[1]Rahbre Islam
[2]Safdar Tanweer
[3]Md Tabrez Nafis
[4]Imran Hussain
[5]Md Onais Ahmad

Enhancing Accuracy and Analyzing Performance of Machine Learning Models Using Random Forest for Heart Disease Prediction

Journal of Electrical Systems

Abstract: **-** Heart related disease widely known as cardiovascular disease (CVD), is a major issue that is getting worse globally. Precautionary steps are necessary before analyzing medical data to detect any disease early on to avoid fatalities. An advanced machine learning technique such as random forest (RF) may be very helpful to medical fraternity in understanding and enhancing their decisions on healthcare. This is critical to establishing a more hygienic setting for patient diagnosis and care. We applied RF, a kind of machine learning method, to real-time patient data in our study. Using a unique method to select significant features, RF outperformed the other models recently developed with an astounding accuracy of 94.56% rate using feature selection technique. This demonstrates the potential of machine learning, particularly RF, which boosts our capacity for effective prediction and treatment of cardiovascular illnesses.

Keywords:Heart disease, machine learning, random forest algorithm, feature selection.

I. INTRODUCTION

Diagnosis and prognosis of any disease are important in the healthcare sector though it is considered a hectic task. It needs to be done diligently so that its automation may be beneficial for stakeholders [1]Sometimes clinical experts like physicians, pathologists and even a pool of experts unable to predict a disease. So, the computer-based information systems play a vital role to reduce clinical expenses and enhance the quality of medical care. To ensure that computer systems are working well, we should test different techniques. In most cases hospitals use these systems for patients and their data, but sometimes they produce lots of irrelevant information that are not very useful for making decisions in the healthcare sector. Currently CVDs are the most common illness the world is experiencing [2].World Health Organization reports that more than 17.9 million heart disease casualties took place as per the fact sheet of 2019 which is an estimation of 32% of global deaths [3]. Many organizations use data mining (DM) techniques in the medical field for decision-making and identifying patterns of complex datasets [4].

Nowadays many scientists are using a data mining approach to identify how heart disease evolves. Our aim in this study is also to collect vital features for a doctor to make better decisions. Most of them used a proven method to predict and detect heart disease in the earliest stages. So researchers created many Machine Learning (ML) tools to enhance assistance [5]One ML classifier, Random Forest (RF) has been applied with FS to compare outcomes of the other models studied recently.

II. RELATED WORK

In the context of machine learning approaches G. S. Reddy et al. [6] employed logistic regression (LR) and random forest (RF) classifiers to predict heart disease and found RF with 87.64% mean accuracy and outperformed LR with a difference of 7.64%. T. Poojitha et al [7] used a novel Random Forest (RF) and K Nearest Neighbor (KNN) and tested for their predictive power of cardiovascular disease (CVD) and demonstrated that RF is better with an accuracy of 90.16%. Khan and colleagues [8] employed Decision tree, random forest , logistic regression, Naïve Bayes, and support vector machine algorithms to correctly predict and make decisions for CVD patients from Khyber Teaching Hospital. The performance assessment of the algorithms considered based on various conditions to identify the most suitable model. The RF algorithm

[1] Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India rahberislam@gmail.com
[2] Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India safdartanweer@yahoo.com
[3] Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India tabrez.nafis@gmail.com
[4] Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India ihussain@gmail.com
[5] Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India onaisahmad@gmail.com

shown the highest accuracy, sensitivity, and ROC curve at 85.01%, 92.11%, and 87.73%, respectively, for CVD.

In order to forecast cardiac disease and create decision rules that would make the correlations between input and output variables more clear, Peker et al.[9] employed Classification and Regression Tree (CART) algorithm in this study. The study also assigns a priority ranking to the characteristics that affect heart disease. Taking into account all performance characteristics, the model's reliability is proven with an accuracy of 87% in prediction. Bhatt et al. [10] used DT, XGBoost, RF and MLP on a real-dataset of 70,000 instances from Kaggle and found RF with an accuracy of 87.05% applying with cross-validation (CV) and 86.92% without CV.

Paramasiva and associates [11] studied, diverse ML algorithms, including Decision Tree, Discriminant Analysis, Logistic Regression, Naïve Bayes, Support Vector Machines, K-Nearest Neighbors ,and Ensemble classifiers trained on the Cleveland heart disease dataset.10-fold cross-validation used, both with and without the application of Principal Component Analysis and found an accuracy of 85.8% by LR with PCA where they used 9 components. Using the Framingham dataset from the Kaggle platform as the research sample, Wang and associates [12]conducted empirical analysis. AUC value, ROC curve, accuracy, error rates, and support vector machines (SVM), CNN, Random Forest (RF), and CART are the usual performance evaluation measures used in the study to compare the C-RF model with these models. The results show that the C-RF model has an 85% classification accuracy, which is an improvement of 8%, 9%, 4%, and 3% over CART, SVM, CNN, and RF, in that order.

In order to forecast anomalies associated to the heart, Upadhyay et al.[13] used logistic regression (LR) and random forest (RF). Their random forest accuracy was 80%, which is 5% less accurate than their logistic regression accuracy. Yahya et al. [14] conducted a predictive analysis on heart disease patient data to identify potential risk factors associated with their heart disease status. Two distinct heart disease datasets, Cleveland and Statlog, were utilized for classification model building and results validation, respectively. The Cleveland data underwent a thorough exploratory analysis using the Chi-square test, revealing strong associations ($p < 0.001$) between certain bio-clinical categorical variables and heart disease conditions. Ten classification models were trained, where support vector machine (SVM) shown an accuracy of 85% surpassing all with the best predictive performance.

Three AutoML programs (PyCaret, AutoGluon, and AutoKeras) were examined by Paladino et al. [15] using three different datasets (Cleveland, Hungarian, and a mixed dataset). They examined the effectiveness of automated machine learning (AutoML) techniques for diagnosing heart illness, and the results showed that AutoGluon outperformed the competition with accuracy rates ranging from 78% to 86%.

