HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

P.Chitra[1]*, T. Sathis Kumar[2], M. Mary Shanthi Rani[3]

Abstract

The heart is the next big organ with more importance in the human body relative to the brain, which pumps the blood and supplies it to all organs of the whole body. Prediction of occurrences of heart diseases in the medical field is significant work. Data analytics is helpful for prediction from more information, and it helps the medical centre predict various diseases.

In this paper, different techniques of mining for forecasting heart risk discussed. Heart disease cause millions of death every year, it is rapidly increasing mining methods one too much helpful detect and diagnose heart risk. Different mining methods have used to abstract knowledge for forecasting heart disease. In this paper, the survey is carried on various single data mining techniques to achieve high accuracy in predicting heart disease. Here we use many classifications, namely Random Forest Classifier (RFC), K-Nearest Neighbor Classifier (KNN), Gradient Boosting Classifier (GBC), Extra Tree Classifier (ETC), Extreme Gradient Boosting Classifier (XGB), the approach of classifiers. Analysis of various methods proved that techniques based on classification obtain high accuracy compared to previous methods. The performance of the classifier model is confirmed to outperform its counterparts progressively. The improved accuracy of various classifiers experimented in this reported research work vouches for its application in Heart Disease classification (HD).

Keywords: Cardiac Disease, RFC, XGB, KNN, ETC, GBC Classifier, Clinical Data, Confusion Matrix, ROC curve, AUC, Normal and Abnormal HD.

[1*,2,3]Department of Computer Science and Engineering, School of Engineering and Technology Dhanalakshmi Srinivasan University, Samayapuram, Trichy

***Correspondence Author:** P. Chitra
*Department of Computer Science and Engineering, School of Engineering and Technology Dhanalakshmi Srinivasan University, Samayapuram, Trichy

DOI: - 10.48047/ecb/2023.12.si5a.0232

I. Introduction

Cardiovascular disease is one of the common diseases that can reduce human life nowadays. 17.5 million people die each year as a result of heart disease. Life depends on the heart's component because the heart is a necessary part of our body. Heart disease is a disease that affects the function of the heart [1]. For many health promotions and clinical medicine aspects, estimating a person's risk of coronary heart disease is essential. Due to the rapid growth of digital technologies, healthcare centres store a very complex and challenging amount of data in their database. In analysing different data in medical centres, data mining techniques and machine learning algorithms play vital roles.

The SPECT data from heart disease is used for evaluation and has downloaded from the UCI machine's learning repository [2]. This paper's primary objective is to classify the different cardiac SPECT diagnosis stages using the most favorable feature set. This feature vector then predicted using different algorithms for the classification. The performance has evaluated using metrics similar to Accuracy, Precision, Recall and FI-score, Area Under Curve (AUC), Receiver Operating Characteristic (ROC) Curve. The remaining sections of this article are section II presents the review on SPECT heart disease classification methods, the methodology has described in section III, and the respective descriptions are given in the IV the conclusion on the heart disease diagnosis in given section V.

II. Related Work

Numerous works have done related to disease prediction systems using different data mining techniques and machine learning algorithms in medical centres.

K. Polaraju et al., [3] proposed Prediction of Heart Disease using Multiple Regression Model, and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work has performed using a training data set consisting of 3000 instances with 13 different attributes mentioned earlier. The data set has divided into two parts that are 70% of the data are used for training, and 30% used for testing. Based on the results, it is clear that the regression algorithm's classification accuracy is better compared to other algorithms.

Marjia et al. [4] developed heart disease prediction using KStar, j48, SMO, and Bayes Net and Multilayer perception using WEKA software. Based on performance from different factor, SMO and Bayes Net achieve optimum performance than KStar, Multilayer perception and J48 techniques using k-fold cross-validation. The accuracy performances achieved by those algorithms are still not satisfactory. Therefore, the accuracy's performance is improved more to give the better decision to diagnose disease.

S. Seema et al. [5] focuses on techniques that can predict chronic disease by mining the data containing in historical health records using Naïve Bayes, Decision tree, Support Machine(SVM) and Artificial Neural Network(ANN). A comparative study is performed on classifiers to measure the better performance on an accurate rate. SVM gives the highest accuracy rate from this experiment, whereas for diabetes, Naïve Bayes gives the highest accuracy.

