JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

STUDY ON HEART DISEASE PREDICTION MODEL USING MACHINE LEARNING ALGORITHMS AND HYBRID MACHINE LEARNING MODEL.

Akshata Ayyalasomayajula
Department of Computer Science and Engineering
Gandhi Institute of Technology and Management
GITAM (Deemed to be University)
aayyalas@gitam.in

***Abstract*—Machine Learning is used in various fields these days, The health care industry is no exception. Heart Disease is one common type of deadly disease these days which people are prone to. Early detection of it would reduce the risk created by the disease. Although there is lot of data that is collected by the health care industry , the data collected isn't efficiently made use of due improper methods of discovering underlying patterns of present in the data and the decision making. In this project we'll be making a model wherein we'll be building a model that can be used in prediction of a potential heart disease. Here we use different Machine Learning algorithms for the prediction of a heart disease using 13 different attributes like Blood Pressure , Cholesterol , Resting Heart rate , Gender , Age etc. which tell the likelihood of a patient being effected by a heart disease.**

***Index Terms*—Machine Learning , Heart Disease Prediction , Classification Algorithm , Hybrid Machine learning Model, Accuracy.**

I INTRODUCTION

In the project we create a system which can detect the presence of the presence of Heart Disease using Machine Learning Algorithms. The algorithms include Naïve Bayes , Support Classifier Decision Tree Classifier , Random Forest classifier , Logistic Regression , K Nearest Neighbors Classifier , XG Boost Classification and Neural Network. The dataset has been taken Kaggle , UCI Repository. The dataset has 13 attributes namely Age , Blood Pressure , Sex , Cholesterol that are used in prediction of presence of a Heart Disease. In this method different readings have been taken not only by using one algorithm but by using two or more algorithms in detecting the likelihood of a heart disease. This model of using two or more algorithms together is commonly known as Hybrid Model. The data uses Sequential Classifier using Keras to build the Neural Network. The classifier uses 70% of the data as training data and 30% data for Classification. The introduced model is built using Machine Learning and Neural Network concepts also known as hybrid Machine Learning model.

Neural Networks are generally regarded as one the best methods for disease Prediction model like Brain Disease or Heart Disease Prediction. The model has showed an enhanced level of Performance when compared to existing models like the Data Mining models that used usually prediction of diseases.

We have also seen a lot of development in Healthcare Industry using Machine Learning and Internet of things. Machine Learning Algorithms on network traffic data have shown to provide accurate Identification when connected to IoT device.

In This model we Introduce a Hybrid Model using Random Forest using a Sequential Classifier. The main objective to build this model is to improve the performance of accuracy in Heart Disease Prediction.

II RELATED WORK

There has been ample of work done in this field related to the current model. ANN and HRFLM models have been introduced to produce high accuracy models. The data set used is collected form the UCI Laboratory which has collected from patients having various types of heart diseases. Data Mining models and Machine Learning Algorithms like Naïve Bayes , Support Vector Machine have been used and the accuracies generated by the these models have been compared to the Hybrid Model that is generated here. The model tends have a accuracy of 86.7% which is effective than the existing models and also can be implemented on larger data attribute datasets unlike the conventional existing models. There have been various research models that have been but the models built using Machine Learning models and Deep Learning models have been considered to be more effective methods the Prediction of heart diseases.