Bakar et al. [16] conducted a comparative analysis between the Random Forest (RF) and Artificial Neural Network (ANN) performance on the Heart Disease Prediction Model dataset from the UCI repository. After employing K-Fold Cross-Validation and dataset splitting they found both ANN and RF with accuracy rates of 67.9% and 64.6% respectively. In their investigation into cardiovascular disease prediction models, Mahmud and colleagues [17] emphasized the potency of Random Forest when its hyper parameters are fine-tuned. In order to increase predictability, ensemble methods and a hybrid bag-and-stacking strategy are advised. Here Kaggle dataset with 70,000 unique variables used to develop an ML and found an accuracy of 84.03%.

Carrying the same concept Stonier and his colleagues [18] compared ML techniques such as Random Forest, Regression models, K-nearest neighbor imputation, Naïve Bayes algorithm, and more and assessed,. It is evident from the results that the Random Forest algorithm works better than the others, with a higher accuracy of 88.52% in predicting the risk of heart attacks. Table 01 depicts the concise overview of reviewed literature section.

A. Inference from the related work

After reviewing the recently developed ML models in the literature I came to know that the primary goal of a researcher would be continuously enhance computer models that forecast cardiac disease. To improve these models and increase their applicability in practical scenarios, researchers will also investigate various approaches, datasets, and automated processes. It's critical to ascertain which variables lead to precise forecasts and to ensure that these models are reliable and equitable. Incorporating the theme the goal my proposed work is to develop an ML model with enhanced accuracy rate comparing to the studied model’s accuracies.

III. MATERIALS AND METHODS

Using the Random Forestalgorithm to create a trustworthy and an accurate prognostic model for heart disease based on real-time dataset.

- Data collection and description
- Data pre-processing
- Classification

A. Data Collection and Data Description

Real-time dataset for this study have been acquired from Varun Arjun Medical College and Rohilkhand Hospital, Shahjahanpur, U.P, India. This CVD dataset consists of 820 instances with 28 attributes, including the target attribute ("Target").

Here, Target is categorized as either "0" or "1" that represents having no CVD and having CVD as shown with a percentage in Fig. 02. Age, Ht.m2. (Height in square meters), Wt.kg. (Weight in kilograms), BMI .kg/m2. (Body Mass Index in kilograms per square meter), SBP (Systolic Blood Pressure), DBP (Diastolic Blood Pressure), HR (Heart Rate), PP (Pulse Pressure), RBP (Resting Blood Pressure), Chol (Cholesterol), MHR (Maximum Heart Rate), OPK (old peak), CPT (Cardiopulmonary Testing), FBS (Fasting Blood Sugar), RES (Resting Energy Expenditure), EX (Exercise), Slope (heart rate slope during ECG), VCA (Vascular Compliance Assessment), THA (Thalliumscan), Physical_Act (Physical Activity), Smoking, Alcohol_Drinking, HTN (Hypertension), Family History of CVD, Stress, Sex, Diabetes, and Target out of 28 including biomarkers. One attribute, Target, is taken as the dependent variable so as to predict the disease, whether a patient has CVD or not.

Attribute	Description
Age	Age of the individual (years)
Ht.m2	Height in square meters
Wt.kg	Weight in kilograms
BMI.kg/m2	Body Mass Index
SBP	Systolic Blood Pressure
DBP	Diastolic Blood Pressure
HR	Heart Rate
PP	Pulse Pressure
RBP	Resting Blood Pressure
Chol	Cholesterol levels in the blood
MHR	Maximum Heart Rate
OPK	Old Peak (cardiology measure)
CPT	Cardiopulmonary Testing
FBS	Fasting Blood Sugar
RES	Resting Energy Expenditure
EX	Exercise (physical activity)
Slope	Heart rate slope during ECG
VCA	Vascular Compliance Assessment
THA	Thalliumscan (assess blood flow to the heart)
Physical_Act	Physical Activity level
Smoking	Smoking status
Alcohol_Drinking	Alcohol consumption status
HTN	Hypertension status
Family History of CVD	Family History of Cardiovascular Disease
Stress	Stress status
Sex	Gender of the individual
Diabetes	Diabetes status
Target	Target variable (outcome or condition to be predicted)

B. Data pre-processing

Preparing raw data for analysis and modelling is known as data pre-processing, and it is an essential stage in machine learning. It guarantees that the data is consistent, clean, and appropriate for the methods of choice.

- **Loading Dataset**
  First of all, the CVD dataset is loaded using a read.csv () function from R Studio. After loading, a preliminary exploration of the dataset is done to visualize the nature and structure of the dataset with **str ()** function that helps to trace variable types, and any missing values as well.

- **Missing Values Treatment**
  Any missing values may hamper our analytic or modeling attempt so in order to decide whether to remove or impute missing values, the dataset is inspected for any missing values. Utilizing **na.omit()** function dataset cleaned for further executable form. To ensure the data integrity redundant records are also identified and cleaned.

- **Feature Selection using Boruta algorithm**

- **Categorical Data**

  Target variable converted into factors using **as.factor ()** from R Studio for ML model building.

- **Splitting of Dataset**

  The dataset is split into training and testing sets for model evaluation with 80:20.

- **Structure of dataset**

Age	Ht.m2.	Wt.kg.	BMI .kg.m2.	SBP	DBP	HR	PP	RBP	chol	MHR	OPK	CPT	FBS	RES	EX	slope	VCA	THA	Physical_Act	Smoking	Alcohol Drinking	HTN	Family. History.of. CVD	Stress	Sex	Diabeties	Target
int	num	num	num	int	int	int	int	int	int	int	num	in	int	in	int	int	int	int	int	int	int	int	int	int	factorial	int	factotial

Fig.1. Structure of the data set

C Glimpse of Target variables

Data balancing is essential for accurate results before and after the classifier is being applied. The below graph shows whether the target classes are equal or not, where "1" represents heart disease patients and "0" represents no heart disease.

