JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 09, ISSUE 02, 2022

Heart Disease Prediction Using Machine Learning Techniques.

Dr. Vadhri Suryanarayana
Professor, Department of Computer Science and Engineering,
Ramachandra College of Engineering, Eluru, AP, INDIA
vs@rcee.ac.in

Dr. Satyabrata Dash,
Associate Professor ,Department of Computer Science and Engineering,
Ramachandra College of Engineering, Eluru, AP,INDIA
Satyabrata.cse@rcee.ac.in

Dr. Shameena Begum
Professor, Department of Computer Science and Engineering,
Ramachandra College of Engineering, Eluru, AP, INDIA
sameenamz@gmail.com

Sujata Chakarvarty
Professor Department of Computer Science and Engineering,
Centurion University of Technology & Management, Odisha, INDIA
sujata.chakravarty@cutm.ac.in

Y. Nagendra Kumar
Assistant Professor, Department of Computer Science and Engineering,
Ramachandra College of Engineering, Eluru, AP, INDIA
nagendrayakkala@rcee.ac.in

K. Venkatesh
Assistant Professor, Department of Computer Science and Engineering,
Ramachandra College of Engineering, Eluru, AP, INDIA
venkatesh.kondaveti@gmail.com

ABSTRACT

Heart disease is now days one of the primary basis of death in world. Prediction of heart diseases requires more precision, perfection and correctness because a little mistake can cause death of the person and also it associates with many risky factors. To deal with the problem there is essential need of prediction system to get accurate and reliable about diseases. Machine learning provides the way to predict any kind of event which take training from natural events. In this paper we have implemented supervise machine learning classification algorithms like Logistic Regression, K-Nearest Neighbour, SVM, Decision Tree, and Random Forest and calculate accuracy by using existing dataset from the Cleveland database of UCI repository of heart disease patients

***Keywords*-**Logistic Regression, SVM, Decision Tree, Random Forest, KNN

1. INTRODUCTION

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researchers have been conducted in attempt to pinpoint

ISSN- 2394-5125 VOL 09, ISSUE 02, 2022

the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithms. Over the last decade, heart disease or cardiovascular remains the primary basis of death worldwide. An estimate by the World Health Organization, that over 17.9 million deaths occur every year worldwide because of cardiovascular disease, and of these deaths, 80% are because of coronary artery disease and cerebral stroke [1]. The vast number of deaths is common amongst low and middle-income countries [2]. Many predisposing factors such as personal and professional habits and genetic predisposition accounts for heart disease. Various habitual risk factors such as smoking, overuse of alcohol and caffeine, stress, and physical inactivity along with other physiological factors like obesity, hypertension, high blood cholesterol, and pre-existing heart conditions are predisposing factors for heart disease. The efficient and accurate and early medical diagnosis of heart disease plays a crucial role in taking preventive measures to prevent death.Data mining refers to the extraction of required information from huge datasets in various fields such as the medical field, business field, and educational field. Machine learning is one of the most rapidly evolving domains of artificial intelligence. These algorithms can analyse huge data from various fields, one such important field is the medical field. It is a substitute to routine prediction modelling approach using a computer to gain an understanding of complex and non-linear interactions among different factors by reducing the errors in predicted and factual outcomes [3]. Data mining is exploring huge datasets to extract hidden crucial decision-making information from a collection of a past repository for future analysis. The medical field comprises tremendous data of patients. These data need mining by various machine learning algorithms. Healthcare professionals do analysis of these data to achieve effective diagnostic decision by healthcare professionals. Medical data mining using classification algorithms provides clinical aid through analysis. It tests the classification algorithms to predict heart disease in patients [4]. Data mining is the process of extracting valuable data and information from huge databases. Various data mining techniques such as regression, clustering, association rule and classification techniques like Naïve Bayes, decision tree, random forest and K-nearest neighbor are used to classify various heart disease attributes in predicting heart disease. A comparative analysis of the classification techniques is used [5]. In this research, I have taken dataset from the UCI repository. The classification model is developed using classification algorithms like Logistic Regression, SVM, Decision Tree, Random Forest, and KNN for prediction of heart disease. In this research, we have analysed of these algorithmsfor heart disease prediction, comparison among the existing systems is made. It also mentions further research and advancement possibilities in the paper.

2. PROBLEM DEFINITION

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance

ISSN- 2394-5125 VOL 09, ISSUE 02, 2022

of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

3. LITERATURE REVIEW

With growing development in the field of medical science alongside machine learning various experiments and researches has been carried out in these recent years releasing the relevant significant papers. The paper [1] propose heart disease prediction using KStar, J48, SMO, and Bayes Net and Multilayer perceptron using WEKA software. Based on performance from different factor SMO (89% of accuracy) and Bayes Net (87% of accuracy) achieve optimum performance than KStar, Multilayer perceptron and J48 techniques using k-fold cross validation. The accuracy performance achieved by those algorithms are still not satisfactory. So that if the performance of accuracy is improved more to give batter decision to diagnosis disease. [2]In a research conducted using Cleveland dataset for heart diseases which contains 303 instances and used 10-fold Cross Validation, considering 13 attributes, implementing 4 different algorithms, they concluded Gaussian Naïve Bayes and Random Forest gave the maximum accuracy of 91.2 percent. [3]Using the similar dataset of Framingham, Massachusetts, the experiments were carried out using 4 models and were trained and tested with maximum accuracy K Neighbors Classifier: 87%, Support Vector Classifier: 83%, Decision Tree Classifier: 79% and Random Forest Classifier: 84%. Heart disease affects millions of people, and it remains the chief cause of death in the world. Medical diagnosis should be proficient, reliable, and aided with computer techniques to reduce the effective cost for diagnostic tests. Data mining is a software technology that helps computers to build and classify various attributes. This research paper uses classification techniques to predict heart disease. This section gives a portrayal of the related subjects like machine learning and its methods with brief descriptions, data pre-processing, evaluation measurements and description of the dataset used in this research.

4. DATASETS

The dataset is publicly available on the Kaggle Website at [4] which is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. It provides patient information which includes over 4000 records and 14 attributes. The attributes include: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting, sugar blood, resting electrocardiographic results, maximum heart rate, exercise induced angina, ST depression induced by exercise, slope of the peak exercise, number of major vessels, and target ranging from 0 to 2, where 0 is absence of heart disease. The data set is in csv (Comma

ISSN- 2394-5125 VOL 09, ISSUE 02, 2022

Separated Value) format which is further prepared to data frame as supported by pandas library in python.