Sairabi H.Mujawar et al., [6] used k-means and Naïve Bayes to predict heart disease. This paper builds the system using a historical heart database that gives a diagnosis. 13 attributes have considered building the system. Extraction from the database, data mining techniques such as clustering, classification methods have used 13 attributes with total of 300 records were used from the Cleveland Heart Database. This model predicts whether the patient has heart disease or not based on the values of 13 attributes.

Sharan Monica. L et al. [7] proposed an analysis of cardiovascular disease. This paper proposed data mining techniques to predict the disease. It intends to provide a survey of current techniques to extract information from the dataset, and it will be helpful for healthcare practitioners. The performance can be obtained based on the time taken to build the decision tree for the system. The primary objective is to predict the disease with a fewer number of attributes.

Sharma Purushottam et al. [8] proposed c45 rules and partial tree technique to predict heart disease. This paper can discover a set of rules to predict patients' risk levels based on given parameter about their health. The performance has calculated in measures of accuracy classification, error classification, rules generated and the results. Then comparison has done using C4.5 and partial tree. The result shows that there is potential prediction and more efficient. Table 2 describes the accuracy of heart disease with different techniques shown below.

III. Methodology

In this paper, we discussed about a general framework for analysing and predicting Heart

disease (HD). The proposed method flowchart has given in fig 1. The proposed model based on the ensemble model [9]. They are two types of terms describing the ensemble models:

A. Bagging: To decrease the model's variance

B. Boosting: To decrease the model's bias

Bagging

The bagging model is used to decrease the variance and increase the prediction accuracy. The model is creating several subsets of data from the training set, which is randomly selected.

Boosting

The boosting model is another ensemble technique that creates a group of predictions. The model learned sequentially through the early learner's fitting model. This model is used to analyse data and data errors. If the input data is irrelevant to the hypothesis, its weight is improved to attain better classification [10]. The gradient boosting is an extension of the Gradient Descent over boosting method.

$$Gradient\ Boosting = Gradient\ Descent + Boosting$$

In this model used in gradient descent algorithm, achieve different loss function. The ensemble of tree models has built one by one, and the tree entities have included successively. If the tree tends to improve the loss, it has depicted by the dissimilarity between actual and predicted values.

Table 1: Comparison of Bagging, Boosting and Stacking

Feature	Bagging	Boosting
Data from the subset	Randomly Selected	Misclassified samples gave higher preference
Objective	Minimises variance	Increases predictive force
Mechanics	Random subspace	Gradient descent
Model	(weighted) average	Weighted majority vote

An outline of the basic functionalities of XGB, GBM, RFC, ETC, and KNN has given in the following section.

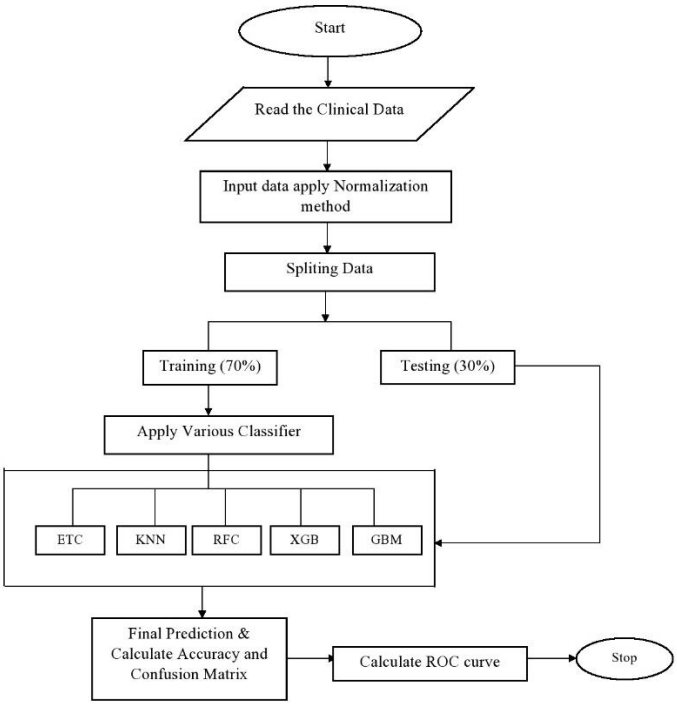

Fig 1: Flow chart of the proposed method

1. Random Forest Classifier

A random forest algorithm is a supervised classification algorithm that creates a forest with several trees. Generally, a random forest acts as a meta estimator that fits many decision tree classifiers to test the different sub-samples of the dataset and uses averaging to improve predictive accuracy and control over-fitting—a random forest classifier used in Banking, Medicine, Stock marker, E-commerce. A random forest algorithm is

used in the medical field to identify the correct combination of a set of components to validate the medicine. Random forest algorithm also helpful for identifying the disease by analysing the patient's medical records.

