a Kaggle dataset.

SVM, Decision Tree, Random Forest, Logistic Regression, and AdaBoost are five different machine learning techniques utilised to quantify performance in this work. The expected attributes leading to heart disease in patients are available in the dataset which contains 14 attributes that are useful to evaluate the system are selected among them. If all the features taken into the consideration, then the efficiency of the system get decreased. To increase efficiency, feature selection is done. In this n features must be selected for evaluating the model which gives more accuracy. The correlation of some features in the dataset is almost equal and so they are removed.

All the five machine learning methods after tuning the hyper parameters their accuracies are compared based on which, one prediction model is generated. Hence, the aim is to use various evaluation metrics like kappa score, recall, confusion matrix, precision, f1-score and accuracy which predicts the disease efficiently. Comparing all five, Random forest and Decision Tree classifier gives the highest accuracy of 90.73%, recall of 95.33%, precision of 87.93% and F1-score of 91.48%.

V. REFERENCES

1. Jayshril S. Sonawane, D.R Patil, "**Prediction Of Heart Disease Using Multilayer Perceptron Neural Network"**, IEEE Access *on Information Communication and Embedded Systems (ICICES2014)*, 2014, pp. 1-6, doi: 10.1109/ICICES.2014.7033860.
2. Ketut Agung Enriko, Muhammad Suryanegara, and DadangGunawan," **Heart Disease Diagnosis System with k-Nearest Neighbors Method Using Real Clinical Medical Records",** Proceedings of the 4th International Conference on Frontiers of Educational Technologies, Association for Computing Machinery, New York, NY, USA,

127–131.https://doi.org/10.1145/3233347.3233386

3. AH Chen, SY Huang, PS Hong, CH Cheng, and EJ Lin,2011, "**HDPS: Heart Disease Prediction System", Computing in Cardiology"**, ISSN: 0276-6574, pp.557- 560.
4. Sibo Prasad Patro, Gouri Sankar Nayak, Neelamadhab Padhy,**"Heart disease prediction by using novel optimization algorithm: A supervised learning prospective"** Informatics in Medicine Unlocked Volume 26, 2021, 100696 https://doi.org/10.1016/j.imu.2021.100696
5. Mohammad Shafenoor Amin, Yin Kia Chiam, Kasturi DewiVarathan, "**Identification of significant features and data mining techniques in predicting heart disease"**, Telematics Inform. 36 (2019) 82–93.
6. Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, Emmanuel Andrès, "**Classification models for heart disease prediction using feature selection and PCA",** Informatics in Medicine Unlocked,Volume 19,2020,100330,ISSN 2352-9148,https://doi.org/10.1016/j.imu.2020.100330.
7. Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, and Vijay K. Mago,2016,**"Building a Cardiovascular Disease predictive model using Structural Equation Model & Fuzzy Cognitive Map,"** 2016 IEEE Access on Fuzzy Systems (FUZZ-IEEE), 2016, pp. 1377-1382, doi: 10.1109/FUZZ-IEEE.2016.7737850.
8. Kathleen H Miao and Julia H. Miao, **"Coronary Heart Disease Diagnosis using Deep Neural Networks"** International Journal of Advanced Computer Science and Applications(IJACSA), 9(10), 2018. http://dx.doi.org/10.14569/IJACSA.2018.091001
9. Dr.K.Thirumoorthy, Dr.K.Muneeswaran published "**Feature selection using hybrid poor and rich optimization algorithm for text classification**" Pattern Recognition Letters Vol.147, Issue.1, pp.63-70, July-2021 , DOI: https://doi.org/10.1016/j.patrec.2021.03.034
10. Dr.K.Thirumoorthy, Dr.K.Muneeswaran published "**Feature Selection for Text Classification Using Machine Learning Approaches**" National Academy Science Letters Vol.1, Issue.1, pp.1-7, March-2021 , DOI: https://doi.org/10.1007/s40009-021-01043-0
11. Mr.B.Lakshmanan, Ms.T.Jenitha published "**Optimized Feature Selection and Classification in Microarray Gene Expression Cancer Data**" Indian Journal of Public Health Research & Development Vol.11, Issue.1, pp.347-352, January-2020 , DOI: 10.37506/v11/i1/2020/ijphrd/193842
12. Miss.Pavithra D, Mr.B.Lakshmanan published "**Feature Selection and Classification in Gene Expression Cancer data**" IEEE sponsored International Conference on Computational Intelligence in Data Science , June-2017
13. Dr.M.Jansi Rani, Dr.M.Karuppasamy, Mrs.M.Prabha published "**Bacterial foraging optimization algorithm based feature selection for microarray data classification**" Materials Today: Proceedings , December-2020 , DOI: 10.1016/j.matpr.2020.11.325
14. Safial Islam Ayon, Md. Milon Islam & Md. Rahat Hossain (2020): "**CoronaryArtery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques"**,IETE Journal of Research, DOI: 10.1080/03772063.2020.1713916https

://doi.org/10.1080/03772063.2020.1713916

15. V. R. Elgin Christo, H. Khanna Nehemiah, J. Brighty&Arputharaj Kannan(2020): "**Feature Selection and Instance Selection from Clinical Datasets Using Co-operativeCo-evolution and Classification Using Random Forest**", IETE Journal of Research, DOI:10.1080/03772063.2020.1713917. https://doi.org/10.1080/03772063.2020.1713917