Predicting Heart Diseases Using Enhanced Machine Learning Techniques

Vidhya Ashok[1], Mohamed Shameem[2], H Shaheen[3], Brumancia Easpin[4]

[1]Assistant Professor, School of Engineering and IT, Manipal Academy of Higher Education, Dubai
[2]Senior Lecturer & Course Leader, department of Artificial Intelligence, University of West London – RAK Campus, United Arab Emirates
[3]Lecturer, Department of Computing and Engineering, University of West London, RAK Branch Campus, UAE
[4]Lecturer, Department of Computing and Engineering, University of West London, RAK Branch Campus, UAE

Received: 09.04.2024 Revised : 18.05.2024 Accepted: 27.05.2024

ABSTRACT
According to recent surveys, cardiovascular disease is a leading global cause of death, resulting in a significant number of fatalities each year. In the field of medicine, data mining is gaining increasing recognition and significance. The medical industry generates a massiveand complex volume of data, making it challenging to handle and analyze using conventional approaches. Consequently, data mining emerges as a critical component to address this challenge. However, the accuracy of predictions is often questioned due to the high rate ofinaccuracy in some forecast computations. Therefore, selecting a prediction approach thatyields higher accuracy with fewer errors becomes crucial in this context. The objective of thisstudy is to develop a dependable system for forecasting heart illness. It is evident that the random forest algorithm consistently outperforms other methods in terms of precision. To further enhance the accuracy of the outcomes, the random forest method is subjected to hyper parameter tuning.

Keywords: Data mining, Heart Disease, Random forest, Accuracy, Hyper parameter tuning.

INTRODUCTION

Anomalies in the heart can lead to a variety of cardiac disorders. Among the leading causes of death globally, cardiovascular disease claims a significant number of lives each year. Heart conditions encompass peripheral artery disease, coronary heart disease, congenital heart disease, and cerebra vascular illness. Many of these ailments can be prevented by addressing specific behavioral risk factors, such as an unhealthy diet, smoking, obesity, lack of physical activity, and excessive alcohol in take. Establishing a system that enables precise early detection of cardiac problems is crucial to implementing timely interventions and preventing the progression of the disease.

Currently, the primary challenge facing the healthcare industry revolves around facility advantage. The quality of service hinges on accurate diagnosis and the administration of effective therapy to patients. Deviations from correct diagnoses can lead to extremely di-reconsequences. Although medical history records or data are extensive, they originate from various sources, and their interpretation by physicians is essential to complete the picture.

Duetoreal-world data being inherently noisy, partial, and inconsistent, pre-processing is necessary when incorporating missing values into the database. Cardiovascular disease waspreviously believed to be the world's leading cause of death, with the assumption that it wasalso the most preventable and manageable condition. Ultimately, the prompt and accurate diagnosis of a nailment plays a crucial role in its successful treatment.

Heartdiseaseis the prevailing chronic ailment globally, yet it is also highly avoidable. Adopting a healthy lifestyle (primary prevention) and timely analysis (secondary prevention) stand as the two main factors in preventing heart disease. Crucially, regular health check-ups(tertiary prevention) play a vital role in identifying and preventing heart disease at an early stage. Various tests, such as angiography, chest x-rays, echocardiography, and exercise tolerance tests, are available to assist in diagnosing this crucial condition. However, these examinations can be expensive and require the use of specialized medical equipment.

The primary objective of this research is to leverage machine learning algorithms in constructing a model for cardiac disease prediction. This model aims to enable clinicians to diagnose the condition earlier, requiring fewer medical tests, and facilitating the provision ofappropriate therapy, potentially saving

numerous lives. In hospitals, where vast amounts of patient data related to heart illnesses and other diseases are generated daily, efficient utilization and management of this data for decision-making can be challenging for doctors inthe absenceofdataminingtools.
Data mining is highly recommended for predicting cardiac problems because it allows for theextraction of more accurate and valuable insights from massive datasets, thereby simplifying the prediction process. It forms the fundamental basis of machine learning by assisting in the handling of large amounts of data with remarkable speed and enabling earlypredictions.