5. PROPOSED SYSTEM

This research aims to foresee the odds of having heart disease as probable cause of computerized prediction of heart disease that is helpful in the medical field for clinicians and patients [5]. To accomplish the aim, we have discussed the use of various machine learning algorithms on the data set and dataset analysis is mentioned in this research paper. This paper additionally depicts which attributes contribute more than the others to anticipation of higher precision. This may spare the expense of different trials of a patient, as all the attributes may not contribute such a substantial amount to expect the outcome [5].

4.1. Data Pre-processing

The real-life information contains large numbers with missing and noisy data. These data are pre-processed to overcome such issues and make predictions vigorously. Figure 1 explains the sequential chart of our proposed model.

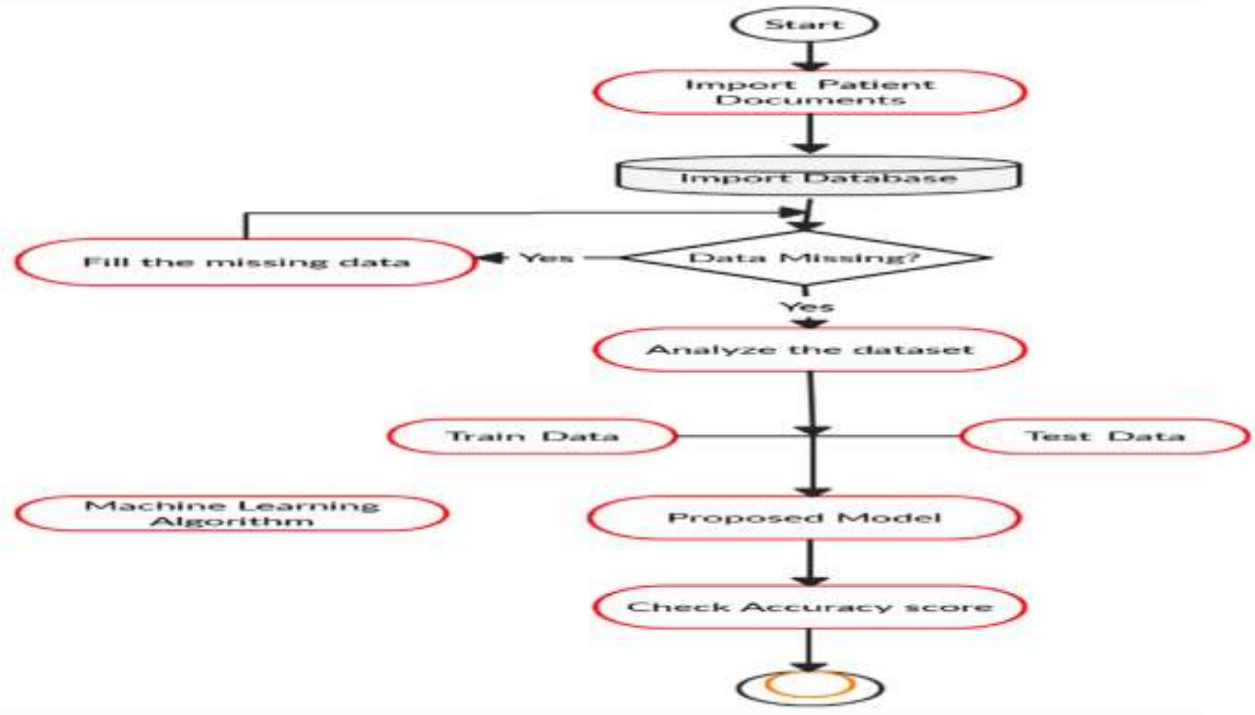

Figure-1. Data Processing

Cleaning the collected data usually has noise and missing values. To get an accurate and effective result, these data need to be cleaned in terms of noise and missing values are to be filled up.

ISSN- 2394-5125 VOL 09, ISSUE 02, 2022

Transformation it changes the format of the data from one form to another to make it more comprehensible. It involves smoothing, normalization, and aggregation tasks. Integration the data may not be acquired from a single source but varied sources, and it has to be integrated before processing. Reduction the data gained are complex and require to be formatted to achieve effective results. The data are then classified and split into training data set and test data set which is run on various algorithms to achieve accuracy score results.

4.2. Machine Learning Algorithms

There are many supervised classification algorithms that have used which are as follows:

a. Logistic Regression

Logistic regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature. Dichotomous means there are only two possible classes. For example, it can be used for cancer detection problems. It computes the probability of an event occurrence.

It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function.

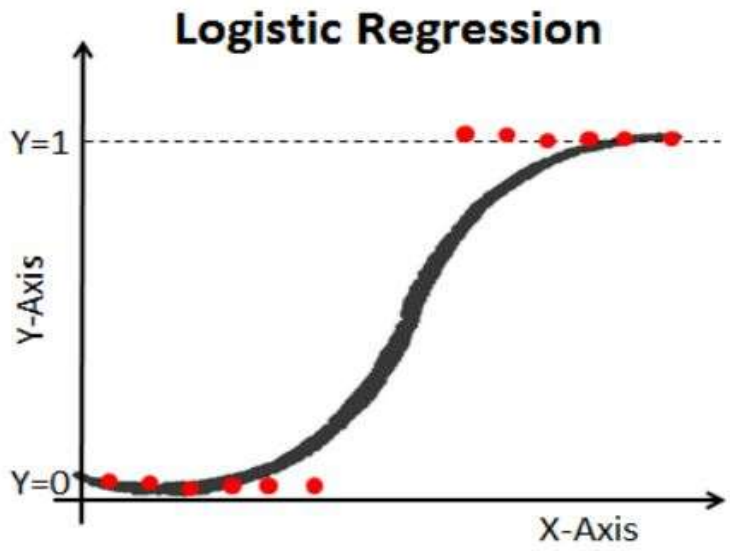

Figure -2 Logistic Regression

There are three types of Logistic Regression:

- Binary Logistic Regression: The target variable has only two possible outcomes such as Spam or Not Spam, Cancer or No Cancer.

ISSN- 2394-5125 VOL 09, ISSUE 02, 2022

- Multinomial Logistic Regression: The target variable has three or more nominal categories such as predicting the type of Wine.

- Ordinal Logistic Regression: the target variable has three or more ordinal categories such as restaurant or product rating from 1 to 5.

Logistic Regression is a simple, easy to implement, and efficient classification algorithm that handles non-linear, complicated data. However, there is a loss of accuracy as it is based on assumption and class conditional independence.

