iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50111

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: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Hybrid Classification Using Ensemble Model to Predict Cardiovascular Diseases

Karthik Bogelly[1], C. Rahul Rao[2], Srikanth Uppari[3], K. Shilpa[4]
Dept. Of CSE, CMR Technical Campus

Abstract: ***Machine Learning is a widely used tool in the healthcare industry. Machine Learning algorithms help to predict and detect the presence of cardiovascular diseases. Such information, if predicted ahead of time, can provide important knowledge to doctors who can then diagnose and deal per patient basis. We work on predicting possible heart diseases in people using Machine Learning algorithms. In this project we perform the comparative analysis of classifiers like Naïve Bayes, SVM, Logistic Regression, Decision trees and Random Forest and we propose an ensemble classifier which perform hybrid classification by taking classifiers(strong and weak) since it can have multiple number of samples for training and validating the data so we perform the analysis of existing classifier and proposed classifier like Ada-boost and XG-boost combined with logistic regression and which can give the better accuracy and predictive analysis.***

Keywords: SVM; Naive Bayes; Decision Tree; Random Forest; Logistic Regression; Adaboost; XG-boost; python programming; confusion matrix; correlation matrix.

I. INTRODUCTION

As reported by the WHO, every year more than 10 million deaths occur globally due to heart diseases. It is one of the biggest causes of morbidity and mortality among the people on Earth. Predicting various cardiovascular disease is a significant topic in the section of data analysis. Heart diseases are on the rise rapidly around the world. It is a silent killer which shows no apparent and obvious symptoms. *Figure 1* show s the difference between diseased and healthy heart. Patients with high risk complications can benefit from early prediction of the heart disease and avoid any delay in treatment. Machine Learning has become a boon to society by assisting medical experts in making decisions and predictions from large data sets of patient data. This project aims to predict future heart diseases by analyzing patients' data which helps in segmenting whether a person has heart disease or not using machine learning algorithms. Common risk factors like exercise, smoking, drinking, genetic history of heart diseases, age, gender, lifestyle habits and ethnicity influence whether the person will have heart disease or not.

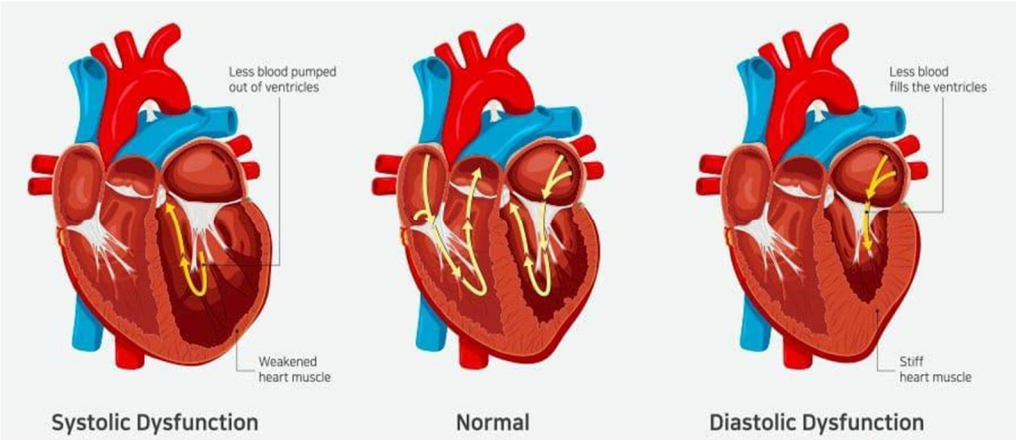


Figure 1: Healthy vs Unhealthy Heart

This investigation is done for identifying the optimal classification algorithm/s for identifying the possibility of heart disease in a patient. In this project we are conducting a comparative analysis using algorithms like Decision Tree, Naïve Bayes, Random Forest at different levels of evaluation. Despite the fact they are commonly used machine algorithms ,the prediction of heart diseases is a significant task involving highest possible accuracy. Hence all the algorithms are evaluated at numerous levels and types of evaluation strategies which provides researchers and medical experts to come to a right conclusion.