LITERATURE SURVEY

Ref No	Author	Technique	Dataset	Accuracy(%)	Limitations	Year
[1]	Pabitra Kumar Bhunia, ArijitDebnath, PoulamiMondal, Monalisa D E, Kankana Ganguly, Pranati Rakshit.	Machine learning algorithms with few features.	Heart.csv	90.32%	Comparative analysis	2021
[2]	Rao Patrof	Various Machine learning algorithms.	Various tools and database	Random forest with best result	Comparative analysis	2021
[3]	JayshrilS.Sonawane	Multilayer perception neural network	Cleveland Heart disease	80%	For boolean functions, the learning time of multilayer networks with back propagation scales exponentially.	2014
[4]	Rifkiwijaya	Machine learning ANN	Various tool and databases	81%	The neural network must betrained before it can function. Large neural networks requires long processing time.	2013
[5]	M.A.Jabbar	Feature subsets selection	UCI repository	80%	It takes too much parameters from the patient's record and generates rules that are irrelevant. This technique is computationally expensive and has poor performance.	2017
[6]	Jyotisoni	Weighted Associated classifiers	UCI machine learning datasets	81.51%	In the prediction model, not all attributes are equally important in predicting accuracy.	2011
[7]	SenthilKumar	Different combination of feature	Hospital databases	88.7%	It is difficult to be adopted by less experienced people.	2016
[8]	Avinash Golande	Parameters tuning	Hospitals database	comparative analysis	The accuracy of the structure can be improved further by combining various data mining techniques.	2019

[9]	Sellappan Palaniappan	Data mining algorithms or parameter technique	Heart.csv	web-based, user-friendly, mountable, dependable, stretchable, and justified	The algorithm's accuracy should be increased further.	2019
[10]	m. Anbarasi	Features subsets selections using genetical algorithm	Hospital database	70%	The language used to define a potential solution must be resilient. A poor fitness function generates major problems.	2017

METHODOLOGY

The entire machine learning approach is described here, including how the input datasets is entered in to the software and how it will be processed in both text and block diagram forms.

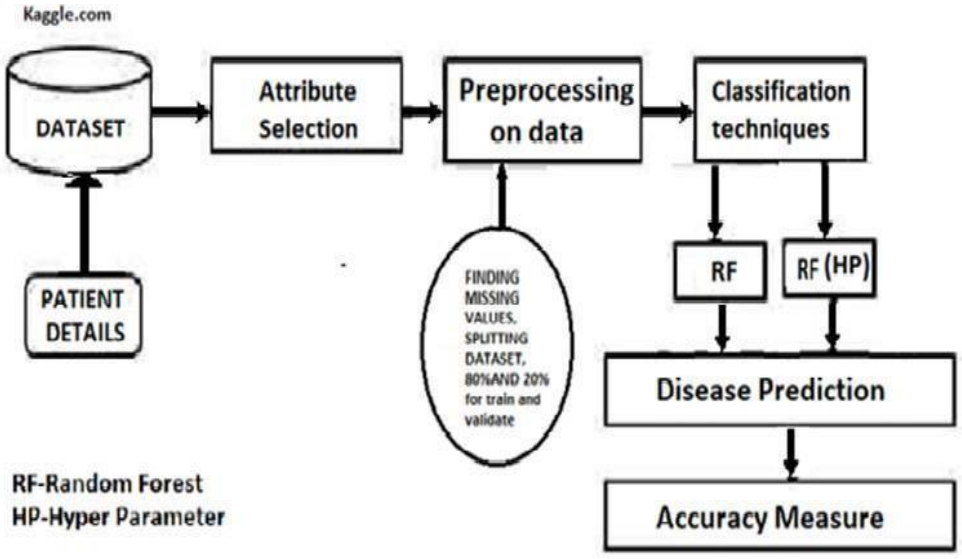


Fig 1: Initial System Design for Heart Disease Prediction

The "Heart.csv" data set was taken from the Kaggle website. It is comprises of 303 records and a subset of 14 characteristics:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

Fig 2. Pre-processed Dataset

Random Forest Classifier

Random Forest is a classifier that combines certain selection timber on many subsets of a given dataset and uses the common to improve the predicted accuracy of that dataset. Following algorithm illustrates random forest method. Random Forest operates in two stages: the first is to generate the random forest by mixing N decision trees, and these condition is to make predictions for each tree generated in the first phase. The following stages will walk you through the working process:

Step-1:Choose K data points at random from the training set.

Step-2: Create decision trees based on the data points you've chosen (Subsets).

Step-3:Choose the number N for the number of decision trees you wish to create.

Step-4:Repeat Step1&2.

Step-5: Find the forecasts of each decision tree for new data points and allocate the new data points to the category that receives the most votes.