An accuracy of 86.81% has been achieved in Naïve Bayes with 10 most important predictors chosen using Logistic Regression.

b. Support Vector Machine(SVM)

Generally, Support Vector Machines is considered to be a classification approach, it but can be employed in both types of classification and regression problems. It can easily handle multiple continuous and categorical variables. SVM constructs a hyperplane in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane(MMH) that best divides the dataset into classes.

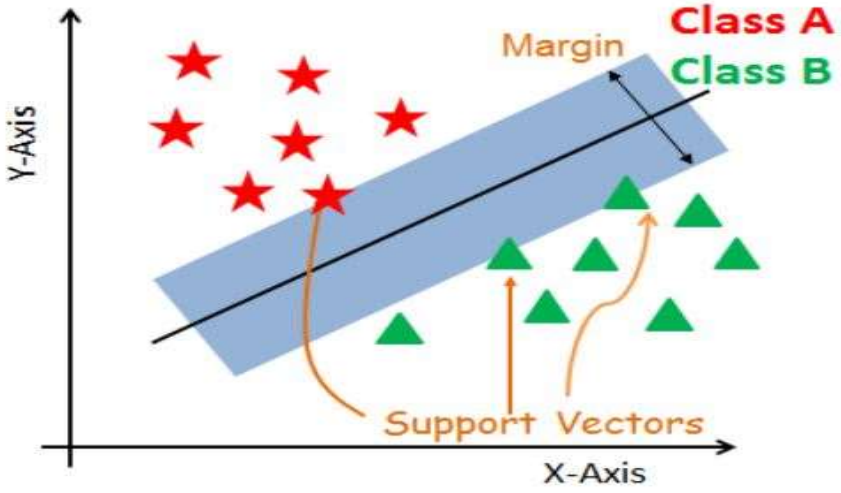

Figure-3. Support Vector Machines

c. Decision Tree

Decision tree is a classification algorithm that works on categorical as well as numerical data. Decision tree is used for creating tree-like structures. Decision tree is simple and widely used to handle medical dataset. It is easy to implement and analyse the data in tree-shaped graph. The decision tree model makes analysis based on three nodes.

- Root node: main node, based on this all other nodes functions.

ISSN- 2394-5125 VOL 09, ISSUE 02, 2022

- Interior node: handles various attributes.
- Leaf node: represent the result of each test.

This algorithm splits the data into two or more analogous sets based on the most important indicators. The entropy of each attribute is calculated and then the data are divided, with predictors having maximum information gain or minimum entropy:

Entropy(S)=∑i=1c−Pilog2Pi,Entropy(S)=∑i=1c−Pilog2Pi,
Gain (S,A)=Entropy(S)−∑v∈Values(A)|Sv||S|Entropy (Sv).Gain (S,A)=Entropy(S)−∑v∈Values(A)|Sv||S|Entropy (Sv).
The results obtained are easier to read and interpret [3]. This algorithm has higher accuracy in comparison to other algorithms as it analyzes the dataset in the tree-like graph. However, the data may be over classified and only one attribute is tested at a time for decision-making.

An accuracy of 71.43% has been achieved by the decision tree by Chauhan et al. [9], whereas the accuracy obtained was very poor about 42.8954% in [10].

d. Random Forest Algorithm

Random forest algorithm is a supervised classification algorithmic technique. In this algorithm, several trees create a forest. Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model's forecast. In the random forest classifier, the more number of trees give higher accuracy. The three common methodologies are:

- Forest RI (random input choice);
- Forest RC (random blend);
- Combination of forest RI and forest RC.

It is used for classification as well as regression task, but can do well with classification task, and can overcome missing values. Besides, being slow to obtain predictions as it requires large data sets and more trees, results are unaccountable.

Random forest algorithm has obtained an accuracy of 91.6% with Cleveland dataset in [11]. Using People's dataset, an accuracy of 97% was achieved.

e. K-Nearest Neighbor (K-NN)

The K-nearest neighbors algorithm is a supervised classification algorithm method. It classifies objects dependant on nearest neighbor. It is a type of instance-based learning. The calculation of distance of an attribute from its neighbors is measured using Euclidean distance [3]. It uses a group of named points and uses them on how to mark another point. The data are clustered based on similarity amongst them, and is possible to fill the missing values of data using K-NN. Once the missing values are filled, various prediction techniques

apply to the data set. It is possible to gain better accuracy by utilizing various combinations of these algorithms.

K-NN algorithm is simple to carry out without creating a model or making other assumptions. This algorithm is versatile and is used for classification, regression, and search. Even though K-NN is the simplest algorithm, noisy and irrelevant features affect its accuracy. In a study by Pouriyeh et al., 83.16% accuracy was achieved with value $K = 9$ [8].

5. SIMULATIONS

5.1.Data set and attributes

The data is collected from the UCI machinelearning repository. The data set is named Heart Disease DataSet and can be found in the UCI machine learning repository. The UCI machine learning repository contains a vast and varied amount of datasets which include datasets from various domains. These data are widely used by machine learning community from novices to experts to understand data empirically. Various academic papers and researches have been conducted using this repository. This repository was created in 1987 by David Aha and fellow students atUCI Irvine. Heart disease dataset contains data from four institutions [18].

1. Cleveland Clinic Foundation.
2. Hungarian Institute of Cardiology, Budapest.
3. V.A. Medical Centre, Long Beach, CA.
4. University Hospital, Zurich, Switzerland.

For the purpose of this study, the data set provided by the Cleveland Clinic Foundation is used. This dataset was provided by Robert Detrano, M.D, Ph.D. Reason to choose this dataset is, it has less missing values and is also widely used by the research community [19].