Random forest classifier advantages: i) The overfitting problem will never come when we use the random forest algorithm in any classification problem. B. The random forest algorithm has used for future engineering, which means identifying the most important features out of the available feature from the training dataset.

2. K-Nearest Neighbors

The classification of K-Nearest-Neighbours (KNN) uses case-based classification learning [11]. It is an extension of the neighboring techniques that are basic. The main steps of K-Nearest-Neighbors implementation are:

Similarity assessment: The comparison between the test and training set calculated. It has calculated by the Euclidean distance, Manhattan distance, Jacquard similarity co-efficient, and correlation coefficient. Among these, the Euclidean distance method has mostly used for a given feature test $(x_{j1,}x_{j2},\dots\dots\dots x_{ji})$ and training feature $(x_{j1,}x_{j2},\dots\dots\dots x_{jn})$, the Euclidean distance is calculated as follows:

$$d_j = \sqrt{\sum_{i=1}^{n}(test_{ki} - train_{kn})} \qquad (1)$$

where n the number of the feature vector is k is the number of training and testing and sample data. $d_j$is the Euclidean distance between the j^{th} sample of the training and testing data. The classification has applied to the test sample data, which classified into the classes according to every class's voting results.

3. Gradient Boosting Machine (GBM)

While dealing with boosting algorithms, two buzzwords, Bagging and Boosting, are frequently encountered. The term Bagging refers to constructing algorithms for learning on random data samples and using simple means to assess bagging probabilities. On the other hand, the term Boosting refers to a similar process, but samples' collection takes place more intelligently. It helps us to increase the weights of findings that are difficult to classify. In addition to bagging, the gradient booster classifier improves. The results are not selected based on the method of bootstrap, but the errors. It is possible to select the prediction from among the range of models such as decision trees, regressions, and classifiers. The new forecast learns from previous predictors' inaccuracy. Actual predictions require fewer time/iterations. However, the stop criteria should be carefully selected, or the process may result in overfitting of training data [13].

4. Extra Tree Classifier (ETC)

Extra Trees Classifier is a training system for an ensemble based on the decision tree's design. ETC is more like a method of random forests. The ETC has assumed those sub-data decisions to mitigate over-learning and over-adaptation [14].

Extra Trees has named for (Extremely Randomised Trees). Extra Tree Classifier adds one more step of randomisation to the random forest algorithm. Random forests will use random subsets of features to calculate the optimal split to nodes. For each feature within that random subset, a random split have implemented, and then the best feature to split will be chosen to compare inevitable randomly favored splits. Extremely randomised trees are far more efficient in computation than random forests, and their performance is almost always comparable.

1. Builds multiple trees with **bootstrap** = **False** by default, which means it samples without replacement.
2. Nodes are split based on **random** splits among a **random subset** of the features selected at every node.

5. Extreme Gradient Boosting (XGB)

XGB manages structured data in remarkable ways, in which computing and regularisation. The computational method uses the loss function's second-order gradients. It prepares more information about gradient direction and reaches the minimum function of loss—the loss function has applied simultaneously as average gradient boosting. The base model — decision tree is an alternative to mitigate the inclusive model's error. This model uses the derivative of the second-order as an approximation. Regularisation (L1 & L2) in this equation has observed to boost. The XGB algorithm has additional benefits: fast learning and cluster-wide parallelisation [15].

Algorithm1: Heart Disease Classification Algorithms

Input: Clinical Data set

Output: Classification of Heart Disease (Normal and Disease)

Phase I: Pre-Processing
Step 1: Read the Clinical Data
Step 2: Clinical Data using Pre-Processing Method
Step 3: Pre-Processing Data Apply Transformation
Step 4: Transformation data using Splitting for training and testing data
Phase II: Prediction
Step 5: The training and testing using various classifier algorithms
Step 6: testing data predicted in Heart Disease
Step 7: Calculated Confusion Matrix, ROC curve, AUC and Accuracy for model performance
Step 8: Stop

Dataset Description

This study normal and abnormal clinical data of heart disease from UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets /SPECTF+Heart data [16] have used to evaluate the performance of the various classifier model. It has tested on 187 clinical data in the database.