Hyper parameter tuning techniques

The easiestapproach to consider hyper parameters is to think of them as the settings of a setof rules that may be modified to maximize performance. While version parameters, like as theslope and intercept in a linear regression, are discovered during training, hyper parameter should be established by the facts scientist before to training.

In the case of a random algorithm, hyper parameters include the vast range of selecting bushes inside the wooded area as well as the large range of functions considered by each treewhen splitting a node. (The variables and thresholds used to break apart every node detected at sometime during training are the characteristics of a random wood land area.)

Scikit-Learn provides a set of practical default hyper parameters for every models, although these aren't guaranteed to be the most important for a situation. The optimal hyper parametersare frequently impossible to predict ahead of time, and tweaking a model is the point at which device learning transitions from a science to trial-and-error absolutely engineering.

Few of the hyper parameters are shown below

- n_estimators= the total number of trees in the forest
- max_features=maximum amount of characteristics examined while breaking anode
- max_depth=max number of levels in each decision tree
- min_samples_split=The minimum amount of data points that can be inserted in anode before it is divided.
- min_samples_leaf=a leaf node can only have a certain number of data points.

Proposed System

Our suggested technique focuses on unique machine learning processes for heart disease categorization and prediction, hence addressing the current difficulty. We will build our model using Random Forest to improve performance and accuracy.

For the prediction of cardiac disease, the suggested approach employs the random forestalgorithm and hyper parameter tweaking techniques.

Proposed Algorithm

Step 1: Load the data set on heart disease.

Step2: Data pre-processing and exploration

We have to pre-process the data before we can train a model

- First, we clean up the data by replacing missing values with the mean.
- Second, we'll remove several columns to simplify the model. The third step is to divide the data.
- Using an 80/20 split ratio, we will divide the dataset into training data (xtrain, ytrain) and test data (xtest, ytest).
- There are 242 rows and 14 columns in the train data set.
- There are 61 rows and 14 columns in the test data.
- The dependent variable is the variable target.
- In train and test data, the goal variable is different .We'll have to match the mup.

Step -4: Create a Random Forest Model and put it into action. Using the x test dataset, we train our training dataset model and generate a test prediction. The trained model's performance is then shown in a confusion matrix.

Step5: Using the Random Forest Model with Hyper Parameter Tuning. The range of parameters in our model will be defined as follows:

- max features - This is the maximum amount of features that Random Forest can try in each tree.
- Auto: This will just take all of the characteristics in each tree that make sense.
- sqrt: Takes the square root of the total number of features in each individual run.
- n estimators - Choose the highest value your processor can handle to make your forecasts stronger and more accurate.

Step 6: We must specify the measure the grid search algorithm will use to evaluate the model's performance. The grid search technique is then used.

Step 7: Use the grid search approach and start with the whole feature set, removing low-valuefeatures one by one. Only one feature is eliminated in each iteration, which has the biggest impact on total model accuracy, as long as the accuracy does not increase. Tuning is a technique for identifying high-priority characteristics.

Step 8: On the remaining characteristics in the data set, use the adjusted Random Forest method to optimize classification accuracy.

Step9: We view the findings in a new confusion matrix and calculate the classifier's accuracy.

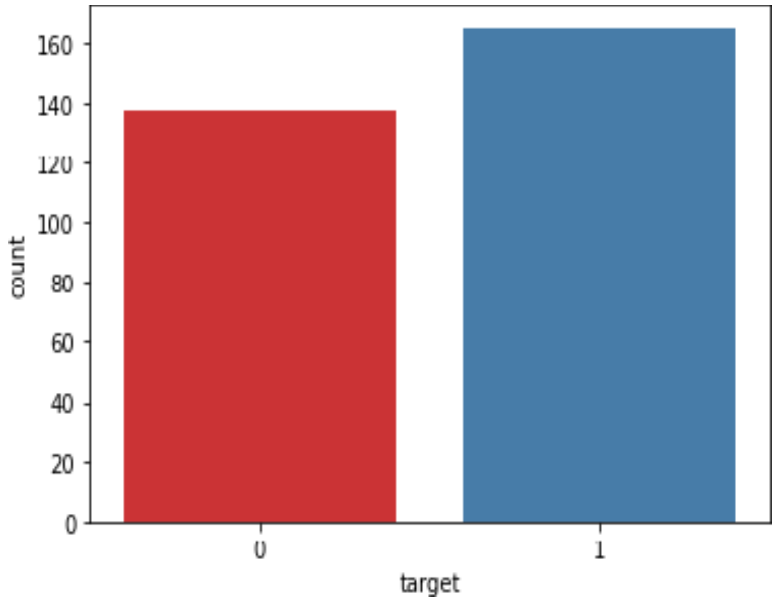


Fig 4. bar chart representing the number of patients with a heart disease and without