We propose an ensemble classifier which perform hybrid classification by taking classifiers (strong and weak) since it can have multiple number of samples for training and validating the data so we perform the analysis of existing classifier and proposed classifier like Ada-boost and XG-boost combined with logistic regression and which can give the better accuracy and predictive analysis. Dataset which contains 76 features and 14 important features that are useful to evaluate the system are selected among them. If all the features are taken into consideration, then the efficiency of the system the author gets is less. To increase efficiency, attribute selection is done.

II. LITERATURE SURVEY

Over the years, many research teams have conducted several tests. The following are a few of the groups:

1) Purushottam ,et ,al proposed a paper "Efficient Heart Disease Prediction System" using hill climbing and decision tree algorithms. They used Cleveland dataset and preprocessing of data is performed before using classification algorithms. The Knowledge Extraction is done based on Evolutionary Learning (KEEL), an opensource data mining tool that fills the missing values in the data set.A decision tree follows top-down order. For each actual node selected by hill-climbing algorithm a node is selected by a test at each level. The parameters and their values used are confidence. Its minimum confidence value is 0.25. The accuracy of the system is about 86.7%.
2) Santhana Krishnan. J ,et ,al proposed a paper "Prediction of Heart Disease Using Machine Learning Algorithms' ' using decision tree and Naive Bayes algorithm for prediction of heart disease. In the decision tree algorithm, the tree is built using certain conditions which give True or False decisions. The algorithms like SVM, KNN are results based on vertical or horizontal split conditions depending on dependent variables. But a decision tree for a tree-like structure having root nodes, leaves and branches based on the decision made in each of tree Decision tree also helps in the understanding of the importance of the attributes in the dataset. They have also used the Cleveland data set. Dataset splits in 70% training and 30% testing by using some methods. This algorithm gives 91% accuracy. The second algorithm is Naive Bayes, which is used for classification. It can handle complicated, nonlinear, dependent data so it is found suitable for heart disease dataset as this dataset is also complicated, dependent and nonlinear in nature. This algorithm gives an 87% accuracy
3) Sonam Nikhar et al proposed paper "Prediction of Heart Disease Using Machine Learning Algorithms" their research gives point to point explanation of Naïve Bayes and decision tree classifiers that are used especially in the prediction of Heart Disease. Some analysis has been led to think about the execution of prescient data mining strategy on the same dataset, and the result decided that Decision Tree has highest accuracy than Bayesian classifier
4) Aditi Gavhane et al proposed a paper "Prediction of Heart Disease Using Machine Learning", in which training and testing of dataset is performed by using neural network algorithm multi-layer perceptron. In this algorithm there will be one input layer and one output layer, and one or more layers are hidden layers between these two input and output layers. Through hidden layers each input node is connected to output layer. This connection is assigned with some random weights. The other input is called bias which is assigned with weight based on requirement the connection between the nodes can be feedforwarded or feedback.
5) Avinash Golande et al, proposed "Heart Disease Prediction Using Effective Machine Learning Techniques" in which few data mining techniques are used that support the doctors to differentiate the heart disease. Usually utilized methodologies are k-nearest Neighbour, Decision tree and Naïve Bayes. Other unique characterization-based strategies utilized are packing calculation, Part thickness, consecutive negligible streamlining and neural systems, straight Kernel self-arranging guide and SVM (Bolster Vector Machine).

III. EXISTING SYSTEM

The most difficult aspect of cardiac disease is detecting it. There are tools that can forecast heart disease, but they are either too expensive or too inefficient to quantify the risk of heart disease in humans. Early identification of heart disorders can reduce mortality and overall consequences. Unfortunately, it is not possible to precisely monitor patients every day in all circumstances, and 24-hour consultation by a doctor is not accessible since it demands more intelligence, time, and skill. We may use various machine learning algorithms to evaluate data for hidden patterns in today's environment since we have a large amount of data. The hidden patterns in medical data can be utilized for health diagnostics.

A. *Disadvantages Of Existing System*

1) Results of cardiovascular illness cannot be accurately predicted
2) Data mining techniques don't aid in making wise decisions.
3) Unable to handle large databases of patient information.