DESCRIPTION OF SPECT DATA

The SPECT database consists of images and clinical patient records. Data contained in the spreadsheet has converted to a relational database. Then it is analysed, and the significant attributes have extracted: encrypted patient ID, sex, weight, height, encrypted date of the study, 22 partial diagnoses, and the overall diagnosis.

All have recorded in a text file. The image database is also analysed. Images are stored in a predetermined directory structure, defined according to the encrypted study date and encrypted patient hospital number. There are two 3D images for each patient, one for each study and six 2D images (three for each study). The database design goal is to simplify maintenance and add new patients records and images when they become available.

In each patient, the study contains two three-dimensional cardiac SPECT image sets of the LV. A cardiologist diagnoses, say, Ischemia, Infarct or Artifact, by comparing these two images. Evaluation of the images is a highly subjective process, with excellent potential for substantial variability. We use a procedure similar to the one described by using [16] to analyse the images. The raw image data taken from multiple planar views has processed by filtered back-projection to create a three-dimensional image. Each of these 3D images has displayed as three sets of two-dimensional images. These 2D images correspond to the following sections of the LV myocardium: short-axis view, long horizontal axis view and long vertical axis view. From these 2D sets of images, a cardiologist selects five slices for each study that constitute the Yale system's final report.

Two slices have a short axis view-one slice near the hearts 'apex, one in the middle of the LV and one near the heart base.

- One slice corresponds to the centre of the LV cavity for horizontal long-axis view.
- One slice corresponds to the centre of the LV cavity for vertical long-axis view.

Each of these five images has divided into four or five regions of interest (ROI), along with the LV myocardium. As a result, for each study, there are 22 regions of interest. The cardiologist evaluates appearance and continues each of these regions. Comparison between corresponding ROIs in stress and rest study has performed. Partial diagnoses are made for each ROI by the cardiologist; they have classified into seven categories: Normal, Reversible, Partially Reversible, Artifact, Fixed, Equivocal and Reverse Redistribution. The cardiologist makes the overall diagnosis based on the partial diagnoses.

IV. RESULTS AND CONCLUSION

The classification report performs M1, M2, M3, M4 and M5 model in Tables 1,2,3,4,5 and 6. The different classifiers have tested using the dataset described, and the classification precision, accuracy, recall, F1-score have determined for each classifier. The results presented here will provide an evaluation of the execution of the different classifiers concerning each metric. The metrics have intended as the percentage of adequately classified samples separated by the sum number of samples. The results obtained for each model are numerically recorded in tables and respectively depicted as graphs.

Table 1: Classification model performance Analysis

MODEL	SCORE (ACCURACY) %
Extra Tree Classifier (ETC)	65.43
K-Nearest Neighbour (KNN)	75.31
Random Forest Classifier (RFC)	79.10
Extreme Gradient Boosting (XGB)	79.01
Gradient Boosting Machine (GBM)	83.95