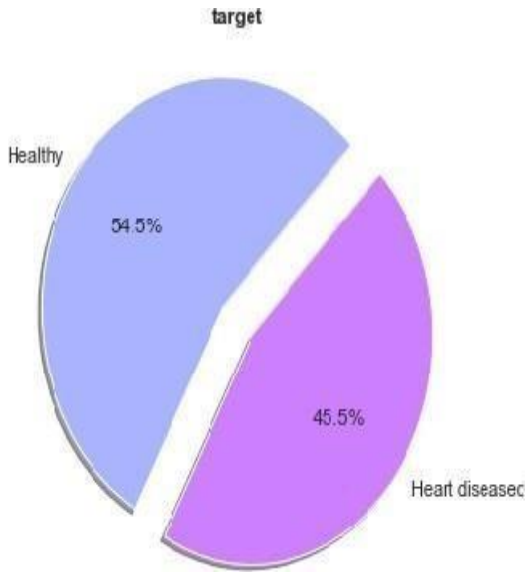


Fig 5. Pie chart representing number of patients with a heart disease and without

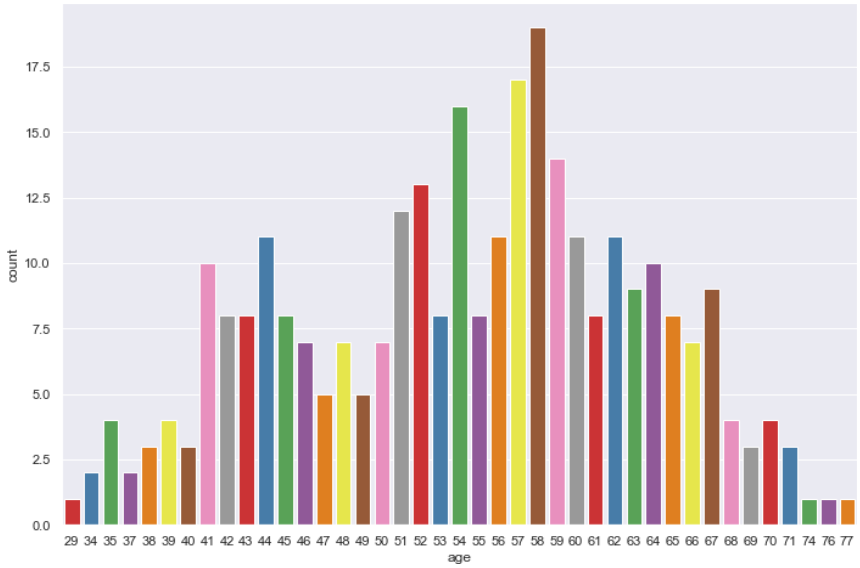


Fig 6. Number of people having various age group

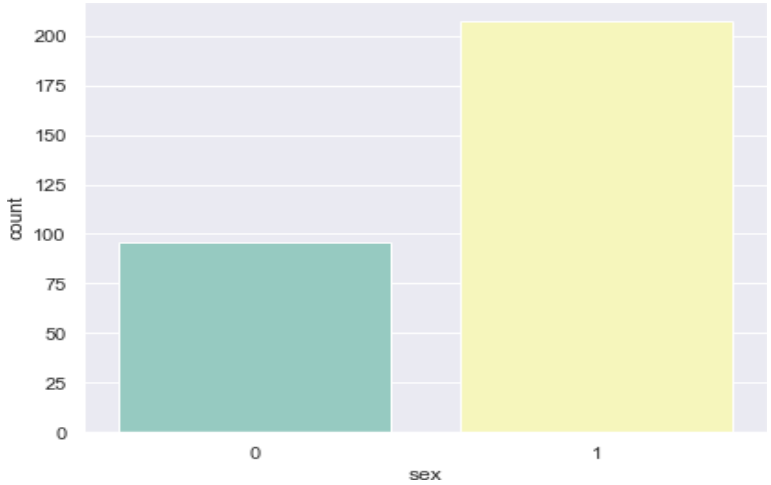
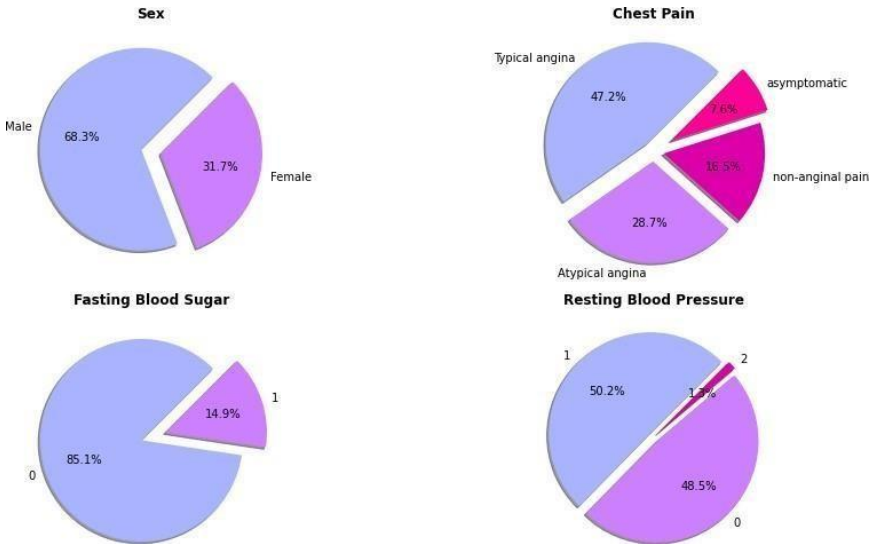


Fig 7. Population of Males Vs Females in the dataset



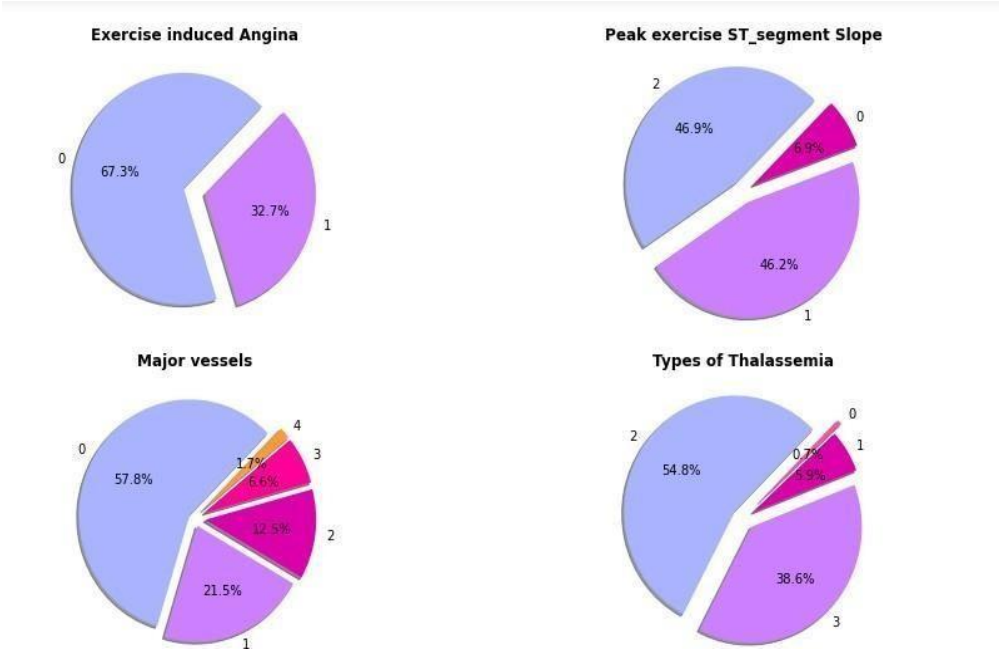


Fig 8. Pie chart representing each column of the dataset

Apply Random Forest Classifier on the data set to get the accuracy

```
max_accuracy = 0


for x in range(50):
    rf = RandomForestClassifier(random_state=x)
    rf.fit(X_train,Y_train)
    Y_pred_rf = rf.predict(X_test)
    current_accuracy = round(accuracy_score(Y_pred_rf,Y_test)*60,2)
    if(current_accuracy>max_accuracy):
        max_accuracy = current_accuracy
        best_x = x

print(max_accuracy)
print(best_x)
```