IV. PROPOSED SYSTEM

Data gathering and the selection of key attributes are the first steps in the system's operation. The necessary data is then preprocessed into the necessary format.

The data is then split into training data and testing data. With the training set of data, the algorithms are used to train the model. The system is tested using test data in order to determine its accuracy. The following modules are used to implement this system.

1) Dataset collection
2) Choose the qualities
3) Preprocessing of Data
4) Data balancing
5) Illness Prognosis

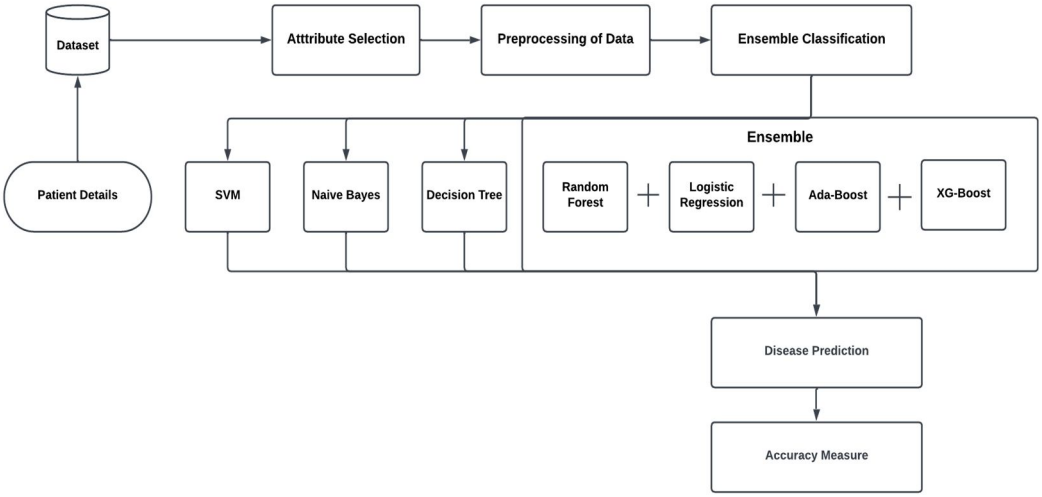


Figure 3: Architecture of Heart Disease Prediction

A. Dataset Collection

For the foundation of our cardiac disease prediction system, we first gather a dataset. We divided the dataset into training and assessment data after it was collected. The learning of the prediction model takes place on the training dataset, and the evaluation of the prediction model occurs on the testing dataset. 30% of the data are used for testing in this endeavour, and 70% are used for training. Heart Disease UCI served as the project's data source. The dataset has 76 attributes, of which the algorithm uses 14 for its operation.

The choice of suitable attributes for the prediction system is included in attribute or feature selection. This is done to make the method more effective. For the prediction, a number of patient characteristics are chosen, including gender, the nature of the patient's chest discomfort, fasting blood pressure, serum cholesterol, and exang. For this approach, attribute selection is done using a correlation matrix.

B. Preprocessing of Data

The pre-processing of data is a critical stage in the development of a machine learning algorithm. Data that isn't initially clean or in the model's necessary format can lead to inaccurate results. Pre-processing involves transforming data into the structure we need. It is used to handle the dataset's noise, duplicates, and missing numbers. Activities like importing datasets, dividing datasets, attribute scaling, etc. are all part of data pre-processing. Preprocessing the data is necessary to increase the model's precision.

C. Data balancing

Balanced datasets can be balanced in two ways. They are Under Sampling and Over Sampling (a) Under Sampling: In Under Sampling, dataset balance is done by the reduction of the size of the ample class. This process is considered when the amount of data is adequate. (b) Over Sampling: In Over Sampling, dataset balance is done by increasing the size of the scarce samples. This process is considered when the amount of data is inadequate.

D. Illness Prognosis

Various machine learning algorithms like SVM, Naive Bayes, Decision Tree, Random Tree, Logistic Regression, Ada-boost, Xg-boost are used for classification. Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy is used for heart disease prediction.