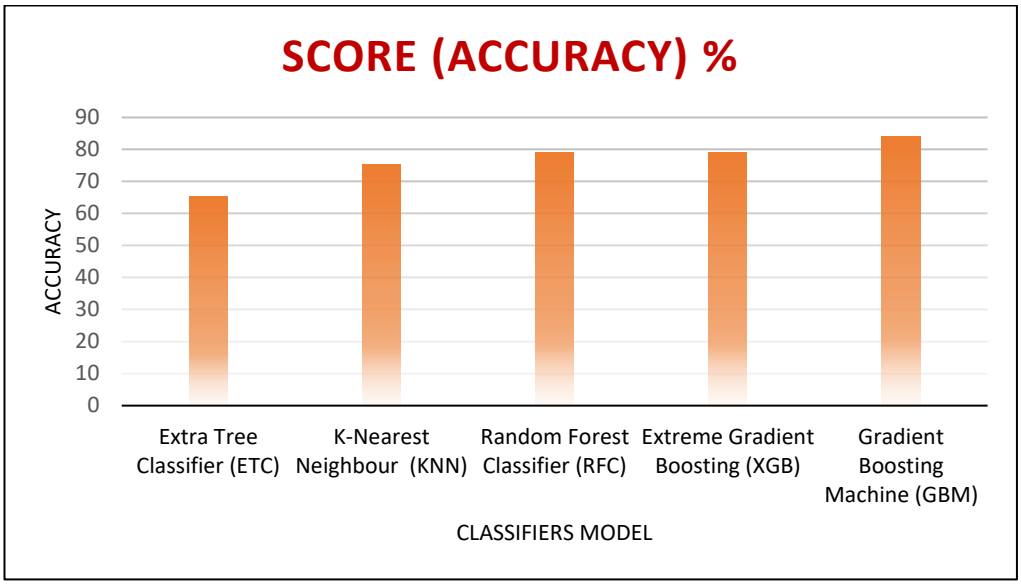

Fig. 2: Performance Analysis of Classifiers (Accuracy)

Fig 2 and Table 1 show a comparison of the classification accuracy obtained by various classifiers.

Case 1:

Table 2 shows a classifier M_1 model Precision, Recall and F1-Score obtained by ETC classifier. The M_1 classifier model accuracy is 65.43%.

Table 4: Performance analysis of ETC classifier

Group	Precision	Recall	F1-Score	Support
Normal	0.26	0.47	0.33	15
Disease	0.85	0.70	0.77	66
Classification model report				
Macro Avg	0.56	0.58	0.55	81
Weighted Avg	0.74	0.65	0.69	81
Accuracy: 0.6543 (65.43%)				

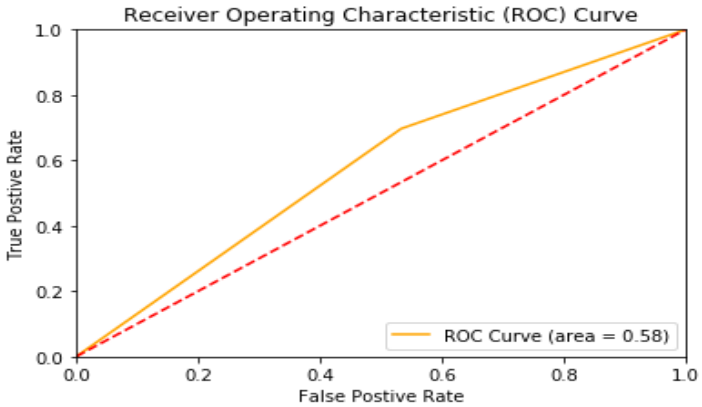

Fig 3: ETC Performance analysis of ROC Curve

ROC curve plot has visualised the performance of a binary classifier. It also specifies the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) at different classification

thresholds. The ROC results have annotated visualising the fig 3 is showing in ETC model ROC this model accuracy is 0.58.

Case 2:

Table 3 shows a classifier M_2 model Precision, Recall and F1-Score obtained by the KNN classifier. The M_2 classifier model accuracy is 75.31%.

Table 4: Performance analysis of KNN classifier

Group	Precision	Recall	F1-Score	Support
Normal	0.33	0.33	0.33	15
Disease	0.85	0.85	0.85	66
Classification model report				
Macro Avg	0.59	0.59	0.59	81
Weighted Avg	0.75	0.75	0.75	81
Accuracy: 0.7531(75.31%)				

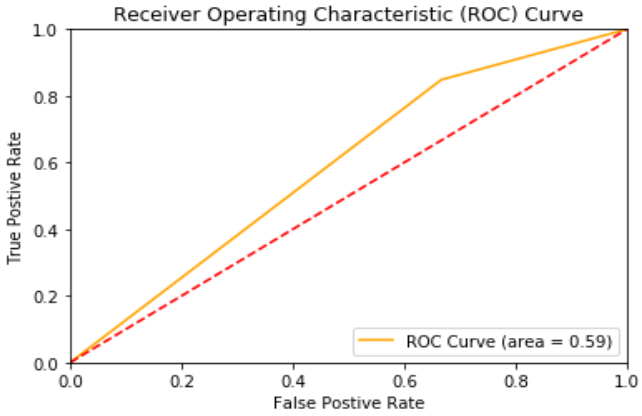

Fig 4: KNN Performance analysis of ROC Curve

It performs the trade-off between the True Positive Rate (TPR) and the False positive rate (FPR) at different classification thresholds. The ROC results have annotated visualising the fig 4 is showing in KNN classifier model ROC this model accuracy is 0.59.