Fig 9. Random Forest Classifier Code Here is the confusion matrix obtained for Random Forest

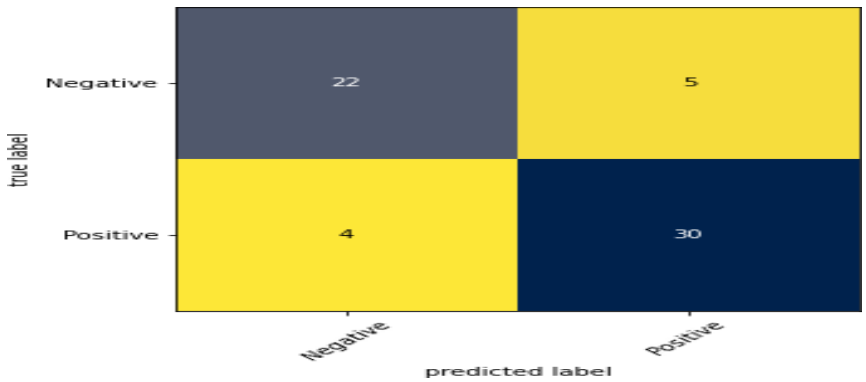


Fig 10. Confusion matrix for Random Forest

A confusion matrix is at able that shows how an algorithm performs. Two rows and two columns (for two class issues) indicate TP, FP, TN, and FN in the confusion matrix.
The confusion matrix isused to evaluate the number of correct and wrong predictions provided by the model with the actual categorization of the heart disease dataset.

Prediction		Disease	
		+	-
	+	True positive	False positive
		TP	FP
	-	False Negative	True Negative
		FN	TN

Accuracy=(TP+TN)/(TP+FP+TN+FN)
Where
TP => Positive tuples that the classifier properly labels.TN=>Negative tuples that the classifier properly labels.
FN => Positive tuples that the classifier has mistakenly labelled.
FP=>Negative tuples that the classifier has improperly labelled.
Here is the ROC Curve obtained from Random Forest Algorithm

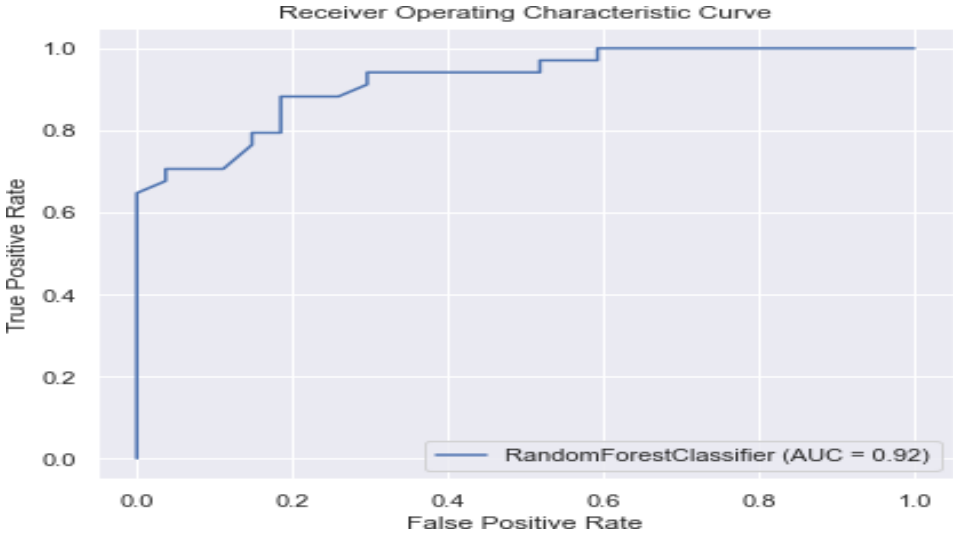


Fig 11. ROC Curve showing accuracy obtained by using Random Forest Classifier Accuracy (RF)=92%.