Table 1. Attributes of the Heart disease dataset

Attribute	Representation	Information Attribute	Description
Age	Age	Integer	Age in years (29 to 77)
Sex	Sex	Integer	Gender instance (0 = Female, 1 = Male)
ChestPainType	Cp	Integer	Chest pain type (1: typical angina, 2: atypical angina, 3: non- anginal pain, 4: asymptomatic)
RestBloodPressure	Trestbps	Integer	Resting blood pressure in mm Hg[94, 200]
SerumCholestoral	Chol	Integer	Serum cholesterol in mg/dl[126, 564]
FastingBloodSugar	Fbs	Integer	Fasting blood sugar > 120 mg/dl (0 = False, 1= True)
ResElectrocardiographic	Restecg	Integer	Resting ECG results (0: normal, 1: ST-T wave abnormality, 2: LV hypertrophy)
MaxHeartRate	Thalach	Integer	Maximum heart rate achieved[71, 202]
ExerciseInduced	Exang	Integer	Exercise induced angina (0: No, 1: Yes)

ISSN- 2394-5125 VOL 09, ISSUE 02, 2022

Oldpeak	Oldpeak	Real	ST depression induced by exercise relative to rest[0.0, 62.0]
Slope	Slope	Integer	Slope of the peak exercise ST segment (1: up-sloping, 2: flat, 3: down-sloping)
MajorVessels	Ca	Integer	Number of major vessels coloured by fluoroscopy (values 0 - 3)
Thal	Thal	Integer	Defect types: value 3: normal, 6: fixed defect, 7: irreversible defect
Class	Class	Integer	Diagnosis of heart disease (1: Unhealthy, 2: Healthy)

The fig-4,5,6 shows the different simulations based on the attributes of the Heart disease dataset.

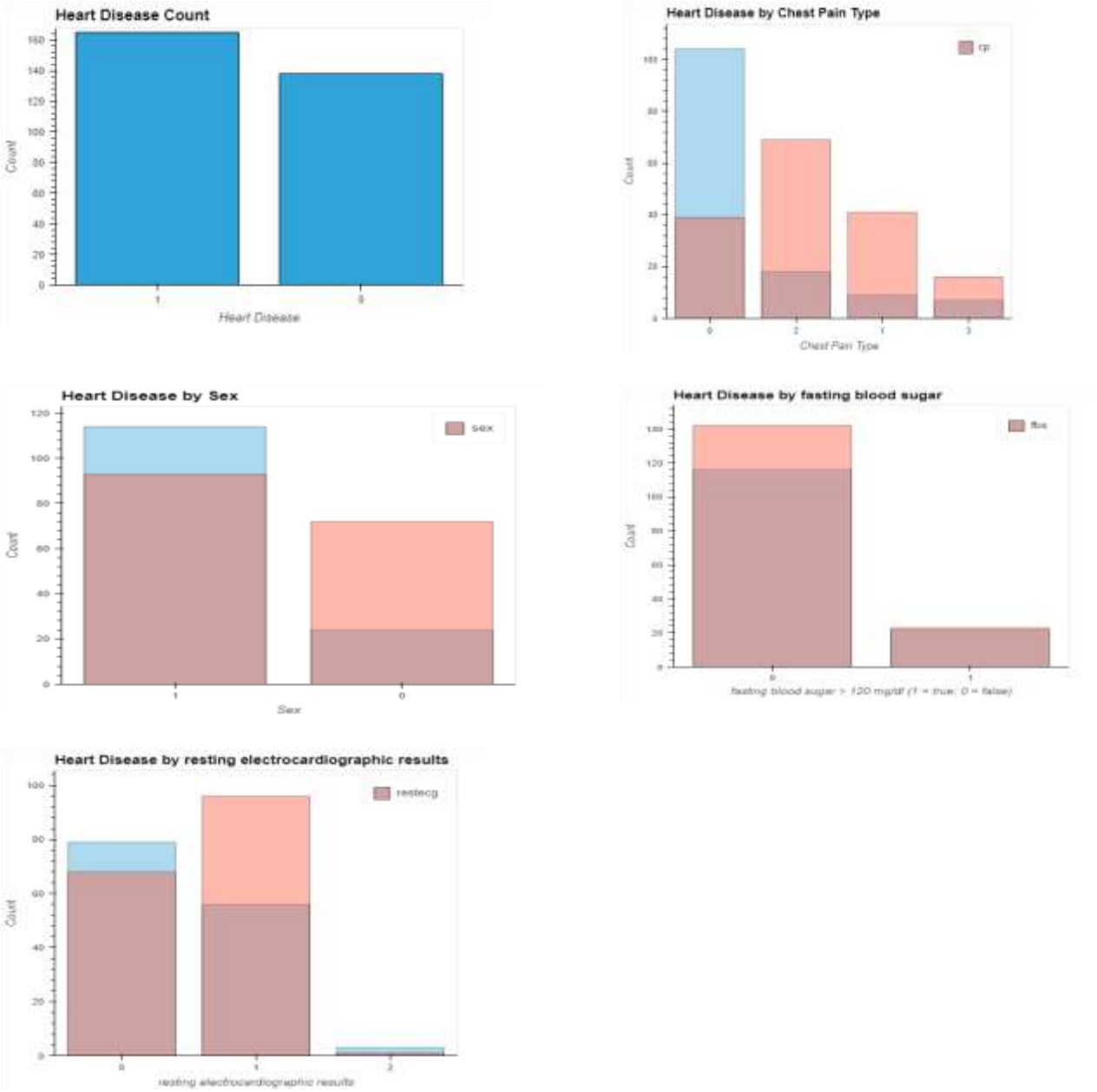

Figure -4 Simulations on Heart disease dataset.