- **Bar Plot**

It uses horizontal or vertical columns to represent data values for different categories or groups. A longer bar indicates higher values, allowing easy comparison of a single variable across multiple groups[19].

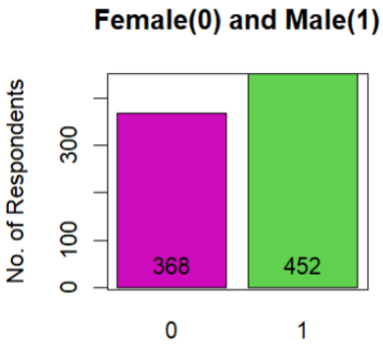

Fig. 02 Overall groups

```
25-30 31-35 36-40 41-45 46-50 51-55 56-60 61-65 66-70 71-75
   26    18    26    18    45    41    50    70    11     9
```

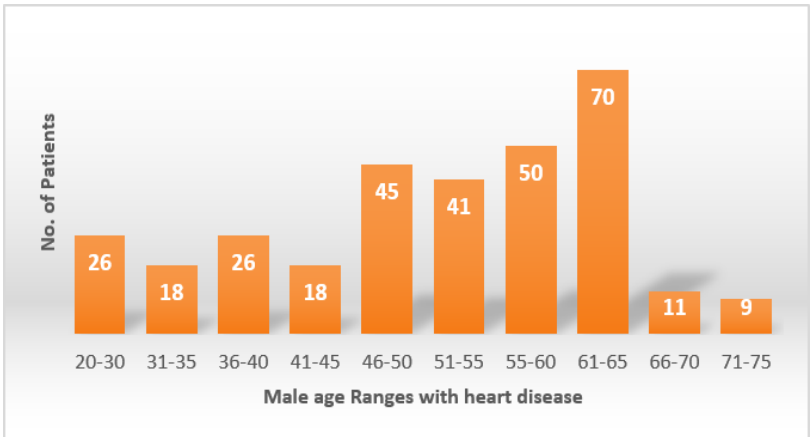

Fig.03 Bar graph of Male age Ranges with heart disease

25-30	31-35	36-40	41-45	46-50	51-55	56-60	61-65	66-70	71-75
15	18	20	32	52	23	28	30	8	5

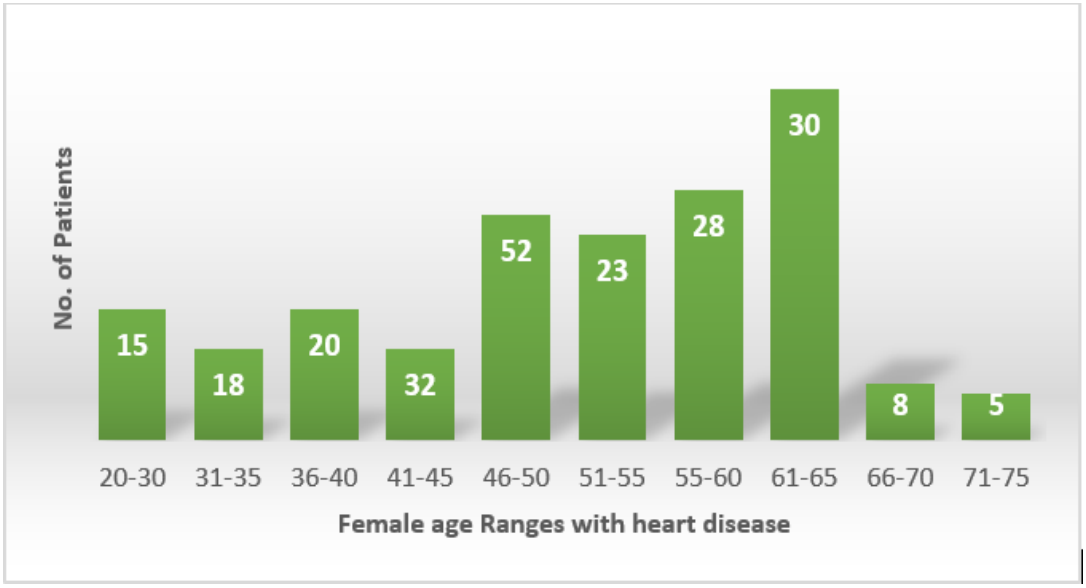

Fig. 04 Bar graph of Female age Ranges with heart disease

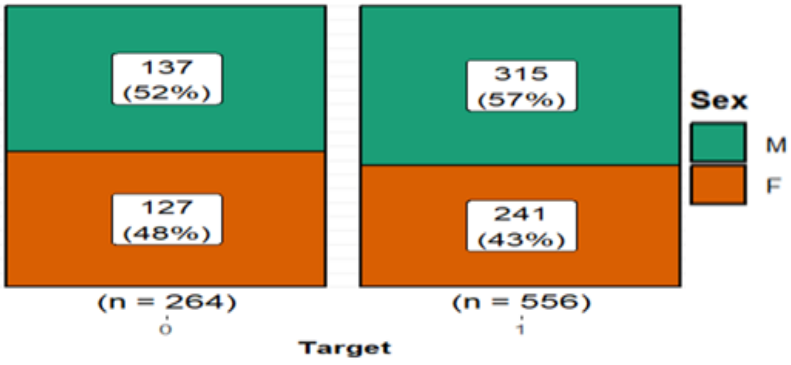

Fig. 05 Number of patients with or without heart disease

IV. METHODOLOGY, FEATURE SELECTION, APPROACH AND SOLUTION

This paper describes an intelligent approach to predict CVD that involves using a reduced set of vital features using Boruta feature selection technique and employs Random Forest (RF) and compares its performance with the recently developed ML models.