III PROPOSED MODEL

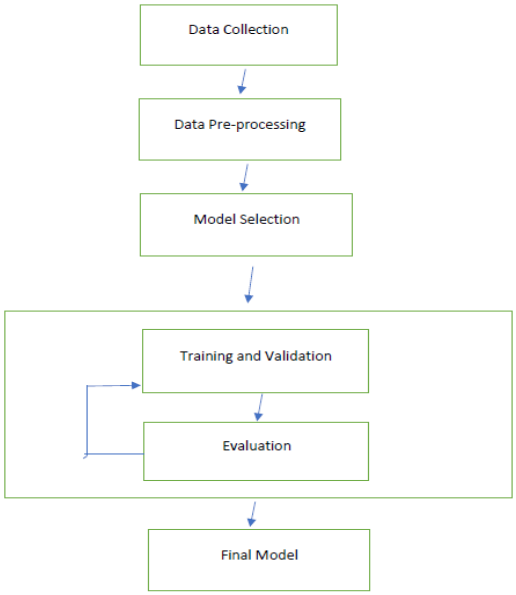

Fig : System Architecture

In this study we build the model using the Cleveland dataset , Using UCI Repository. The data set contains 303 patient records and 14 attributes. With 6 records with few missing values thus making it 296 records that are used in pre processing.

Attribute	Description
Age	Age of individual
Sex	• Gender Of individual(1 = male; 0 = female)
cp	• Chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
Trestbps	• Resting blood pressure (in mm Hg on admission to the hospital)
Chol	• Serum cholesterol in mg/dl

FBS	• Fasting blood sugar level > 120 mg/dl (1 = true; 0 = false)
Restecg	• Fasting blood sugar level > 120 mg/dl (1 = true; 0 = false)
Thalach	• Maximum heart rate achieved
Exang	• Exercise induced angina (1 = yes; 0 = no)
Oldpeak	• ST depression induced by exercise relative to rest
slope	• The slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
ca	• Number of major vessels (0-4) colored by flourosopy
Thal	• Thalassemia is an inherited blood disorder that affects the body's ability to produce hemoglobin and red blood cells. 1 = normal; 2 = fixed defect; 3 = reversable defect
Target	• the predicted attribute - diagnosis of heart disease (angiographic disease status) (Value 0 = < 50% diameter narrowing; Value 1 = > 50% diameter narrowing)

Table : UCI dataset types

The operation performed on the data set are divided as the following steps :

3.1 Data Collection : Data collection can be defined as the process of gathering and measuring information from countless different sources. This data can be numeric (temperature, loan amount, customer retention rate), categorical (gender, color, highest degree earned), or even free text (think doctor's notes or opinion surveys). Collecting data allows you to capture a record of past events so that we can use data analysis to find recurring patterns. From these patterns, we can build the predictive model.

3.2 Data Pre Processing : Data preprocessing refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning model. In our model we analyze the data and also check for any noise/redundancies present in the data. Here we also import all the essential libraries that are required for running of the model. Splitting of the dataset is also done in this stage into Train Set and Test Set.

3.3 Model Selection and Evaluation : Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset that solves the selected problem statement. It is a process that can be applied both across different types of models (e.g. logistic regression, SVM, KNN, etc.) and across models of the same type configured with different model hyperparameters.
The algorithms used in our model are :

3.3.1 Logistic regression : The model applies Logistic regression which is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false. It is represented by using the function –

1 / (1 + e^-value)

3.3.2 Support Vector Machine : The model applies linear SVM to find the optimal hyperplane where Data =(y(i) , x(i)) ; i=1,2,3,…n represents the training data and x(i) represents the target value. The hyperplane is generated by using the formula

$$F(x)=w^{\wedge}T+b$$

Where w is dimensional coefficient and b is the offset.

3.3.3 Random Forest : Random Forest is a ensemble learning model that builds a multitude of decision trees at training time. The output of the random forest is the class selected by most trees. the final feature importance at the Random Forest level, first the feature importance for each tree is normalized in relation to the tree:

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j}$$

Where,

- RFfi sub(i)= the importance of feature i calculated from all trees in the Random Forest model
- normfi sub(ij)= the normalized feature importance for i in tree j
- T = total number of trees

3.3.4 : Naïve Bayes : Naïve Bayes uses the Bayes rules through independent feature values. The model is trained with Gaussian Naïve Bayes Function with Prior Probability P(Xf)=priority (0,1)

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Where Sometimes assume variance

- is independent of Y (i.e., σi),
- or independent of Xi (i.e., σk)
- or both (i.e., σ)

3.3.5 :Decision Tree : Decision Tree classifier in this model is constructed based on high entropy inputs. The trees are constructed in a top down recursive divide and conquer approach. Pruning is performed to remove any noises or irrelevant values present the data.