ISSN- 2394-5125 VOL 09, ISSUE 02, 2022

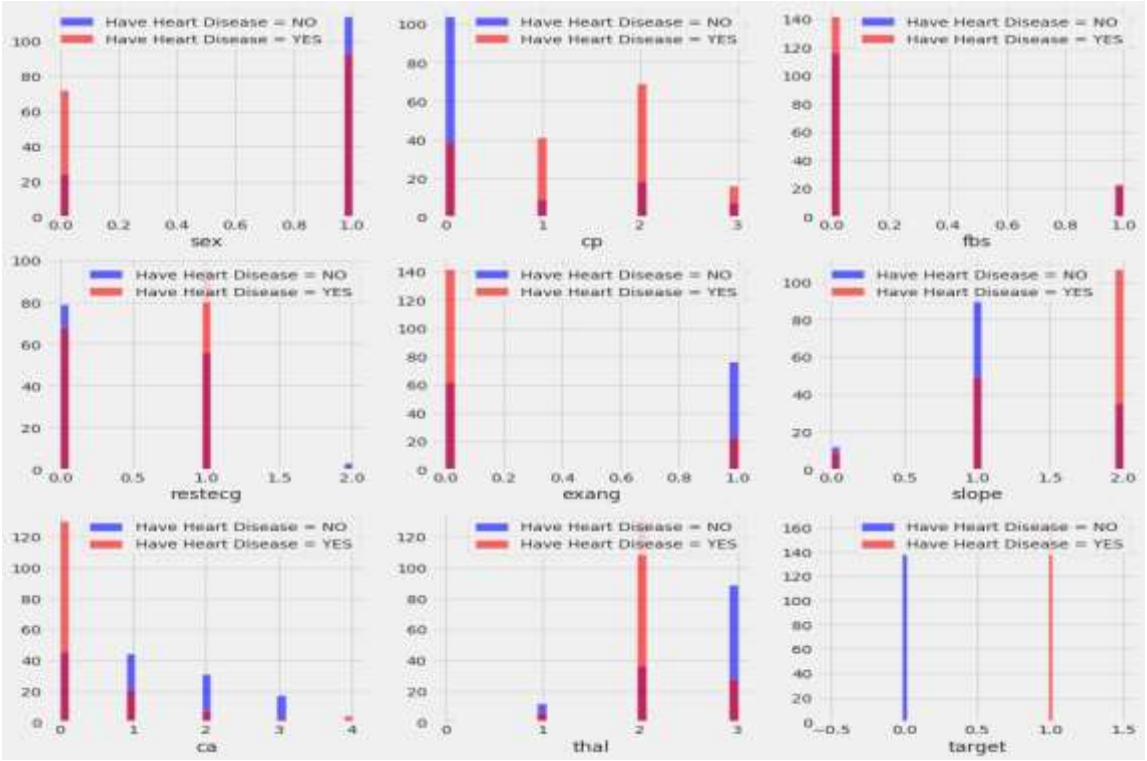

Figure -5 Simulations on Heart disease dataset.

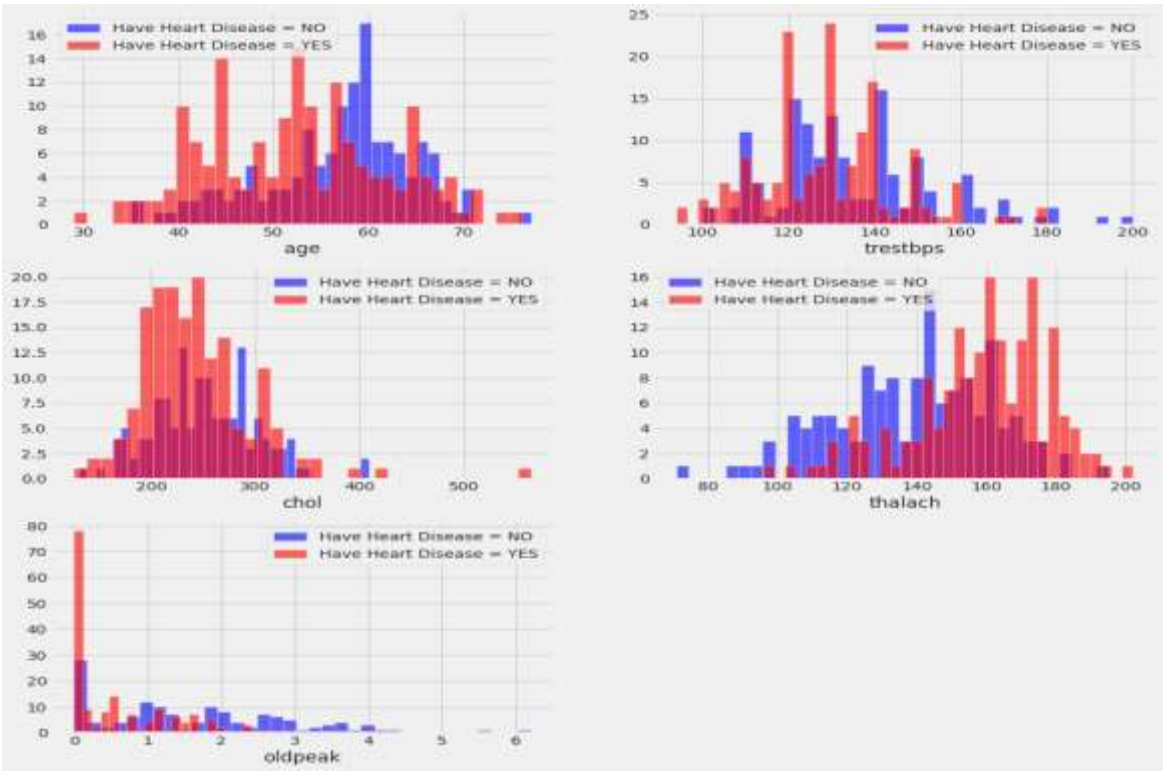

Figure -6 Simulations on Heart disease dataset.

ISSN- 2394-5125 VOL 09, ISSUE 02, 2022

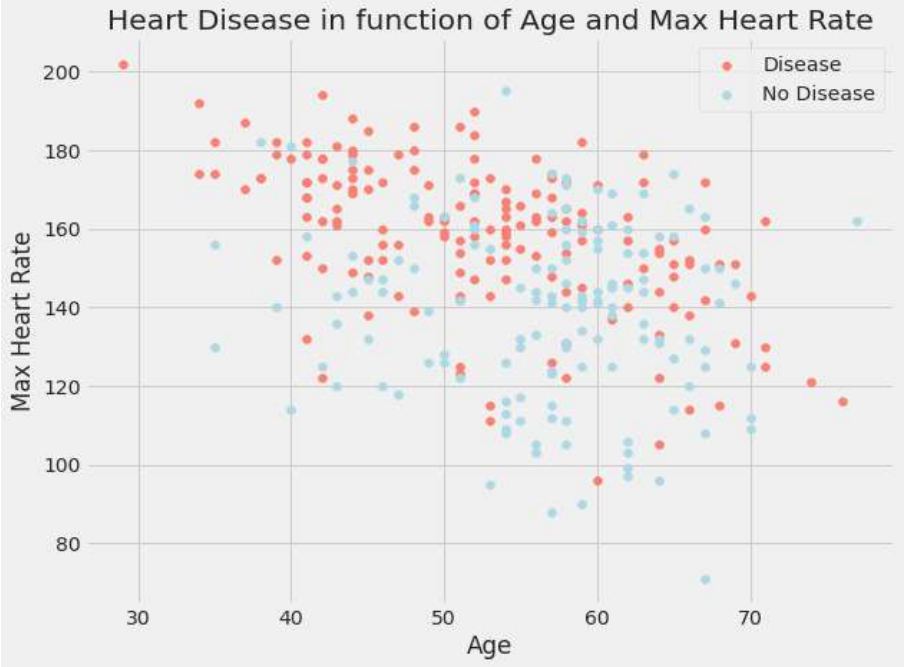

Figure -7 Simulations on Heart disease dataset.