The Random Forest classification algorithm forms the foundation of the Boruta algorithm. Thus, the technique is applied to validate the ultimate feature list as the most crucial one [20]. It can be incorporated via "Boruta package" and using library (Boruta) in R Studio to fetch the important features of a dataset. A graphical representation of features is reflected in Fig. 03.

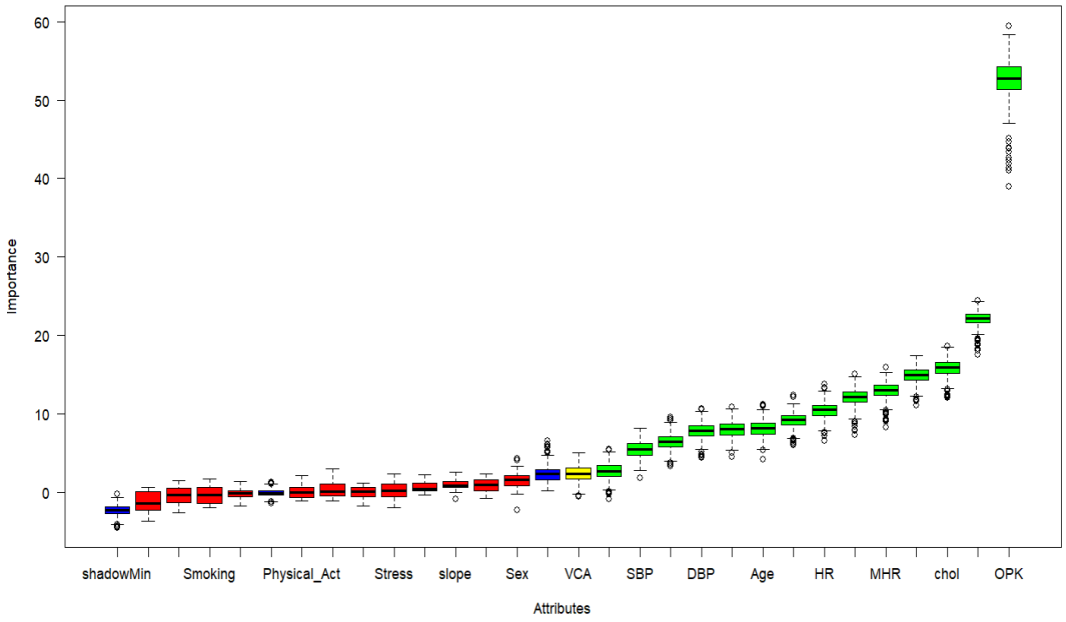

Fig. 06 Feature importance

Here the feature importance is displayed in the box plots using the colors green, red, yellow, and blue. Green: With the highest number of 14 attributes verified key characteristics. Red: With 12 verified inconsequential characteristics. Yellow: With only 1 potential characteristic that requires more investigation. Blue: With 3 attributes shadow characteristics, creating a feature importance baseline. Total 14 variables are verified as confirmed important like Age, Ht.m2., Wt.kg., BMI..kg.m2., SBP,DBP, HR, PP, RBP, chol, MHR,OPK, CPT, and EX. All important attributes came after 499 iterations in 37 seconds.

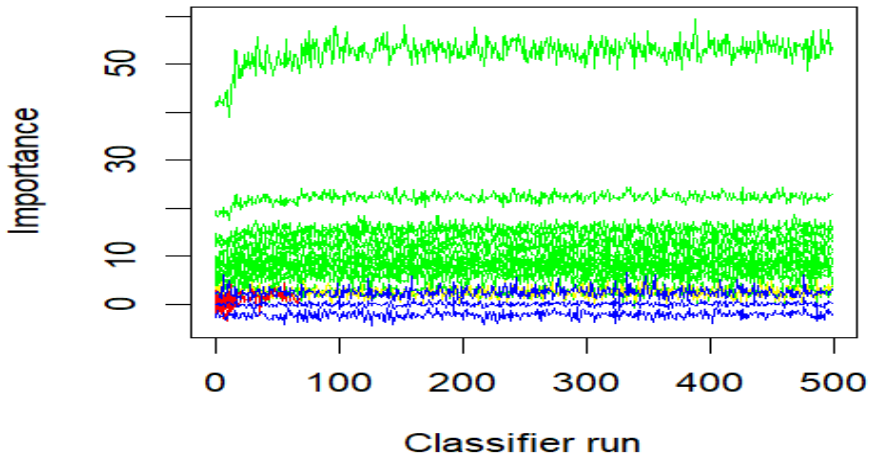

Fig. 07 Important history graph of features

Boruta feature screening is widely used in data mining research as well as clinical treatment. To efficiently isolate the most significant predictors, it uses an iterative procedure to address random fluctuations in relevance ratings within random forests and interactions between components. This approach is also often used in data mining studies to choose features [12].The relationship among confirmed important attributes shown in the below graph with respect to Target attribute.