Case 3:

Table 5 shows a classifier M_3 model Precision, Recall and F1-Score obtained by RFC classifier. The M_3 classifier model accuracy is 79.10%.

Table 5: Performance analysis of RFC classifier

Group	Precision	Recall	F1-Score	Support
Normal	0.44	0.47	0.45	15
Disease	0.88	0.86	0.87	66
Classification model report				
Macro Avg	0.66	0.67	0.66	81
Weighted Avg	0.80	0.79	0.79	81
Accuracy: 0.7910 (79.10%)				

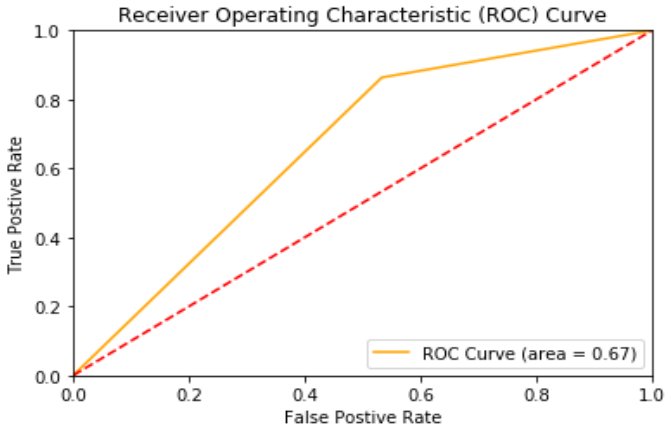

Fig 5: RFC Performance analysis of ROC Curve

The ROC results have annotated visualising the fig 5 is showing in RFC classifier model ROC this model accuracy is 0.67.

Case 3:

Table 6 shows a classifier M_4 model Precision, Recall and F1-Score obtained by XGB classifier. The M_4 classifier model accuracy is 79.01%

Table 6: Performance analysis of XGB classifier

Group	Precision	Recall	F1-Score	Support
Normal	0.43	0.40	0.41	15
Disease	0.87	0.88	0.87	66
Classification model report				
Macro Avg	0.65	0.64	0.64	81
Weighted Avg	0.78	0.79	0.79	81
Accuracy: 0.7910 (79.01%)				

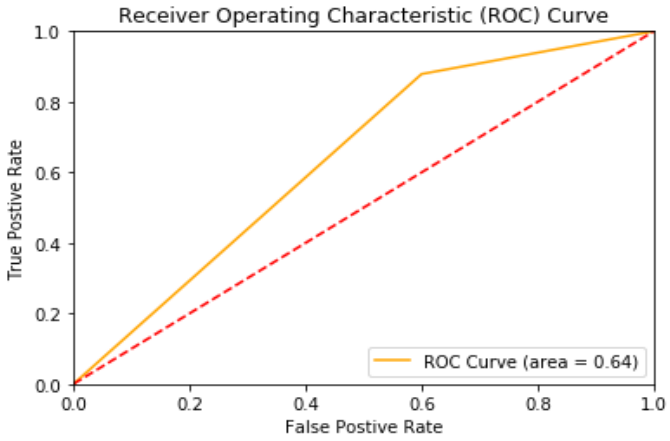

Fig 6: RFC Performance analysis of ROC Curve

The ROC results have annotated visualising the fig 6 is showing in XGB classifier model ROC this model accuracy is 0.64.

Case 4:

Table 7 shows a classifier M_5 model Precision, Recall and F1-Score obtained by GBC classifier. The M_5 classifier model accuracy is 83.95 %

Table 7: Performance analysis of GBC classifier

Group	Precision	Recall	F1-Score	Support
Normal	0.62	0.33	0.43	15
Disease	0.86	0.95	0.91	66
Classification model report				
Macro Avg	0.74	0.64	0.67	81
Weighted Avg	0.82	0.84	0.82	81
Accuracy: 0.8395 (83.95%)				