3.3.6 XG Boost : XG Boost is a Gradient Boosted Decision Tree algorithm where trees are created in a sequential form. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results.

3.3.7 K Nearest Neighbour : It extracts knowledge based on the given input samples and Euclidean distance is calculated as d(xi , yi) and the majority of k nearest neighbours. Mathematically it is represented as

$$d(x,\ y)\ = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Where ,

x,y = two points in Euclidean n-space

Xi , yi = Euclidean vectors, starting from the origin of the space (initial point)

N = n-space

3.3.8 Neural Networks : The Neuron Components include inputs xi and hidden layers and the output layer. The final result is produced using activation filter like ReLu and a bias constant b.

3.4 Final Model : The final model is build using the Cleveland dataset which has 306 records. Although the dataset had 76 attributes we only use 14 attributes. The source of the dataset is Cleveland clinic and is available UCI Repository . First the data is preprocessed and checked if there any missing values or redundances. The dataset has 6 missing records hence 300 patient records are used in pre processing. After pre processing is completed the data analyzed for any patterns that can be observed in the data set or checked if there are any dependent values that are present the dataset. Then the target algorithms that required to find the solution to the problem statement. Each algorithm is run individually on the data , analyzed and the results (i.e., accuracy , f-score) are generated. This method is repeated for the all desired algorithms. A graph is then generated for the accuracy generated for each algorithm. Parameter tuning is also performed to improve the accuracy if possible. (For the current model although hyperparameter tuning is performed there is no change is accuracy that can be observed as the input dataset is small.) The algorithm with the highest accuracy value is taken and is run along with Neural Network model (i.e., as hybrid model where output value of one model is given as input value to another model to generate the desired output.) and the algorithm is run together and the output is generated. The model output are then compared between the machine learning model and the hybrid model.