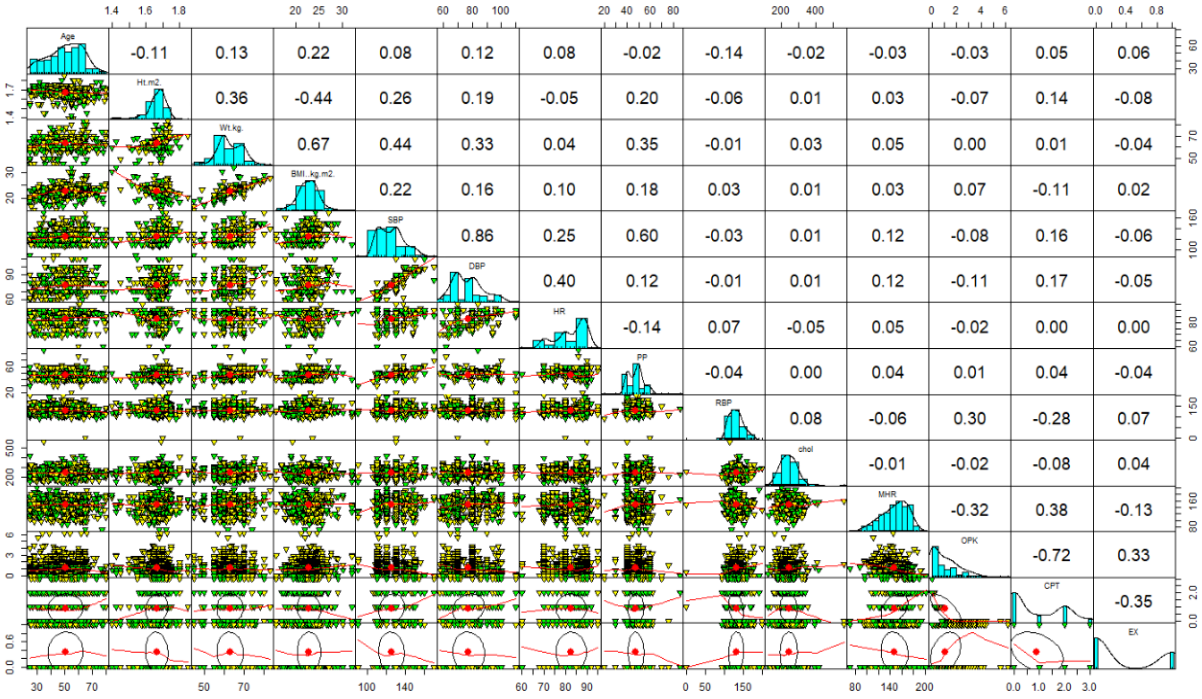

Fig. 08 Pair wise relationship among 14 important attributes with respect to Target

A. Classifier Applied

1) Random Forest (RF)

An ensemble learning technique called random forest that aggregates the results of various decision trees to create a prediction. The forecast is based on the collective will of the individual [21].The mean of the predictions given by the trees inside it is computed. Using Eq. (1), the prediction for a random forest with 'm' trees and individual weights 'W_j' is derived. As a tree goes deeper, the variation increases for a given change in the input.

Each decision tree in RFs aims to categorize a different subset of the input vector, which is vectorised input data.

The RFs incorporate the decision trees' functionality as the input vector passes through each tree in the forest, allowing each tree to classify the input vectors based on the particular subset of the vector it gets as an input [22].

The random forest selects the classification outcome for an input vector based on either the class with the highest number of "votes" or the average of all the "votes" after all decision trees in the random forest have categorized the input vector. The issue of overfitting from a single decision tree is avoided by RFs' usage of several decision trees. If there is a noticeable class disparity in the data, the voting rule may be changed [22].

$$\hat{y} = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} W_j(x_i, x') \qquad \text{Eq. (1)}$$

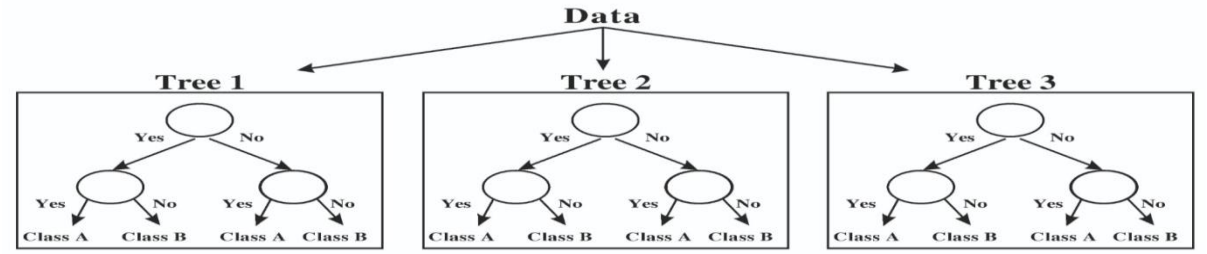

Fig. 09 RF with three different decision trees [23]

Figure 2 illustrates how the three separate decision trees that comprise the RFs identify the data as either class "A" or class "B" by utilizing various input vectors that are drawn from the same subset of the dataset. Compared to conventional classification algorithms, RFs can achieve excellent classification accuracy because their primary method is voting for the class [21]. Given that the imbalance is not too extreme, RFs are also adept at handling class imbalance. Gene selection for investigations on gene expression is one medical application for which RFs are utilized.

2) Applicability prospects of the classifier with and without feature selection

RF was employed on the dataset taking all the attribute in training and testing with Target attribute as dependent one by splitting the dataset with 80:20. Classification with 100 numbers of tree carried out with the base dataset bearing all 28 attributes and produced an accuracy of 91.16% with 83.33% sensitivity and 94.95% specificity.

3) Application of RF on the dataset with FS

The same classifier applied on the same dataset using 80% data for training and 20% for testing. Again Target, as a dependent variable. Classification with 100 numbers of tree carried out with the reduced dataset bearing 14 featured attributes produced after FS and produced an accuracy of 94.56% with 89.58% sensitivity and 96.96% specificity.

B. Performance Metrics

The receiver operating characteristic (ROC) and area under curve (AUC) curve assess how well a classifier model performs at different threshold settings. It is also a visual representation that shows how good the model is at telling things apart. A higher value of an AUC means that model is better and intelligently recognizing '0s' as '0s' and '1s' as '1s', similar to how it is better at identifying patients with a disease from those with no disease [24]. The ROC curve is a graphical representation that shows how good a test is at finding things. It is supposed to find in an ML model (TPR) and how many times it finds things, it shouldn't be (FPR).

Fig.5 and Fig.6 show the Receiver Operating Characteristics (ROC) result for the components that were identified and extracted using the SVM classifier. The ROC curve is a tool used to analyze the performance of classifiers, while the confusion matrix is a measure applied to assess or judge the quality of the ROC curve by examining the area under the curve (AUC).