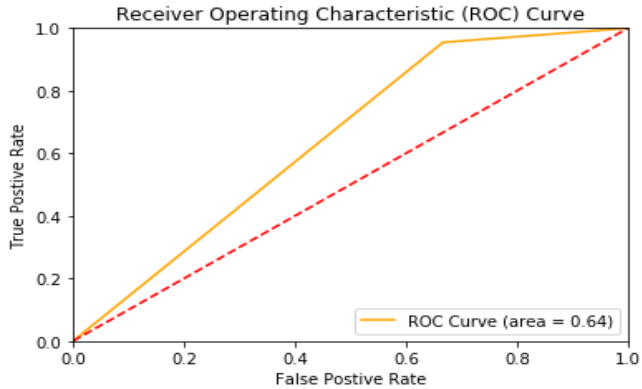

Fig 7: GBC Performance analysis of ROC Curve

The ROC results are being annotated visualising the fig 7 shows in GBC classifier model ROC this model accuracy is 0.66.

Table 8: Performance analysis of ROC (AUC) classifier

MODEL	ROC (AUC)
RFC	0.66
GBC	0.64
KNN	0.59
XGB	0.64
ETC	0.58

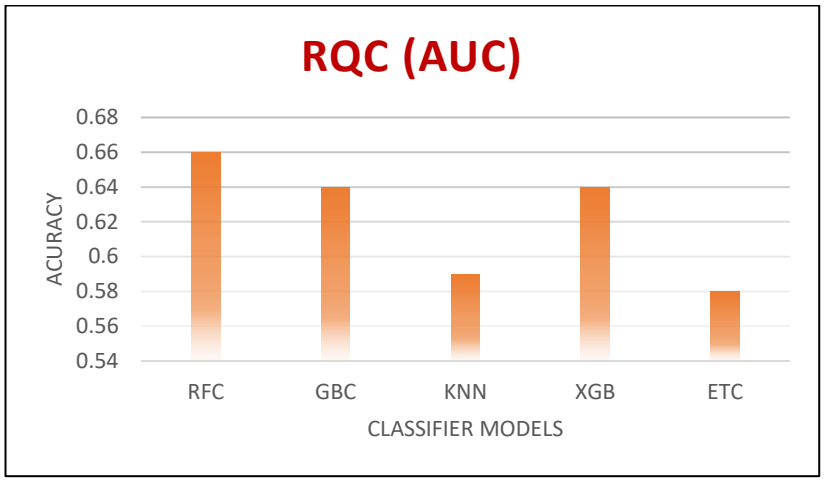

Fig 8: Comparison analysis of ROC curve

Fig 8 and Table 8 show a comparison of the classification accuracy obtained by various classifiers.

Table 8: Comparison analysis of Sensitivity and Specificity

MODEL	SENSITIVITY	SPECIFICITY
RFC	0.1666	0.7986
GBC	0.1111	0.9047
KNN	0.1666	0.8095
XGB	0.3333	0.8550
ETC	0.2222	0.6349

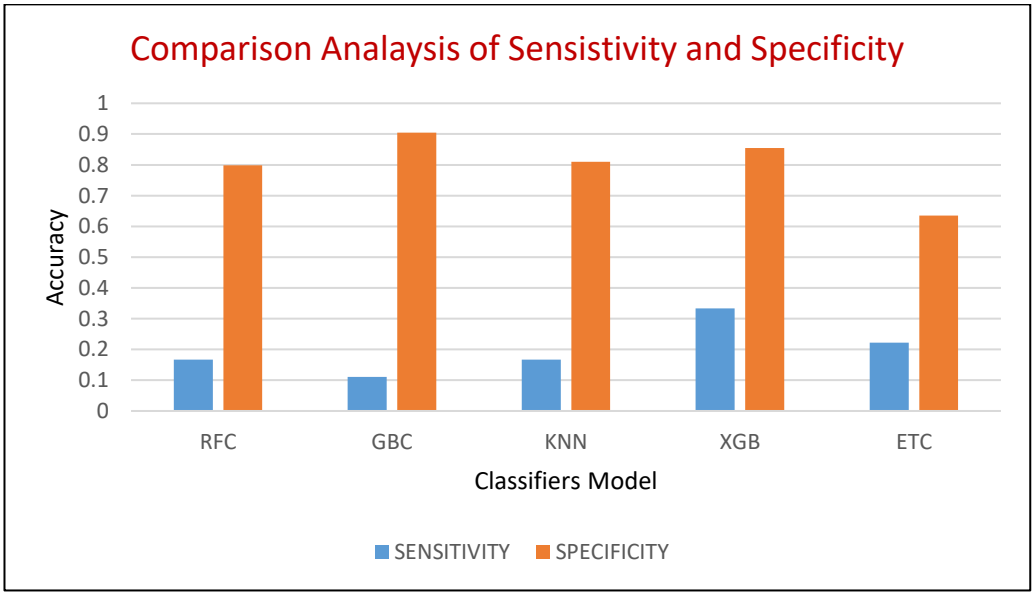

Fig 9: Comparison analysis of Sensitivity and Specificity

Fig 9 and Table 9 show a comparison of the classification sensitivity and specificity accuracy obtained by various classifiers.