IV RESULTS

The results for the Machine Learning model are generated as

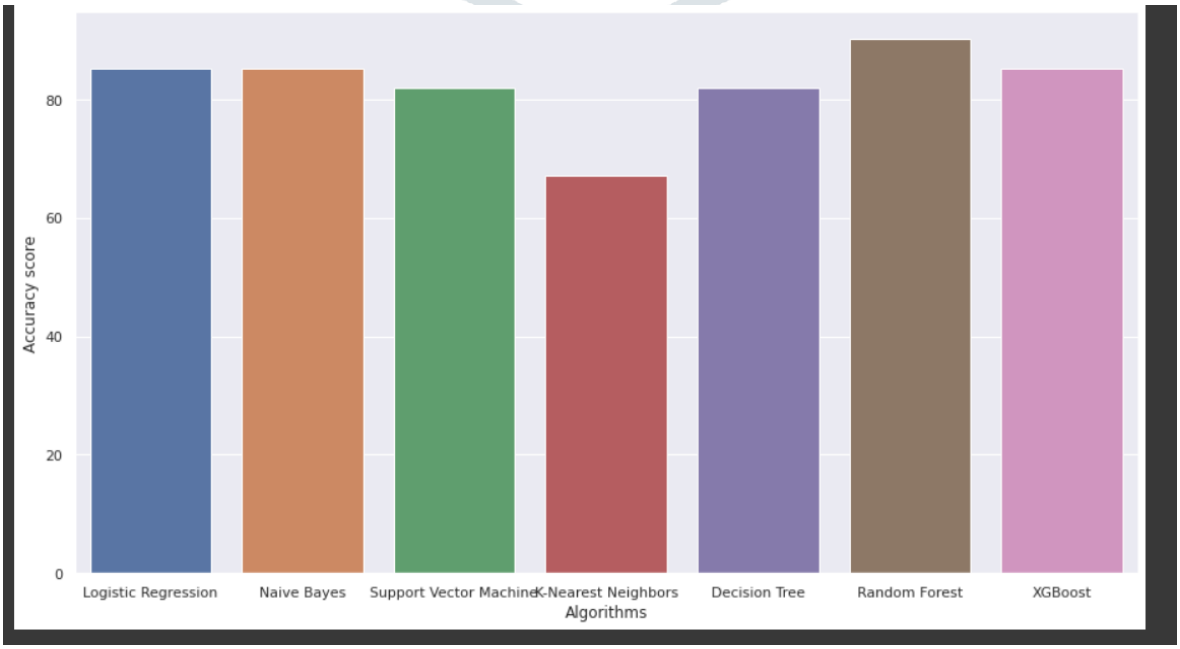

Fig : Accuracy comparison graph of the different algorithms used in the model

From the above graph
The Y axis represents the Accuracy score.
The X axis represents the Algorithms implemented.
From the above figured it can inferred that Random Forest Algorithm generates the highest accuracy in comparison to other algorithms.

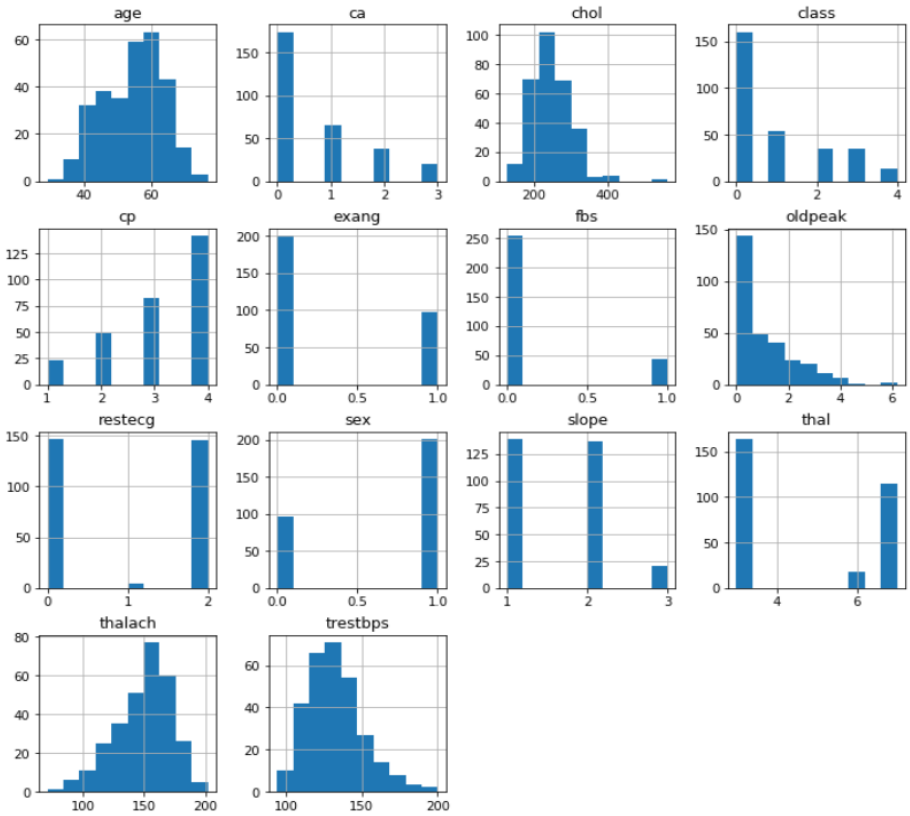

Fig : Histogram of values of all the attributes.