5. Models Building

5.1.EXPERIMENTAL ANALYSIS & RESULTS

Aim of this research is to predict whether or not a patient will develop heart disease. This research was done on supervised machine learning classification techniques using Logistic Regression, decision tree, random forest, and K-nearest neighbor on UCI repository. Various experiments using different classifier algorithms were conducted. Research was performed on 8th generation Intel Corei7 having an 8750H processor up to 4.1 GHz CPU and 16 GB ram. Dataset was classified and split into a training set and a test set. Pre-processing of the data is done and supervised classification techniques such as Logistic Regression, decision tree, K-nearest neighbor, and random forest are applied to get accuracy score. The accuracy score results of different classification techniques were noted using Python Programming for training and test data sets. Percentage accuracy scores are depictedfor different algorithms. Comparison of accuracy score of heart disease prediction in proposed model also we have evaluated.Since our project is a classification problem, we use accuracy, precision, recall and F1 score to evaluate the models. We would like to introduce the meaning of TP,FP,TN and FN. A true positive (TP) is a positive outcome predicted by the model correctly while a false positive (FP) is a positive outcomepredicted by the model incorrectly. A true negative (TN) is a negative outcome predicted by the model correctly while a falsenegative (FN) is a negative outcome predicted by the model incorrectly.

ISSN- 2394-5125 VOL 09, ISSUE 02, 2022

We did not use cross-validation because our dataset is not very sufficient. We split the dataset into 80% for training and 20% for test. Here is the table of results of different methods and we will talk about each evaluation of methods in details.

TABLE 2. RESULT OF DIFFERENT METHODS

Methods	**Train accuracy**	**Test accuracy**	**precision**	**recall**	**F1 score**
Logistic Regression	83.88%	85.25%	0.88	0.78	0.82
SVM	89.26%	86.89%	0.91	0.78	0.84
Naïve Bayes	83.47%	85.25%	0.88	0.78	0.82
Random Forest	100%	91.80%	0.92	0.89	0.91
Neural Network	83.88%	88.52%	0.92	0.81	0.86

A. Logistic Regression

Here is the confusion matrix of the Logistic Regression:

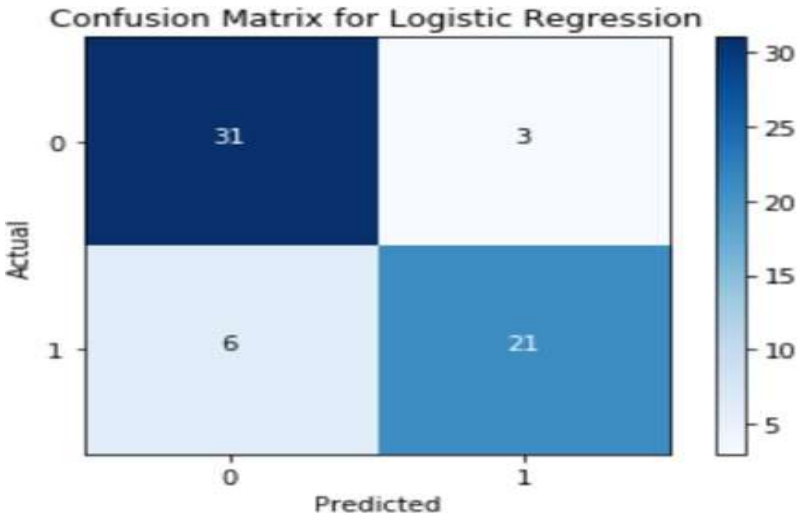

Figure-8. Confusion Matrix for Logistic Regression

I used the L2 penalty, the square of the magnitude of coefficients, supported by Logistic Regression to avoid overfitting. The train accuracy is 83.88% and test accuracy is 85.25%. It performs well but not the best for us. The advantage of the Logistic Regression is that it does not need too much computational resources and it is highly interpretable. So it is easy and sufficient to apply Logistic Regression. However, the limitation of Logistic Regression is that it assumes linearity between the features of the dataset. In the real world, the data is rarely separable, neither as our dataset. That is why we cannot reach a very high accuracy of 90%.

ISSN- 2394-5125 VOL 09, ISSUE 02, 2022

B. SVM

Here is the confusion matrix for SVM:

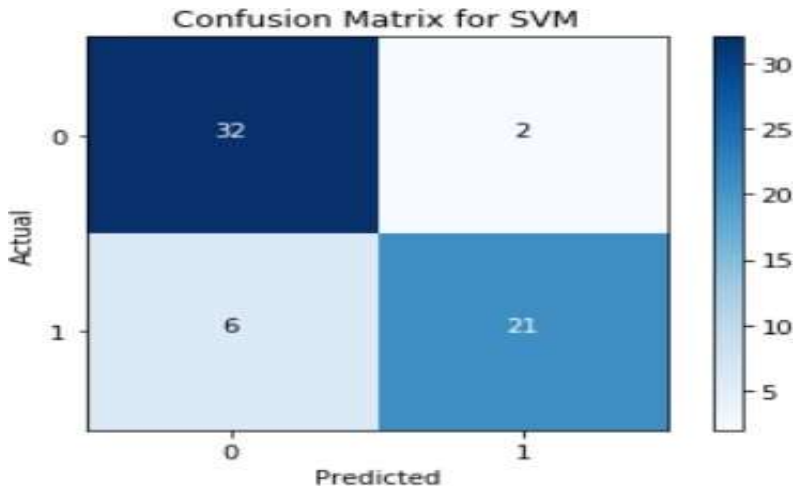

Figure- 9. Confusion Matrix for SVM

According to the tutorial of sklearn, for a small dataset it is better to use sklearn.svm.SVC(). The training accuracy is 89.26% and the test accuracy is 86.89%. The advantage of SVM is that it is very efficient with high dimensional spaces. The main disadvantage is that the SVM has many parameters that needs to be correctly chosen to achieve the best performance. For safety we just use the default parameters of SVM. And the test accuracy of 86.89%, which is better than Logistic Regression.