V. ALGORITHMS

A. Naive Bayes Algorithm

Based on the Bayes theorem, the Naive Bayes algorithm is a supervised learning algorithm used to address classification issues. It is mostly employed in text categorization with a large training set. One of the most straightforward and efficient classification algorithms is the Naive Bayes Classifier, which aids in the development of quick machine learning models capable of making accurate predictions. Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur. Spam filtration, Sentimental analysis, and article classification are a few examples of Naive Bayes algorithms that are frequently used. It is a classification method built on the Bayes Theorem and predicated on the idea of predictor independence The Naive Bayes model is simple to construct and is very beneficial for big data sets. Even with being straightforward, Naive Bayes is known to perform better than even the most complex classification techniques. The words Naive and Bayes, which make up the Naive Bayes algorithm, are as follows: Naïve: Because it presumes that the occurrence of one trait is unrelated to the occurrence of other features, it is referred to as naive. For example, if a fruit is recognized as an apple based on its red, spherical, and delicious fruit, form, and flavor. Hence, without relying on one another, each characteristic helps to recognize it as an apple.

Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

B. Decision Tree Algorithm

Although it may be used to solve classification and regression problems, Decision Tree is a Supervised learning technique that is typically used for classification problems. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result. The Decision Node and Leaf Node are the two nodes of a decision tree.

Whereas Leaf nodes are the results of decisions and do not have any more branches, Decision nodes are used to create decisions and have numerous branches. The given dataset's features are used to execute the test or make the decisions. The most important thing to keep in mind while developing a machine learning model is to select the optimal method for the dataset and task at hand. The two rationales for employing the decision tree are as follows:

1) Decision trees typically reflect how people think while making decisions, making them simple to comprehend.

2) The decision tree's reasoning is clear because it displays a tree-like structure.

C. Ensembled Classifiers

Hybrid classification is done by the usage ensemble model.Here are the algorithms included in the model:

- Random Forest
- Logistic Regression
- XG Boosting
- ADA Boosting

1) Random Forest Algorithm

A supervised learning algorithm is Random Forest. It is a development of machine learning classifiers that incorporates bagging to boost Decision Tree performance. It mixes tree predictors, and trees depend on an individually sampled random vector. All trees are distributed in the same way. Instead of splitting nodes based on variables, Random Forests uses the best among a prediction subset that is randomly selected from the node itself. The worst case of learning with Random Forests has a temporal complexity of O(M(dnlogn)), where M is the number of growing trees, n is the number of occurrences, and d is the data dimension.

Both classification and regression can be done with it. Moreover, it is the most user-friendly and adaptable algorithm. Trees make up a forest. A forest is supposed to be stronger the more trees it has. Using randomly chosen data samples, Random Forests build Decision Trees, obtain predictions from each tree, and then vote on the best answer. Also, it offers a fairly accurate indicator of the feature's relevance.

Applications for Random Forests include feature selection, picture classification, and recommendation engines. It can be used to categorize dependable loan candidates, spot fraud, and forecast sickness. The Boruta algorithm, which chooses significant features in a dataset, is built around it. Random Forest, as the name implies, is a classifier that uses a number of decision trees on different subsets of the provided dataset and averages them to increase the dataset's predictive accuracy. Instead, then depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of most predictions.
Higher accuracy and overfitting are prevented by the larger number of trees in the forest.

2) Logistic Regression

One of the most often used Machine Learning algorithms, within the category of Supervised Learning, is logistic regression. With a predetermined set of independent factors, it is used to predict the categorical dependent variable.
In a categorical dependent variable, the output is predicted via logistic regression. As a result, the result must be a discrete or categorical value. It can be either True or False, Yes or No, 0 or 1, etc., but rather than providing an exact value between 0 and 1, it provides probabilistic values that are in the range of 0 and 1. With the exception of how they are applied, logistic regression and linear regression are very similar. Whereas logistic regression is used to address classification issues, linear regression is used to address regression-related issues. In logistic regression, we fit a "S"-shaped logistic function instead of a regression line, which predicts two maximum values (0 or 1). The logistic function's curve shows the possibility of many events, such as whether or not the cells are malignant, whether or not a mouse is obese depending on its weight, etc. Because it can classify new data using both continuous and discrete datasets, logistic regression is a key machine learning approach

3) ADA Boosting

The first truly successful boosting algorithm created for binary classification was called Adaboost. Adaboost, which stands for Adaptive Boosting, is a very well-liked boosting strategy that turns several "poor classifiers" into one "strong classifier."