Apply Random Forest and Hyper Parameter tuning techniques

```
RFclf = RandomForestClassifier(max_depth=1000, n_estimators=2000)
print(RFclf)
#RFclf.fit(X_train,Y_train)GridSearchCVRandomizedSearchCV
grid = GridSearchCV(estimator=RFclf, param_grid=param_grid, cv = 8, n_jobs=2,verbose=2)
grid_result = grid.fit(X_train, Y_train)
# Summarize results
print("Best: %f using %s" % (grid_result.best_score_, grid_result.best_params_))
```

Fig 12. Random Forest+ Hyper parameter code

Here is the confusion matrix incase of Random Forest+ Hyper parameter tuning

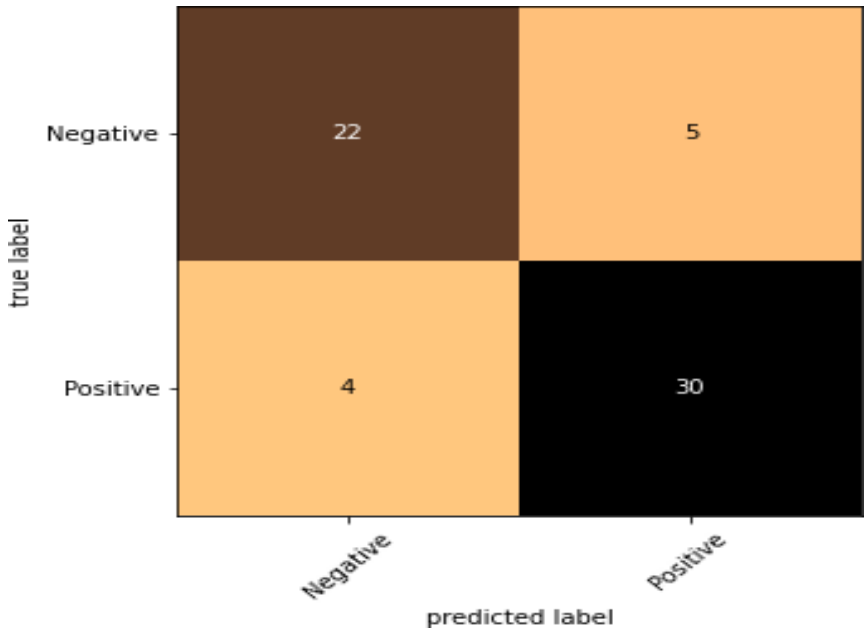


Fig 13. Confusion Matrix for Random Forest Classifier and Hyper Parameter

	precision	recall	f1-score	support
0	0.85	0.81	0.83	27
1	0.86	0.88	0.87	34
accuracy			0.85	61
macro avg	0.85	0.85	0.85	61
weighted avg	0.85	0.85	0.85	61

Fig 14. Accuracy table

Calculation of the Accuracy table

- Precision - Precision indicates how precise/accurate your model is in terms of how many of the anticipated positives are actually positive.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$= \frac{True\ Positive}{Total\ Predicted\ Positive}$$

- Recall –It determines how many Actual Positives our model catches by classifying it as Positive (TruePositive).

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

- F1 Score–It is required when you strive to seek balance between precision or recall.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

- Accuracy
  Accuracy=(TP+TN)/(TP+FP+TN+FN)
  =(22+30)/(22+5+30+4)

Where
TP => Positive tuples that the classifier properly labels.
TN=> Negative tuples that the classifier properly labels.
FN => Positive tuples that the classifier has mistakenly labelled.
FP=>Negative tuples that the classifier has improperly labelled.
The random forest technique provides us the highest accuracy when hyper parameters liken estimators, max features, and max depth are optimized. The ROC graph below depicts the decrease in accuracy with all of these features:-