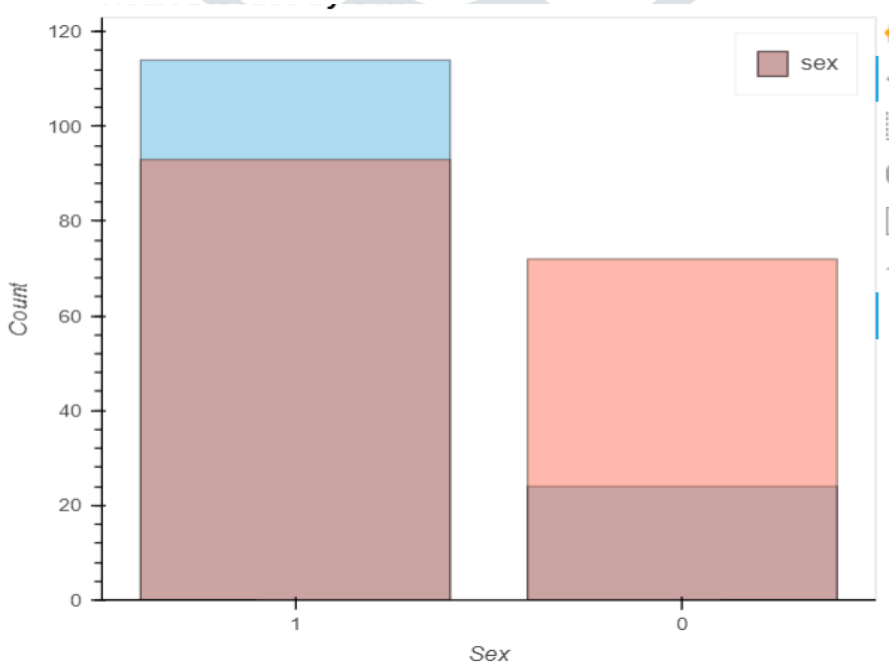

Fig : Probability of presence of heart disease with respect to gender

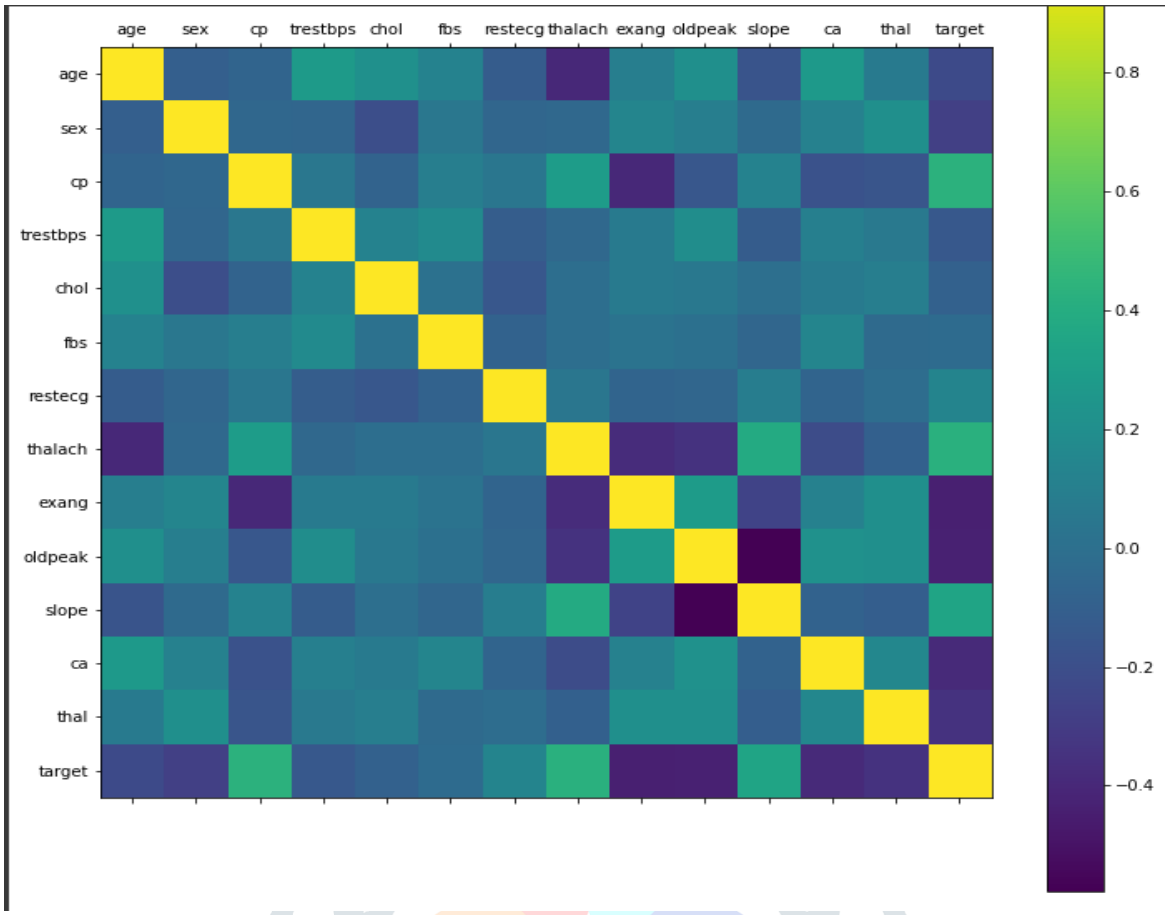

Fig : Correlation matrix generated with respect to attributes in the dataset

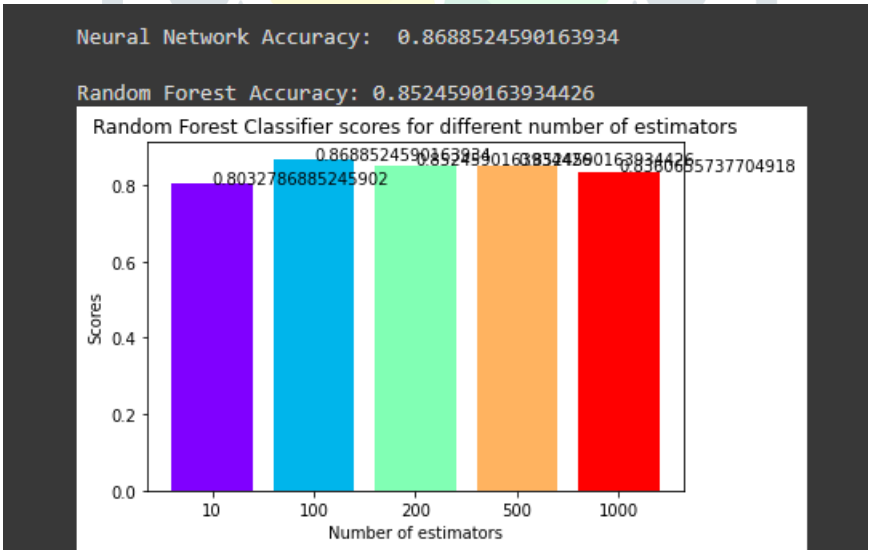

Fig : Classifier score generated for the hybrid model implemented

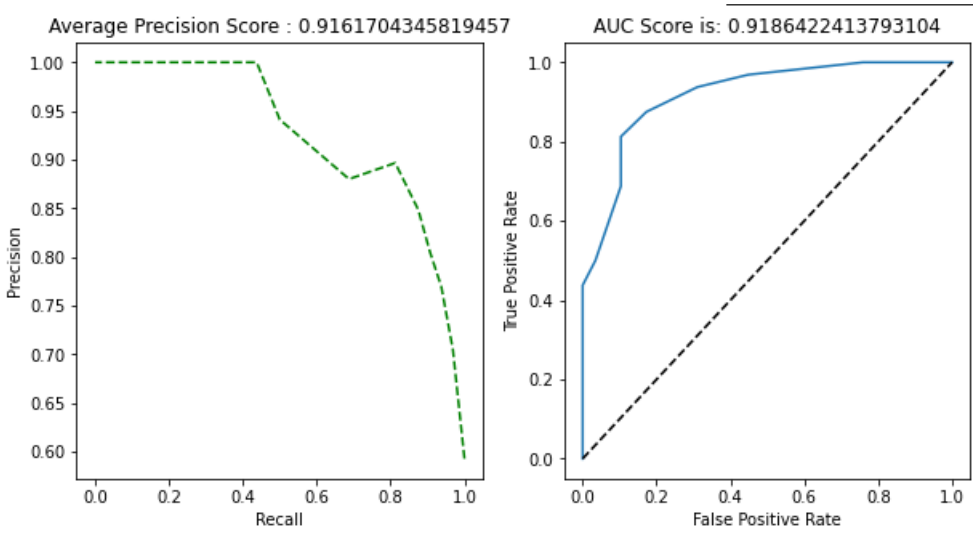

Fig : Average Precision and AUC score of the hybrid model

Where ,
Average Precision Score refers to the is the weighted sum of precisions at each threshold where the weight is the increase in recall. It is calculated according to the next equation. Using a loop that goes through all precisions/recalls, the difference between the current and next recalls is calculated and then multiplied.

AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1 .It is usually used in classification problem to determine the performance.
The ROC curve is plotted with True Positive rate(TPR) against the False Positive(FPR) rate where TPR is on the y-axis and FPR is on the x-axis.

V CONCLUSION

With the large amount of data generated by the medical industry identifying the raw data and the patterns would help in early prediction of the diseases. Thus by reducing the mortality rate caused by the disease. The proposed model is Hybrid models that used Random Forest Algorithm and Neural Network using Sequential classifier. The model has produced quite accurate results. Furthermore the model can be used for large