- Adaboost initially chooses a training subset at random.
- By choosing the training set based on the precise prediction of the previous training, it iteratively trains the Adaboost machine learning model.
- It gives incorrectly classified observations a larger weight so that they will have a higher chance of being correctly classified in the following round.
- It also gives the trained classifier weight in each iteration based on how accurate the classifier is. The classifier that is more precise will be given more weight.
- This approach iterates until the entire training set fits flawlessly or until the given maximum number of estimators has been reached.
- To categories, cast a "vote" using all of the learning algorithms you developed.

a) Advantages

Adaboost offers numerous advantages than SVM algorithms since it is simpler to use and requires less parameter fiddling. Plus, Although technically overfitting is not a feature of Adaboost applications, Adaboost can be utilized with SVM. This is likely because the parameters are not optimized simultaneously, and the learning process is hampered by estimate stage-by-stage. To better grasp maths, use this link.
The adaptable Adaboost may also be used to increase the accuracy of cases and weak classifiers in the categorization of images and texts.

b) Disadvantages

Adaboost employs a learning-based boosting method. Hence, in examples of Adaboost vs. Random Forest, high-quality data is required. Furthermore, before using the data, these characteristics must be removed because it is extremely sensitive to outliers and noise in the data. In addition, it moves significantly more slowly than the XG-boost method.

4) XG Boosting

Gradient Boosted decision trees are implemented using the XG-boost algorithm. It is a kind of software library that was primarily created to increase model performance and speed. Decision trees are generated sequentially in this approach. Weights are significant in XG-boost.

Each independent variable is given a weight before being fed into the decision tree that forecasts outcomes. More weight is applied to factors that the tree incorrectly anticipated, and these variables are subsequently fed into the second decision tree. Then, these distinct classifiers and predictors are combined to produce a robust and accurate model. It can be used for user-defined predict, classification, ranking, and regression.

a) *Regularization:* To prevent overfitting, XG-boost has built-in L1 (Lasso Regression) and L2 (Ridge Regression) regularization. Due to this, XG-boost is also referred to as the regularized form of GBM (Gradient Boosting Machine). We use the Scikit Learn library to feed two regularization-related hyper-parameters (alpha and lambda) to XG-boost. For L1 regularization, alpha is utilized, while for L2 regularization, lambda.

b) *Parallel Processing:* XG-boost makes use of parallel processing's strength, which accounts for its superior speed than GBM. The model is run on many CPU cores. Nthread hyper-parameter is used for parallel processing when using Scikit Learn library. The number of CPU cores available is represented by nthread. If you don't specify a value for nthread, the algorithm will automatically determine if you wish to use all the available cores.

c) *Managing Missing Values:* XG-boost comes with a built-in feature to manage missing data. When XG-boost comes across a missing value at a node, it experiments with both the left and right hand split and learns which method results in a bigger loss for each node. When working with the testing data, it follows suit.

d) *Cross Validation:* XG-boost makes it possible to perform a cross-validation at each stage of the boosting process, making it simple to determine the precise ideal number of boosting rounds to perform in a single run. This contrasts with GBM, where only a small number of variables may be examined without requiring a grid-search.

e) *Efficient Tree Pruning:* When a GBM experiences a loss during a split, it will cease dividing that node. As a result, it is a greedier algorithm. On the other hand, XG-boost makes splits up to the maximum depth given before beginning to prune the tree backwards and removing splits beyond which there is no net gain.

VI. DATASET DETAILS

Of the 76 attributes available in the dataset,14 attributes are considered for the prediction of the output.