Performance Evaluation Matrices

The evaluation metrics is an essential feature for the classifier model and performance assessment [19]. In Table 2 to Table 10, the confusion matrix demonstrates the results of incorrectly and correctly classified instances of each class in the three classes of the problems.

Table 10: Metrics from the confusion matrix

Metric	Computation
Accuracy	(TP+TN)
Misclassification Rate	(FP+FN)= total
True Positive Rate or Recall	TP=Actual Yes
False Positive Rate	FP=Actual No
True Negative Rate	TN= Actual No

The accuracy is mostly an accepted evaluation metric. The effectively measure the correct rates of all the classes. These metrics have defined as:

$$Precision = \frac{tp}{(tp+fp)} \quad (1)$$

$$Recall = \frac{tp}{(tp+fn)} \quad (2)$$

$$F - Measure = \frac{(1+\beta)^2 * Recall * Precision}{\beta^2 * Recall * Precision} \quad (3)$$

Where β is a co-efficient to adjust the importance of precision and recall (usually β =1)

$$FPrate = \frac{FP}{(FP+TN)} \quad (4)$$

$$TPrate = \frac{TP}{(TP+FN)} \quad (5)$$

CONCLUSION

The main goal of this paper aims to provide an insight into heart disease risk diagnosis using classification techniques. From the analysis, many authors used various classification techniques using a different number of attributes for study. It has proven that the proposed work achieves high efficiency compared to other existing works. The study concluded that Gradient Boosting Classifier achieved the highest accuracy and the accuracy level is 83.95%. In future, we can predict various stage predictions in heart disease.

REFERENCES

1. K. Polaraju, D. Durga Prasad, "Prediction of Heart Disease using Multiple Linear Regression Model", International Journal of Engineering Development and Research Development, ISSN:2321-9939, 2017.
2. Marjia Sultana, Afrin Haider, "Heart Disease Prediction using WEKA tool and 10-Fold cross-validation", The Institute of Electrical and Electronics Engineers, March 2017.
3. Dr. S. Seema Shedole, Kumari Deepika, "Predictive analytics to prevent and control chronic disease", https://www.researchgate.net/punlication/316530782, January 2016
4. Sairabi H.Mujawar, P.R.Devale, "Prediction of Heart Disease using Modified K-means and by using Naïve Bayes", International Journal of Innovative Research in Computer and Communication Engineering, vol.3, October 2015, pp.10265-10273.
5. Sharan Monica. L, Sathees Kumar. B, "Analysis of cardiovascular Disease Prediction using Data Mining Techniques", International Journal of Modern Computer Science, vol.4, 1 February 2016, pp.55-58.
6. Sharma Purushottam, Dr Kanak Saxena, Richa Sharma, "Heart Disease Prediction System Evaluation using C4.5 Rules and Partial Tree", Springer, Computational Intelligence in Data Mining, vol.2, 2015, pp.285-294. IJCATM :
7. https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9. Wolpert, D., Stacked Generalisation., Neural Networks, 5(2), pp. 241-259., 1992.
8. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/.
9. https://medium.com/@gabrieltseng/gradient-boosting-and-xgboost-c306c1bcfaf5 Grover, Prince. Gradient Boosting from scratch. https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d.Accessed:2019-01-04.
10. https://medium.com/@namanbhandari/extratreesclassifier-8e7fc0502c7
11. https://towardsdatascience.com/building-a-k-nearest-neighbors-k-nn-model-with-scikit-learn-51209555453a.
12. Precision :https://en.wikipedia.org/wiki/Precision_and_recall. Accessed: 2019-01-04. F1 Score. https://en.wikipedia.org/wiki/F1-Score. Accessed : 2019-01-04.