datasets when compared to previous works using linear classifier models as linear models take each attribute as linearly dependent which may not be true for all values in the scenario of medical data. Hence we use sequential classifier. Additionally new feature selections can be made to get a border perspective in determining the disease. Data can be collected using IoT devices and the model can be implemented on live data unlike the current scenario wherein we have used a existing old medical dataset. Newer data would help in more precise prediction of the disease which would thus lead to building more efficient model.

REFERENCES

1. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in *IEEE Access*, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

2. P. Motarwar, A. Duraphe, G. Suganya and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning," *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 2020, pp. 1-5, doi: 10.1109/ic-ETITE47903.2020.242.

3. V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2020, pp. 177-181, doi: 10.1109/ICACCCN51052.2020.9362842.

4. Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN COMPUT. SCI.* **1,** 345 (2020). https://doi.org/10.1007/s42979-020-00365-y https://doi.org/10.1007/s42979-020-00365-y

5. Combining Machine Learning and Simulation to a Hybrid Modelling Approach: Current and Future Directions https://link.springer.com/chapter/10.1007/978-3-030-44584-3_43

6. Taeshik Shon, Jongsub Moon,A hybrid machine learning approach to network anomaly detection,Information Sciences,Volume 177, Issue 18,2007,Pages 379 3821, ISSN 0020-0255,https://doi.org/10.1016/j.ins.2007.03.025.(https://www.sciencedirect.com/science/article/pii/S0020025507001648)

7. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.258.8158&rep=rep1&type=pdf

8. S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," *2008 IEEE/ACS International Conference on Computer Systems and Applications*, 2008, pp. 108-115, doi: 10.1109/AICCSA.2008.4493524.

9. J. Thomas and R. T. Princy, "Human heart disease prediction system using data mining techniques," *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, 2016, pp. 1-5, doi: 10.1109/ICCPCT.2016.7530265.

10. A. Rairikar, V. Kulkarni, V. Sabale, H. Kale and A. Lamgunde, "Heart disease prediction using data mining techniques," *2017 International Conference on Intelligent Computing and Control (I2C2)*, 2017, pp. 1-8, doi: 10.1109/I2C2.2017.8321771.

11. Jaime Lynn Speiser, Michael E. Miller, Janet Tooze, Edward Ip,A comparison of random forest variable selection methods for classification prediction modeling,Expert Systems with Applications,Volume 134,2019,ISSN 0957-4174,

12. M. Sultana, A. Haider and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, 2016, pp. 1-5, doi: 10.1109/CEEICT.2016.7873142.

13. Maas, A., Appelman, Y. Gender differences in coronary heart disease. *Neth Heart J* **18,** 598–603 (2010). https://doi.org/10.1007/s12471-010-0841-y