C. Naive Bayes

The confusion matrix for Naïve Bayes is

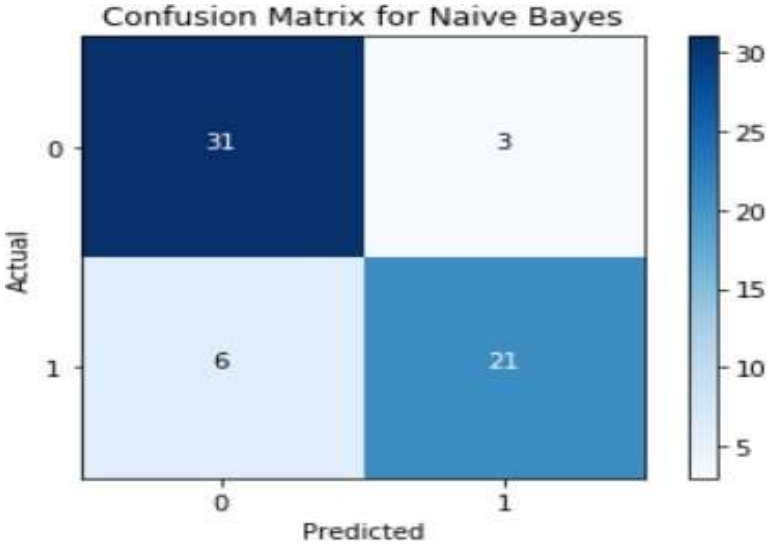

Figure- 10. Confusion Matrix for Naïve Bayes

ISSN- 2394-5125 VOL 09, ISSUE 02, 2022

The train accuracy is 83.47% and the test accuracy is 85.25%. The advantage of Naïve Bayes is that Naïve Bayes is able to make predications given a small amount of training data. The disadvantage of Naïve Bayes is that it assumes all features are mutually independent but in real life we can rarely get a dataset whose attributes are mutually independent and that might be why we cannot reach a very high accuracy of 90%.

D. Random Forest

The confusion matrix of Random Forest is:

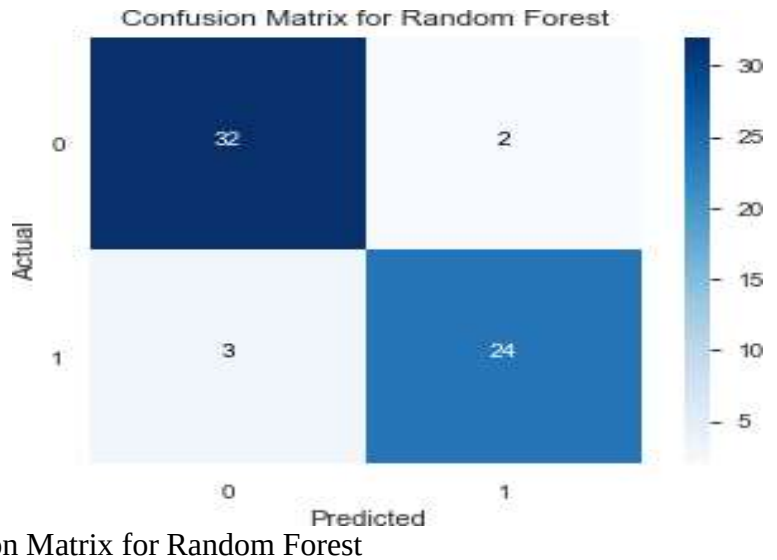

Figure-11. Confusion Matrix for Random Forest

The train accuracy is 100% and the test accuracy is 91.80%. At the first beginning we use the default parameters (n_estimators=100, which means the number of trees in the forest is 100 and max_depth = None, which means the nodes are expanded until all leaves are pure or all leaves contain less than the minimum number of samples required to split an internal node). Though we get 100% test accuracy, we only get 85.25% test accuracy. We guess it might be overfitting. One reason might be the training data is not generalized during the training process so we decide to shuffle the dataset again and we tried the parameter random_state from 1 to 2000. When random_state is 1826, the test accuracy is 91.80%. Then we tried experiments on parameters of n_estimatros(from 10 to 300) and max_depth(from 10 to 300) and the best test accuracy is still 91.80%. This means with random_state=1825, the other default parameters are good enough to get the best test accuracy. For example, the number of trees in the forest is 100, which is appropriate. If the number of trees is small, it will cause underfitting because the model has not been optimized for the training data, let alone the test data. If the number of trees is too big, it will cause overfitting because the model become so complexed and sensitive to new data. The advantage of Random Forest is that it can deal with dataset with high features and balance the variance and it is not sensitive to the noise of the data. Among these 5 models, Random Forest outperforms any

ISSN- 2394-5125 VOL 09, ISSUE 02, 2022

other models.

E. Neural Network

At the first beginning we tried to add 3-4 hidden layers in our neural network but it performs bad. The test accuracy is only 60%. Then we analyzed that the dataset is not big so we decide to make our network simple. At last we have only 1 hidden layer with 31 neuron nodes. For the optimization we use Adam instead of SGD (Stochastic Gradient Descent) because Adam is a combination of RMSprop and SGD with momentum and it takes advantage of momentum by moving average of the gradient. Our learning rate is 0.001, which is appropriate because the loss goes down in a normal speed. Since our dataset is not big, we just choose the batch size to be 200, which is enough for training. And we run 80 epochs to avoid overfitting. The train accuracy is 93.02% and the test accuracy is 88.52%, which is the secondbest. Here is the plot of Accuracy vs Epoch and Loss vs Epoch:

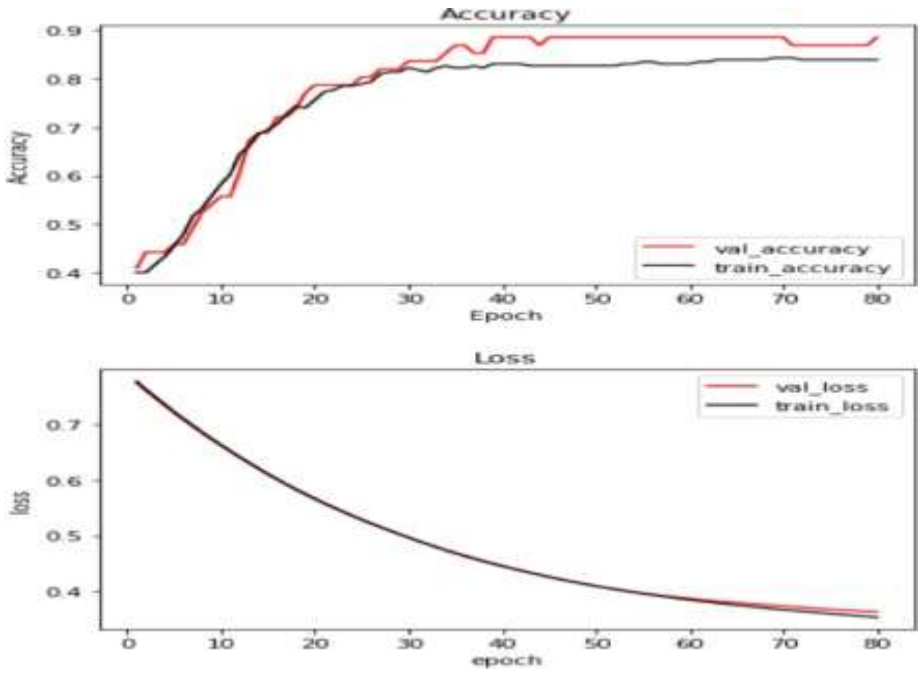

Fig. 7. Accuracy and Loss

As the epochs increase, the loss for test data is reaching 0.35 and test accuracy is reaching 89%. The advantage of the neural network is that neural network can deal with complicated datasets with high dimensional features (e.g. images) and make accurate predictions by building several hidden layers. However, when it comes to small dataset, the neural network does not perform well because it tends to become complicated.

6. CONCLUSION

The overall aim is to define various data mining techniques useful in effective heart disease prediction. Efficient and accurate prediction with a lesser number of attributes and tests is our goal. In this study, we consider only 14 essential attributes. It applied five data mining

classification techniques, K-nearest neighbor, Logistic Regression, SVM, decision tree, and random forest. The data were pre-processed and then used in the model. K-nearest neighbor, Logistic Regression, and random forest are the algorithms showing the best results in this model. I found the accuracy after implementing four algorithms to be highest in K-nearest neighbors ($k = 7$). We can further expand this research incorporating other data mining techniques such as time series, clustering and association rules,and genetic algorithm. Considering the limitations of this study, there is a need to implement more complex and combination of models to get higher accuracy for early prediction of heart disease.

7. REFERENCES

[1] Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol. 2011;3:67.

[2] Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. Growing epidemic of coronary heart disease in low-and middle-income countries. Curr Probl Cardiol. 2010;35(2):72–115.

[3] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4):e0174944.

[4] Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol. 2018;7(2.8):684–7. [5] Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Dis. 2015;7(1):129–37.

[5] Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl. 2017;9:1–16. https://doi.org/10.4236/jilsa.2017.91001.

[6] Pahwa K, Kumar R. Prediction of heart disease using hybrid technique for selecting features. In: 2017 4th IEEE Uttar Pradesh section international conference on electrical, computer and electronics (UPCON). IEEE. p. 500–504.

[7] Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In: 2017 IEEE symposium on computers and communications (ISCC). IEEE. p. 204–207.

[8] Chauhan R, Bajaj P, Choudhary K, Gigras Y. Framework to predict health diseases using attribute selection mechanism. In: 2015 2nd international conference on computing for sustainable global development (INDIACom). IEEE. p. 1880–84.

[9] Bouali H, Akaichi J. Comparative study of different classification techniques: heart disease use case. In: 2014 13th international conference on machine learning and applications. IEEE. p. 482–86.

[10] Xu S, Zhang Z, Wang D, Hu J, Duan X, Zhu T. Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework. In: 2017 IEEE 2nd international conference on big data analysis (ICBDA). IEEE. p. 228–32.

[11] Otoom AF, Abdallah EE, Kilani Y, Kefaye A, Ashour M. Effective diagnosis and monitoring of heart disease. Int J Softw Eng Appl. 2015;9(1):143–56.

[12] Abhishek, t., 2013. Heart Disease Prediction System Using Data Mining Techniques. Oriental Scientific Publishing Co., India, 6(4), pp. 457-466.

[13] Afroz, R. R. a. F., 2013. Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis. Journal of Software Engineering and Applications, 6(3), pp. 85-97.

[14] Aparna K., D. R. N. C. S. P. I. S. S. D. K. V., 2014. Disease Prediction in Data Mining Techniques. InternatIonal Journal of Computer SCIenCe and teChnology, 5(2), pp. 246- 249

[15] Soni, Jyoti, et al. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17.8 (2011): 43-48.

[16] Dangare, Chaitrali S., and Sulabha S. Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47.10 (2012): 44-48.

[17] Uyar, Kaan, and Ahmet İlhan. "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks." Procedia computer science 120 (2017): 588-593.

[18] Kim, Jae Kwon, and Sanggil Kang. "Neural network-based coronary heart disease risk prediction using feature correlation analysis." Journal of healthcare engineering 2017 (2017).

[19] Baccouche, Asma, et al. "Ensemble Deep Learning Models for Heart Disease Classification: A Case Study from Mexico." Information 11.4 (2020): 207.

[20] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[21] https://www.kaggle.com/ronitf/heart-disease-uci

[22] https://www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf

[23] https://nthu-datalab.github.io/ml/labs/03_Decision-Trees_RandomForest/03_Decision-Tree_Random-Forest.html

[24] https://www.kaggle.com/jprakashds/confusion-matrix-in-python-binaryclass

[25] scikit-learn, keras, pandas and matplotlib

[26] Dash, S., Chakravarty, S., Mohanty, S. N., Pattanaik, C. R., & Jain, S. (2021). A Deep Learning Method to Forecast COVID-19 Outbreak. New Generation Computing, 1-25.

[27] COVID-19 Outbreak in Orissa: MLR and H-SVR based Modelling and Forecasting, Satyabrata Dash, Hemraj Saini, S. Chakravarty. International Journal of Computer Applications in Technology ,Inderescience, Vol. 66, No. 3-4, pp 401–414, January 15, 2022

[28] Social Distancing Inspection To Mitigate COVID-19 Using K- Nearest Neighbour, Rutuparnna Mishra, Anshit Ransingh, Sujata Chakravarty, Satyabrata Dash,6th IEEE International Conference on Signal Processing, Computing and Control (ISPCC 2k21**)** 10.1109/ISPCC53510.2021.9609407

[29] Hyperspectral Image Classification using SVM with PCA, Mounika.K, K.Aravind,M.Yamini,P.Navyasr, Satyabrata Dash,V.Surryanarayan, 6th IEEE International Conference on Signal Processing, Computing and Control (ISPCC 2k21) 10.1109/ISPCC53510.2021.9609461

[30] A Mathematical Model for Analysis of COVID-19 Outbreak Using Von Bertalanffy Growth Function (VBGF), Biswajit Brahma, Satyabrata Dash, Prangya Parmita Kar, Subhendu Kumar Pani, Turkish Journal of Computer and Mathematics Education Vol.12 No. 11 (2021), 6063- 6075.