S. No.	Attribute	Description	Type
1	Age	Patient's age (29 to 77)	Numerical
2	Sex	Gender of patient(male-0 female-1)	Nominal
3	Cp	Chest pain type	Nominal
4	Trestbps	Resting blood pressure(in mm Hg on admission to hospital ,values from 94 to 200)	Numerical
5	Chol	Serum cholesterol in mg/dl, values from 126 to 564)	Numerical
6	Fbs	Fasting blood sugar>120 mg/dl, true-1 false-0)	Nominal
7	Resting	Resting electrocardiographics result (0 to 1)	Nominal
8	Thali	Maximum heart rate achieved(71 to 202)	Numerical
9	Exang	Exercise included agina(1-yes 0-no)	Nominal
10	Oldpeak	ST depression introduced by exercise relative to rest (0 to .2)	Numerical
11	Slope	The slop of the peak exercise ST segment (0 to 1)	Nominal
12	Ca	Number of major vessels (0-3)	Numerical
13	Thal	3-normal	Nominal
14	Targets	1 or 0	Nominal

VII. PERFORMANCE ANALYSIS

This project uses a variety of machine learning techniques to predict cardiac illness, including SVM, Naive Bayes, Decision Trees, Random Forests, Logistic Regression, Adaboost, and XG-boost. Only 14 of the 76 features in the Heart Disease UCI dataset are taken into account when predicting heart disease. For this project, a number of patient characteristics are taken into account, including gender, the type of chest pain, fasting blood pressure, serum cholesterol, exang, etc. The accuracy of each algorithm must be measured, and whichever method provides the best accuracy is taken into consideration for the prediction of heart disease. Several assessment criteria, including accuracy, confusion matrix, precision, recall, and f1-score, are taken into account when evaluating the experiment.

Accuracy- Accuracy is the proportion of correctly predicted events to all of the dataset's inputs.

Accuracy = (TP + TN) /(TP+FP+FN+TN)

Confusion Matrix- It gives us a matrix as output and gives the total performance of the system.

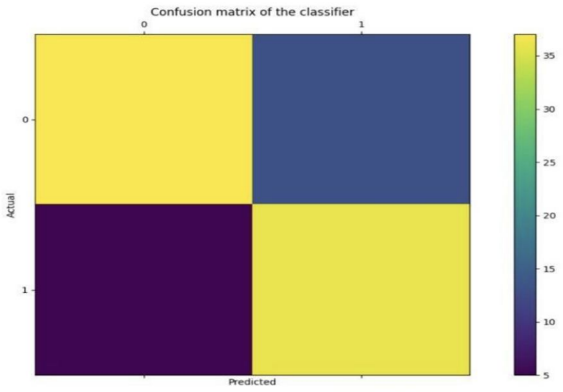


Figure 4: Confusion Matrix of the Ensemble Classifier

Where TP: True positive

FP: False Positive

FN: False Negative

TN: True Negative

Correlation Matrix: The correlation matrix in machine learning is used for feature selection. It represents dependency between various attributes.

Precision- It is the ratio of correct positive results to the total number of positive results predicted by the system.

It is expressed as: Recall-It is the ratio of correct positive results to the total number of positive results predicted by the system. It is expressed as:

F1 Score-It is the harmonic mean of Precision and Recall. It measures the test accuracy. The range of this metric is 0 to 1.