C. About R Studio Environment

R is a programming language and environment that is primarily used for statistical computing, data analysis, and data visualization. R has become a popular choice among statisticians, data scientists, researchers, and analysts for its powerful capabilities in handling data and conducting statistical analyses. Since it has a pool of useful statistical packages, which are used to manipulate and get outcomes after employing ML techniques. Some statistical packages like pROC(),ROCR(), and stats() applied to perform and analyze the ML model.

D. Accuracy Analysis

The following analysis is done based on the attributes of the dataset taken as 673 in training set and 147 in test set. For the same confusion matrix was evaluated to get the outcome from the proposed ML model. A confusion matrix is an evaluating tool that is used to assess how well a classification system works by showing how often it gets things wrong or right [25]. The accuracy of ML algorithms depends on four key components such as true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) generated by the confusion matrix (Table-2) of an ML model.

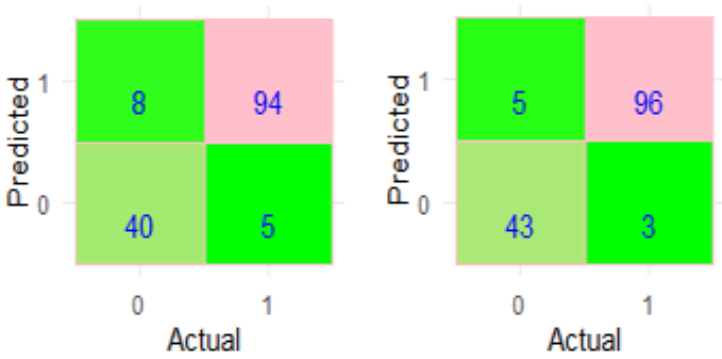

Table-2. Confusion matrix

The below table shows the performances of these models with and without FS technique, Result1 represents without FS and Reults2 with FS.

Metrics	Derivations	Calculation1	Result1	Calculation2	Result2
Accuracy	(TP + TN) / (P + N)	(94 + 40) / (45 + 102)	91.16%	(96 + 43) / (102 + 45)	94.56%
Precision	TP / (TP + FP)	94 / (94 + 8)	92.16%	96 / (96 + 8)	92.30%
Specificity	TP / (TP + FN)	94 / (94 + 5)	94.94%	96 / (96 + 3)	96.96%
Sensitivity	TN / (FP + TN)	40 / (8 + 40)	83.33%	43 / (5 + 43)	89.58%
F1 Score	2TP / (2TP + FP + FN)	2 * 94 / (2 * 94 + 8 + 5)	93.53%	2 * 96 / (2 * 96 + 8 + 5)	93.65%
Error Rate	1 - Accuracy	1 - 0.9116	8.84%	1 - 0.9455	5.44%

a

Table-3. Different measurements of performance

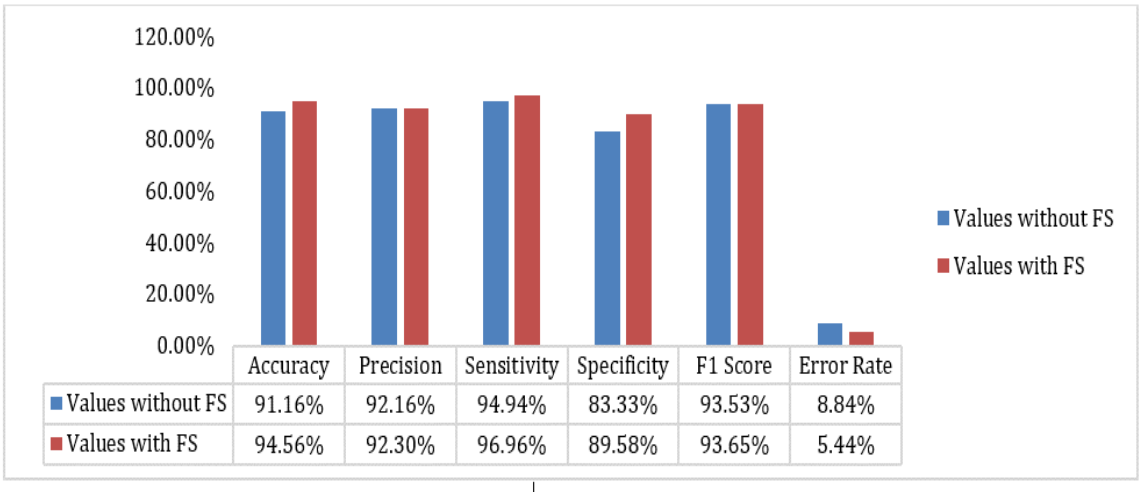

Fig.10 Model's performance metrics

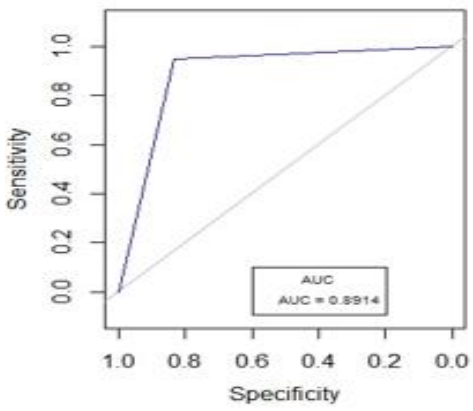

Fig. 11. ROC curve without feature selection

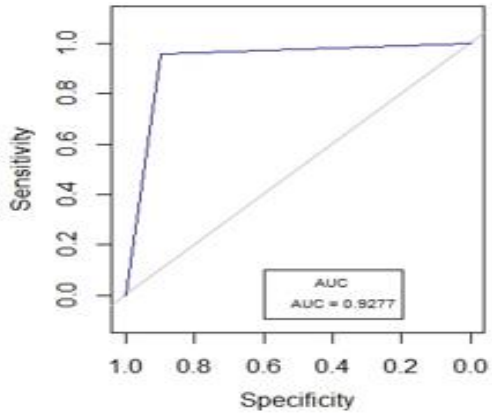

Fig. 12. ROC curve without feature selection

By observing the above figures it may be analyzed that the model is doing better because the curves are above the line. The AUC value is 89.14% found without FS and 92.77% after FS. The model showed only 3.63% difference of AUC value before FS so the curve of Fig.6 is little bit away from the boundary compared to Fig. 5.

Where these features of a confusion matrix denote as TP (True Positives):-The number of people identified correctly with heart disease. TN (True Negatives):- The number of people correctly identified as without heart disease. FP (False Positives):-The number of people incorrectly identified as having heart disease when they don't. FN (False Negatives):-The number of people incorrectly identified as not having heart disease in actuality.

V. REUSLT

After using RF, a machine learning approach with and without FS on the dataset by splitting it in 80:20 ratio of training and testing we found that accuracy with FS is better than without. Accuracy was calculated applying a confusion matrix of both models shown in Table.02 it is concluded that RF with FS showed an accuracy of 94.56% shown in Table.02 as proposed.

Author	Year	Algorithms Used	Best Algorithm	Highest Accuracy
[16]	2022	Random Forest (RF), Artificial Neural Network (ANN)	RF 3.3% <ANN	64.60%
[13]	2023	Logistic Regression (LR), Random Forest (RF)	LR	80.00%
[17]	2023	Random Forest (RF)	RF	84.03%
[12]	2023	SVM, CNN, RF, CART, C-RF	C-RF	85.00%
[14]	2023	Support Vector Machine (SVM)	SVM	85.00%
[8]	2023	Decision Tree (DT), Random Forest (RF), LR, NB, SVM	RF	85.01%
[11]	2021	DT, DA, LR, NB, SVM, KNN, Ensemble classifiers	LR (with PCA)	85.80%
[15]	2023	AutoML (PyCaret, AutoGluon, AutoKeras)	AutoGluon	86%
[9]	2023	Classification and Regression Tree (CART)	CART	87.00%
[10]	2023	Decision Tree (DT), XGBoost, Random Forest (RF), MLP	RF (with CV)	87.05%
[6]	2023	Logistic Regression (LR), Random Forest (RF)	RF	87.64%
[18]	2023	Random Forest (RF), Regression models, KNN, Naïve Bayes	RF	88.52%
[7]	2023	Random Forest (RF), K Nearest Neighbor (KNN)	RF	90.16%
Proposed	**2023**	**RF with FS**	**RF**	**94.56%**

Table-03 comparative performances of existing and proposed ML models

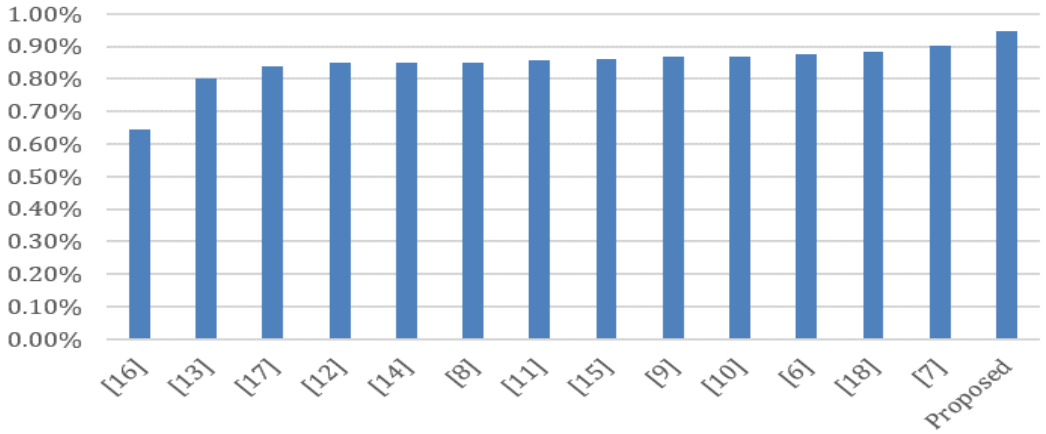

Fig. 08 Graphical representation of the existing and proposed ML models

VI. SUMMARY AND DISCUSSION

This study leverages machine learning (ML) techniques, Random Forest with and without FS, to predict cardiovascular disease (CVD) using patient data taken from Rohilkhand Hospital for research purposes. Here an ML model, RF with FS produces better accuracy compare to RF alone, achieving a higher accuracy rate 94.56% with FS and v/s. 91.16% without FS. The research emphasizes the importance of early CVD detection and the role of machine learning in healthcare analytics.

VII. VII CONCLUSION AND FUTURE AVENUES

This research paper proposes an enhanced accurate ML model for CVD prediction. The CVD dataset used in this work evaluated based on vital features and developed an ML model. In order to improve the performance of the suggested model, Boruta was used for feature selection. After evaluating the proposed model and its performance keeping view in mind for accuracy, precision, recall, F-measure, ROC, and AUC scores. All the way, the ensemble models and other researched models were surpassed by the proposed RF ML model with feature selection. The proposed model shown a considerable enhancement of accuracy of 94.56% with test data.

We tested the model both with and without feature selection (FS) (Fig. 1) using a real-time dataset, and discovered that FS in conjunction with Random Forest (RF) produced good results.

Our long-term objective is to improve the model's accuracy by using Artificial Neural Network (ANN) approaches on a larger dataset, which will include an image datasets. As a result, we will eventually be able to save more lives by improving our ability to predict heart illnesses early. Additionally, it will give the stakeholders a greater understanding of their health issues.

Data Availability

Data will be made available on demand

ACKNOWLEDGEMENT

This study did not receive any grant from any agencies

REFERENCES

[1] S. Mishra, P. K. Mallick, H. K. Tripathy, A. K. Bhoi, and A. González-Briones, “Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier,” *Appl. Sci.*, vol. 10, no. 22, pp. 1–35, 2020, doi: 10.3390/app10228137.

[2] D. Chicco and G. Jurman, “Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone,” *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 1–16, 2020, doi: 10.1186/s12911-020-1023-5.

[3] W. news Room, “https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).” [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[4] R. Rastogi and M. Bansal, “Measurement : Sensors,” vol. 25, no. December 2022, 2023.

[5] M. G. El-Shafiey, A. Hagag, E. S. A. El-Dahshan, and M. A. Ismail, "A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest," *Multimed. Tools Appl.*, vol. 81, no. 13, pp. 18155–18179, 2022, doi: 10.1007/s11042-022-12425-x.

[6] R. B. and R. S. G. S. Reddy Thummala, "Prediction of Heart Disease using Random Forest in Comparison with Logistic Regression to Measure Accuracy," in *IEEE Access*, Chennai: IEEE, 2023. doi: 10.1109/ACCAI58221.2023.10199851.

[7] T. Poojitha and R. Mahaveerakannan, "Prediction Analysis of Novel Random Forest Algorithm and K Nearest Neighbor Algorithm in Heart Disease Prediction with an Improved Accuracy Rate," *Cardiometry*, no. 25, pp. 1554–1561, 2023, doi: 10.18137/cardiometry.2022.25.15541561.

[8] A. Khan, M. Qureshi, M. Daniyal, and K. Tawiah, "A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction," *Health Soc. Care Community*, vol. 2023, no. Cvd, pp. 1–10, 2023, doi: 10.1155/2023/1406060.

[9] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthc. Anal.*, vol. 3, no. November 2022, p. 100130, 2023, doi: 10.1016/j.health.2022.100130.

[10] T. G. 1 and P. L. M. 2 Chintan M. Bhatt 1,* , Parth Patel 1, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, 2023, doi: https://doi.org/10.3390/a16020088.

[11] K. V. V. R. I. E. A. A. A. S. Paramasiva, "Heart Disease Risk Prediction using Machine Learning with Principal Component Analysis," in *2020 8th International Conference on Intelligent and Advanced Systems (ICIAS)*, Kuching, Malaysia: IEEE, 2021. doi: 10.1109/ICIAS49414.2021.9642676.

[12] M. et al. Wang, J., Rao, C., Goh, "Risk assessment of coronary heart disease based on cloud-random forest," *Artif. Intell. Rev.*, vol. 56, pp. 203–232, 2023, doi: https://doi.org/10.1007/s10462-022-10170-z.

[13] A. Y. S. K. D. Upadhyay, "A Statistical Analysis for Heart Disease Prediction System for Next-Gen Software," in *2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*, IEEE, 2023, pp. 71–76. doi: 10.1109/CICTN57981.2023.10140436.

[14] E. A. O. & W. B. Yahya, "Performance analysis of supervised classification models on heart disease prediction," *Innov. Syst. Softw. Eng.*, vol. 19, pp. 129–144, 2023, doi: https://doi.org/10.1007/s11334-022-00524-9.

[15] L. M. Paladino, A. Hughes, A. Perera, O. Topsakal, and T. C. Akinci, "Evaluating the Performance of Automated Machine Learning (AutoML) Tools for Heart Disease Diagnosis and Prediction," *AI*, vol. 4, no. 4, pp. 1036–1058, 2023, doi: 10.3390/ai4040053.

[16] W. A. W. A. Bakar, N. L. N. B. Josdi, M. B. Man, and Y. S. Triana, "An Evaluation of Artificial Neural Networks and Random Forests for Heart Disease Prediction," *J. Hunan Univ. Nat. Sci.*, vol. 49, no. 2, pp. 41–49, 2022, doi: 10.55463/issn.1674-2974.49.2.4.

[17] S. D. and N. S. T. Mahmud, A. Barua, M. Begum, E. Chakma, "An Improved Framework for Reliable Cardiovascular Disease Prediction Using Hybrid Ensemble Learning," in *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Chittagong, Bangladesh: IEEE, 2023. doi: 10.1109/ECCE57851.2023.10101564.

[18] A. A. Stonier, R. K. Gorantla, and K. Manoj, "Cardiac disease risk prediction using machine learning algorithms," *Healthc. Technol. Lett.*, 2023, doi: 10.1049/htl2.12053.

[19] G. E. Newman and B. J. Scholl, "Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias," *Psychon. Bull. Rev.*, vol. 19, no. 4, pp. 601–607, 2012, doi: 10.3758/s13423-012-0247-5.

[20] S. Fahimifar, K. Mousavi, F. Mozaffari, and M. Ausloos, "of highly cited scholarly papers through 3 (i . e ., Ridge , Lasso ," vol. 3, pp. 3685–3712, 2023.

[21] M. Zhu *et al.*, "Class weights random forest algorithm for processing class imbalanced medical data," *IEEE Access*, vol. 6, pp. 4641–4652, 2018, doi: 10.1109/ACCESS.2018.2789428.

[22] A. Liaw and M. Wiener, "The R Journal: Classification and regression by randomForest," *R J.*, vol. 2, no. 3, pp. 18–22, 2002, [Online]. Available: http://www.stat.berkeley.edu/

[23] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–16, 2019, doi: 10.1186/s12911-019-1004-8.

[24] F. S. Nahm, "Receiver operating characteristic curve: overview and practical use for clinicians," *Korean J. Anesthesiol.*, vol. 75, no. 1, pp. 25–36, 2022, doi: 10.4097/kja.21209.

[25] M. Heydarian and T. E. Doyle, "MLCM : Multi-Label Confusion Matrix," pp. 19083–19095, 2022.