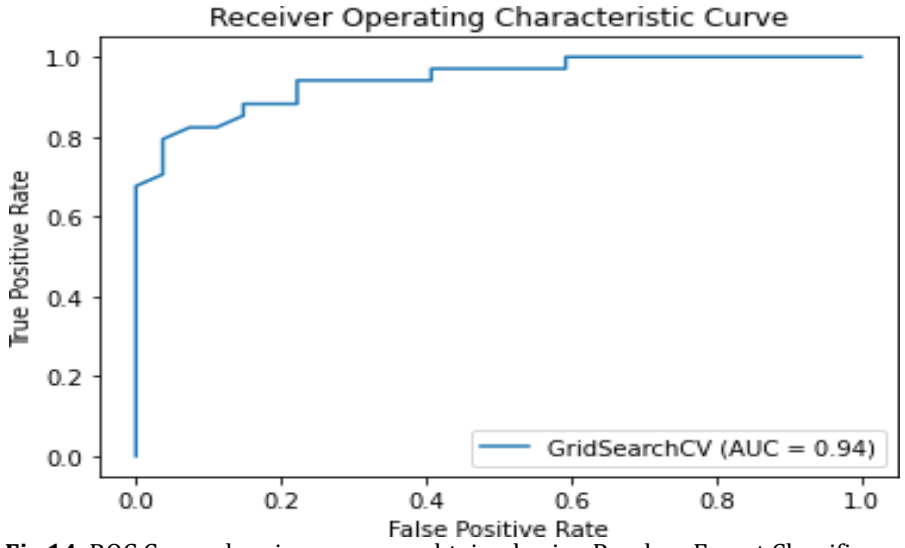


Fig 14. ROC Curve showing accuracy obtained using Random Forest Classifier and HP

Accuracy=94%

Experimental Results

Table 2: Accuracy comparison

Approach	Accuracy
Random Forest Classifier(Existing approach)	92%
Random Forest Classifier with Hyper Parameter tuning (Proposed approach)	94% (increased by2%)

The accuracy gained with RF and HP is shown in Table 2. The RF+HP model improvesaccuracy by 2% over the RF model without HP. As a result, it is obvious from the table that our technique out performs the existing Random Forest Classifier model.

CONCLUSION

In this research study, we developed an effective methodology for forecasting heart diseaseusing a combination of random forest and hyperparameter techniques. Data mining played a crucial role in the heart disease prediction process. We employed feature selection and modified measurements to categorize cardiac llnesses. Our suggested strategy, which involved utilizing random forest and hyper parameter fitting, achieved an impressive accuracy of 94 percent when applied to the heart disease dataset. The incorporation of both random forest and hyper parameter approaches significantly enhanced the accuracy of cardiac disease forecasting. Notably, we demonstrated that fine-tuning the hyper parameters leads to improved classification accuracy compared to standard techniques. Our approach highlights the importance of using a combination of tuning methods and classification algorithms to achieve accurate cardiac disease forecasts. Additionally, to further enhance the prediction model's

accuracy, it is beneficial to explore various data mining approaches and features election algorithms in tandem.

REFERENCES

[1] Himans hu Sharma and M A Rizvi, "Prediction of Heart Disease Using Machine Learning Algorithms: A Survey, "International Journal on Recent and Innovation Trends in Computing and Communication, vol.5, no.8, Himan shuSharma and MARizvi," Prediction of Heart Disease Using Machine Learning Algorithms: A Survey," International Journal on Recent and Innovation Trends in Computing and Communication, vol.5,no.8,

[2] Himan shuSharma and MARizvi|Y

[3] Pavan Kumar T and Avinash Golande," Heart Disease Prediction Using Effective Machine Learning Techniques, "International Journal of Recent Technology and Engineering, vol.8, no. 4, pp.944-950, 2019.

[4] Jayshril S. Sonawane and D.R Patil, "Prediction Of Heart Disease Using Multilayer Perceptron Neural Network, "IEEE International Conference on Information Communication and Embedded Systems, 2014.

[5] Rifki Wijaya, Ary Setijadi Prihatmanto, and Kuspriyanto, "Preliminary design of estimating heart disease with in one year using machine learning ANN,"IEEEJoint.

[6] Jyoti Soni, Uzma Ansari, Dipesh Sharma, and Sunita Soni, "Intelligent and EffectiveHeartDisease Prediction System Using Weighted Associative Classifiers," International Journal of Computer Science and Engineering ,June2011. (IJCSE).

[7] Krishnan J Santhana and SGeetha, "Prediction of Heart Disease Using Machine Learning Algorithms, "ICIICT| Year:2019| Conference Paper|Publisher:IEEE,ICIICT

[8] |Year:2019| Conference Paper | Publisher: IEEE, ICIICT, ICIICT, ICIICT, ICIICT, ICIICT,ICIICT,ICIICT,ICIICT

[9] "Heart Disease Prediction Using Machine Learning,"International Journal of Engineering Research and Technology, Pabitra Kumar Bhunia, Arijit Debnath, Poulami Mondal, Monalisa DE, Kankana Ganguly, and PranatiRakshit2021.(IJERT).

[10] "Heart Disease Prediction Using Machine Learning,"International Journal of Engineering Research & Technology, 2021. ApurbRajdhan, Milan Sai, and Avi Agarwal,"Heart Disease Prediction Using Machine Learning," International Journal of EngineeringResearch&Technology,2021.(IJERT).