VIII. PERFORMANCE MEASURES

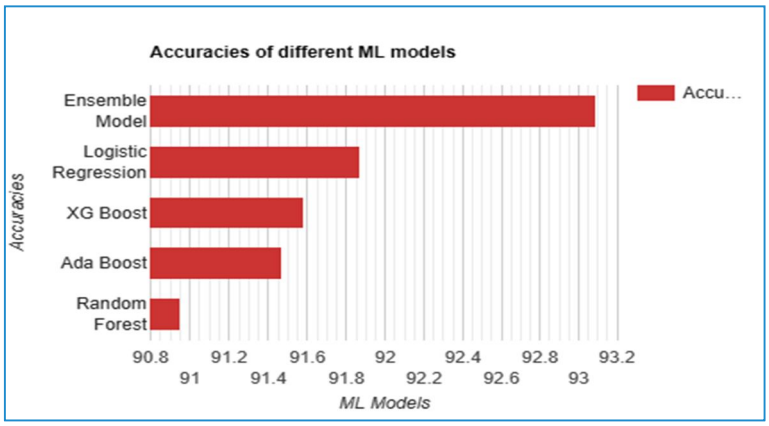


Figure 5: Comparison of accuracies of Ensemble model with other algorithms

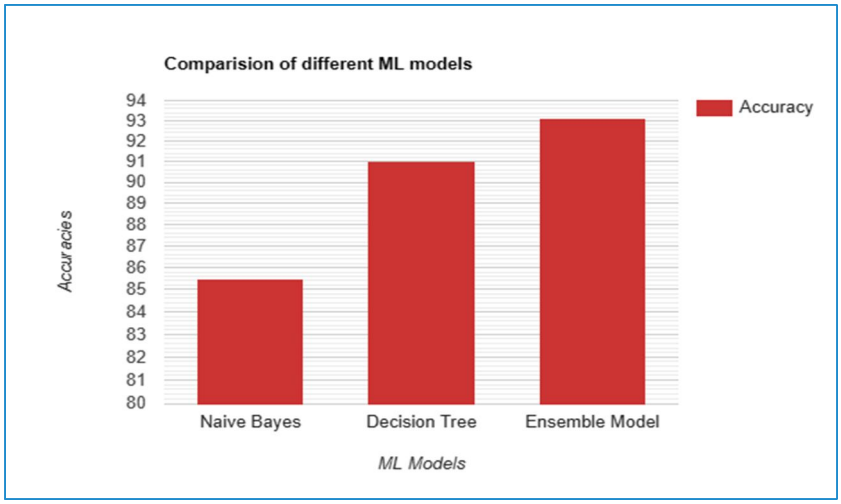


Figure 6: Comparison of accuracies of different algorithms

IX. RESULT

After performing the machine learning approach for training and testing we find that accuracy of the Ensemble Model is better compared to other algorithms. Accuracy is calculated with the support of the confusion matrix of each algorithm, here the number count of TP, TN, FP, FN is given and using the equation of accuracy, value has been calculated and it is concluded that Ensemble Model(Logistic Regression+Random Forest+XG boosting+ADA boosting) is best with % accuracy and the comparison is shown below.

ALGORITHM	ACCURACY
Decision Tree	85.31
Naïve Bayes	90.03
Ensemble Model • Logistic Regression • Random Forest • XG boosting • ADA boosting	93.23

X. CONCLUSION

Heart diseases are a major killer in India and throughout the world, application of promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society. The early prognosis of heart disease can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. The number of people facing heart diseases is on a raise each year. This prompts for its early diagnosis and treatment. The utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity and patients.

In this paper, the seven different machine learning algorithms used to measure the performance are Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, Adaptive Boosting, and Extreme Gradient Boosting applied on the dataset. The expected attributes leading to heart disease in patients are available in the dataset which contains 76 features and 14 important features 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 the author gets is less.

To increase efficiency, attribute selection is done. In this n features have to 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. If all the attributes present in the dataset are taken into account, then the efficiency decreases considerably.

All the seven machine learning methods accuracies are compared based on which one prediction model is generated. Hence, the aim is to use various evaluation metrics like confusion matrix, accuracy, precision, recall, and f1-score which predicts the disease efficiently.

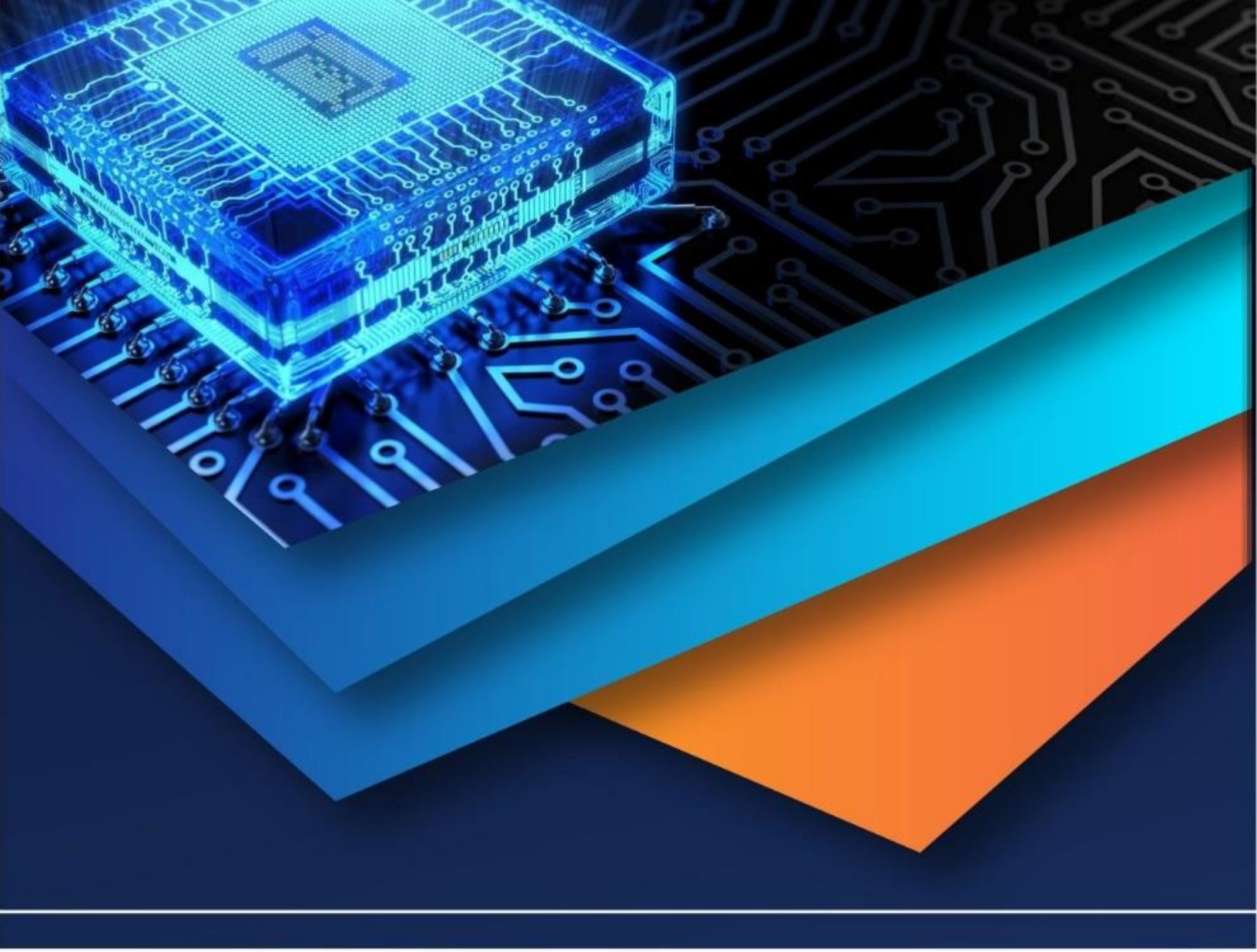
The highest accuracy of 93.23 is given by Ensemble classifier which perform hybrid classification by taking classifiers (strong and weak) since it can have multiple number of samples for training and validating the data so we perform the analysis of existing classifier and proposed classifier like Ada-boost and XG-boost combined with logistic regression and Random Forest.

XI. ACKNOWLEDGEMENT

We thank CMR Technical Campus for supporting this paper titled “Hybrid Classification Using Ensemble Model To Predict Cardiovascular Diseases”, which provided good facilities and support to accomplish our work. Sincerely thank our Chairman, Director, Deans, Head Of the Department, Department Of Computer Science and Engineering, Guide and Teaching and Non-Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

[1] Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8
[2] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.
[3] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.
[4] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.
[5] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259- 64). IEEE. ICCRDA 2020 IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012072 IOP Publishing doi:10.1088/1757- 899X/1022/1/012072 9
[6] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY