Heart Disease Prediction using Machine Learning

Sourav Joshi
Department of Computer Science andInformation Technology Dronacharya College of Engineering, Farukh Nagar, Gurgaon, India
Email- souravjoshi33@gmail.com

Ashwani *Kumar*
Faculty of Computer Science andInformation Technology
Dronacharya College of Engineering,Farukh Nagar, Gurgaon, India

***Abstract*—** Heart disease (HD) is one of the most common diseases nowadays, and people who provide health care, it is very necessary to work with them to take care of their patient's health and save their life. In this paper, different classifiers were analyzed by performance comparison to classify the Heart Disease dataset to classify it correctly and or to Predict Heart Disease cases with minimal attributes.

Large amounts of data that contain some secret information were collected by the healthcare industries. This data collection is useful for making effective decisions. Some advanced data mining techniques are used to make proper results and make effective decisions on data. In this case, a Heart Disease Prediction System (HDPS) is developed using Logistic Regression, K Nearest Neighbour, Decision Tree, Random Forest Classifier, and Support Vector Machine algorithms to predict the heart disease risk level.

The results reveal that the Random Forest Classifier and Support Vector Machine obtained the highest accuracy of 91.23%, whereas 86.79%, 71.69%, and 83.47% accuracy scores are obtained by logistic regression, KNN classifier, and decision tree respectively.

Keywords*— *Machine learning, Logistic regression, Heart disease, Support vector machine, accuracy

1. INTRODUCTION

"Machine Learning is a way of Manipulating and extraction of implicit, previously unknown/known and potential useful information about data" [1]. Machine Learning is a very vast and diverse field and its scope and implementation is increasing day by day. Machine learning Incorporates various classifiers of Supervised, Unsupervised and Ensemble Learning which are used to predict and Find the Accuracy of the given dataset. We can use that knowledge in our project of HDPS as it will help a lot of people.

The majority of people today experience an unhealthy and fast living style that accordingto the studies if giving jolt to the heart. The heart is the organ that pumps blood into various parts of the body through the vessels with a proper amount of oxygen and other essential nutrients. The survival of any organism relies solely on the proper functioning of the heart and, if the heart's pumping operation is troublesome, the body's main bodies such as brain and kidneys will undergo adverse effects. When the heart's work ceases, the death of the person takes place within minutes. Various diseases which can be attributed to our unhealthy lifestyles heart disease, angina pectoris, congestive heart failure, cardiomyopathy, congenital heart defects, arrhythmias, myocarditis, cardiac attack; cardiac cancer, etc. For coronary heart disease, the cardiovascular does not have enough blood to deliver the blood to the heart because of cholesterol and fat within its arterial wall. In case of heart attacks, where the direction of the coronary artery is blocked due to the coagulation of the blood on the heart's wall. During angina, pain in the chest is caused by a blood flow that does not function

properly in the heart. Other causes of cardiac disease include coronary artery disease, heart valvular disease, stroke, high blood pressure, etc.

The World Health statistics 2021 highlights the issue that every one in three adult age group showed prone to high blood pressure- a situation that results in half of the deaths from heart issues and strokes. Disease-related to the heart, also known as cardiovascular disease (CVD), discusses various conditions that affect the heart not just the disease. This juncture proved fatal for one person in every 34 seconds in the United States.

The objective of this project is to check whether the patient is likely to be diagnosed with any cardiovascular heart diseases based on their medical attributes such as gender, age, chest pain, fasting sugar level, etc. A dataset is selected from the UCI repository with patient's medical history and attributes. By using this dataset, we predict whether the patient can have a heart disease or not. To predict this, we use 14 medical attributes of a patient and classify him if the patient is likely to have a heart disease. These medical attributes are trained under three algorithms: Logistic regression, KNN and Random Forest Classifier. Most efficient of these algorithms is KNN which gives us the accuracy of 88.52%. And, finally we classify patients that are at risk of getting a heart disease or not and also this method is totally cost efficient.

2. LITERATURE REVIEW

A quiet Significant amount of work related to the diagnosis of Cardiovascular Heart disease using Machine Learning algorithms has motivated this work. This paper contains a brief literature survey. An efficient cardiovascular disease prediction has been made by using various algorithms some of them include Logistic Regression, KNN, Random Forest Classifier Etc. It can be seen in Results that each algorithm has its strength to register thedefined objectives **[7].**

The model incorporating IHDPS had the ability to calculate the decision boundary using the previous and new model of machine learning and deep learning. It facilitated the important and the most basic factors/knowledge such as family history connected with any heart disease. But the accuracy that was obtained in such IHDPS model was far more less than the new upcoming model such as detecting coronary heart disease using artificial neural network and other algorithms of machine and deep learning. The risk factors of coronary Heart disease or

atherosclerosis is identified by McPherson et al.**,[8]** using the inbuilt implementation algorithm using uses some techniques of Neural Network and were just accurately able to predict whether the test patient is suffering from the given disease or not.

Diagnosis and prediction of Heart Disease and Blood Pressure along with other attributes using the aid of neural networks was introduced by R. Subramanian et al.,[**24].**A deep Neural Network was Built incorporating the given attributes related to the disease which were able to produce a output which was carried out by the output perceptron and almost included 120 hidden layers which is the basic and most relevant technique of ensuring a accurate result of having heart disease if we use the model for Test Dataset. The supervised network has been advised for diagnosis of heart diseases [16]. When the testing of the model was done by a doctor using an unfamiliar data, the model used and trained from the previous learned data and predicted the result thereby calculating the accuracy of the given model.

3. DATA SET INFORMATION

An Organized Dataset of individuals had been selected Keeping in mind their history of heart problems and in accordance with other medical conditions **[2].** Heart disease are the diverse conditions by which the heart is affected. According to World Health Organization (WHO), the greatest number of deaths in middle aged people are due to cardiovascular diseases. We take a data source which is comprised of medical history of 304 different patient of different age groups. This dataset gives us the much-needed information i.e. the medical attributes such as age, resting blood pressure, fasting sugar level etc. of the patient that helps us in detecting the patient that is diagnosed with any heart disease or not. This dataset contains 13 medical attributes of 304 patients that helps us detecting if the patient is at risk of getting a heart disease or not and it helps us classify patients that are at risk of having a heart disease and that who are not at risk. This Heart Disease dataset is taken from the UCI repository. According to this dataset, the pattern which leads to the detection of patient prone to getting a heart disease is extracted. These records are split into two parts: Training and Testing. This dataset contains 303 rows and 14 columns, where each row corresponds to a single record. All attributes are listed in 'Table 1'.

Table 1. Various Attributes used are listed

S. No	Observation	Description	Values
1.	Age	Age in Years	Continuous
2.	Sex	Sex of Subject	Male/Female
3.	CP	Chest Pain	Four Types
4.	Trestbps	Resting Blood Pressure	Continuous
5.	Chol	Serum Cholesterol	Continuous
6.	FBS	Fasting Blood Sugar	< ,or > 120 mg/dl
7.	Restecg	Resting Electrocardiograph	Five Values
8.	Thalach	Maximum Heart Rate Achieved	Continuous
9.	Exang	Exercise Induced Angina	Yes/No
10.	Oldpeak	ST Depression when Workout compared to the Amount of Rest Taken	Continuous
11.	Slope	Slope of Peak Exercise ST segment	up/ Flat /Down
12.	Ca	Gives the number of Major Vessels Coloured by Fluoroscopy	0-3
13.	Thal	Defect Type	Reversible/Fixed/Normal
14.	Num(Disorder)	Heart Disease	Not Present /Present in the Four Major types.

The name of the dataset is heart.csv. There are 303 instances in this dataset, where the cases are either people having heart disease or they are healthy. Among 303, 165 (54.45%)cases are people with heart disease and 138 (45.54%) are people without heart disease. Thenumber of attributes is 14. There are no missing values in the data set nor any null values.

Features include age, sex, chest-pain type, rest BP, cholesterol, blood sugar level, ECG result, maximum heart rate achieved, exercise-induced angina, ST depression, the slope of peak exercise ST segment, number of major vessels, and defect in heart as of 3-normal, 6-fixed defect and 7- reversible defect. Bar graph (Fig.1) showing the positive and negative cases (1=positive, 0=negative) Scatter plot (Fig.2) showing the positive and negative cases depending on age

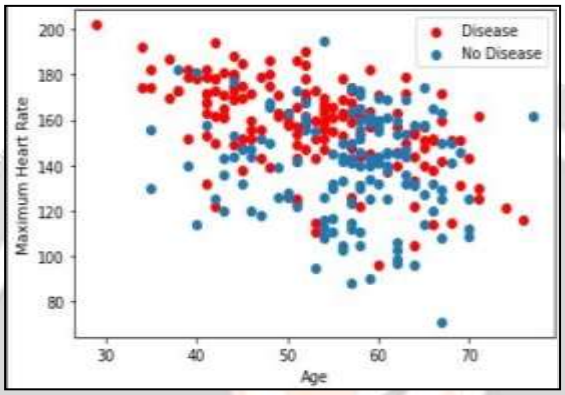


Fig.1 Positive and negative cases

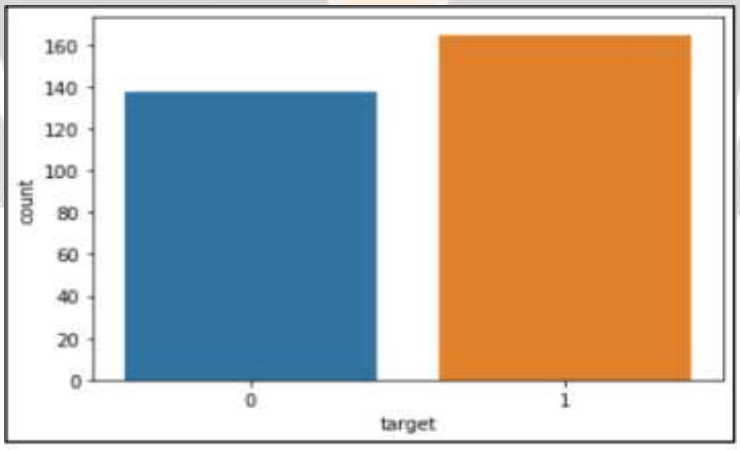


Fig.2 Positive and negative cases depending on age

The dataset contains various parameters of the patients like age, sex, blood pressure, type of chest pin etc. As shown in the figure the dataset is first preprocessed and cleaned to remove any missing values and uniform the parameters. Then it is passed on to feature extraction stage in which the features are extracted for all the individuals. The classifiers arethen utilized to classify the various features based on ground truth taken from other sources for classification. Machine learning models can be developed using numerous techniques like SVM, KNN, Artificial Neural Network etc. These machine learning models are trained based on the features to classify the dataset as belonging to either healthy or affected patients.

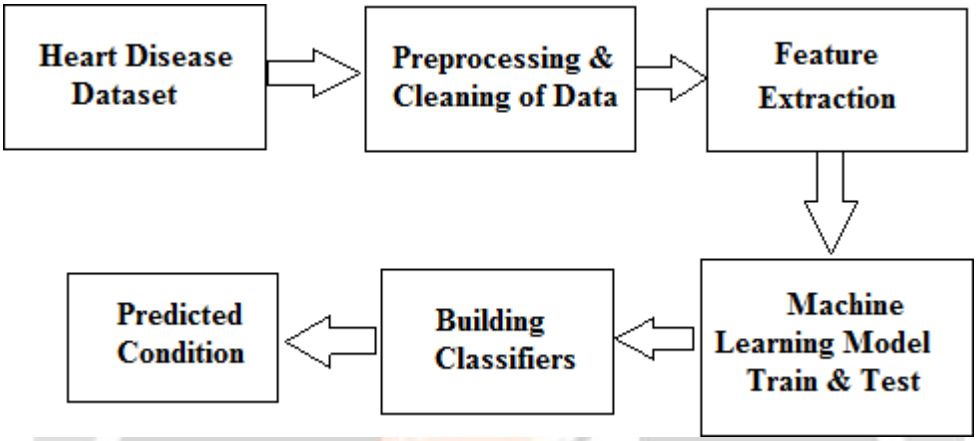


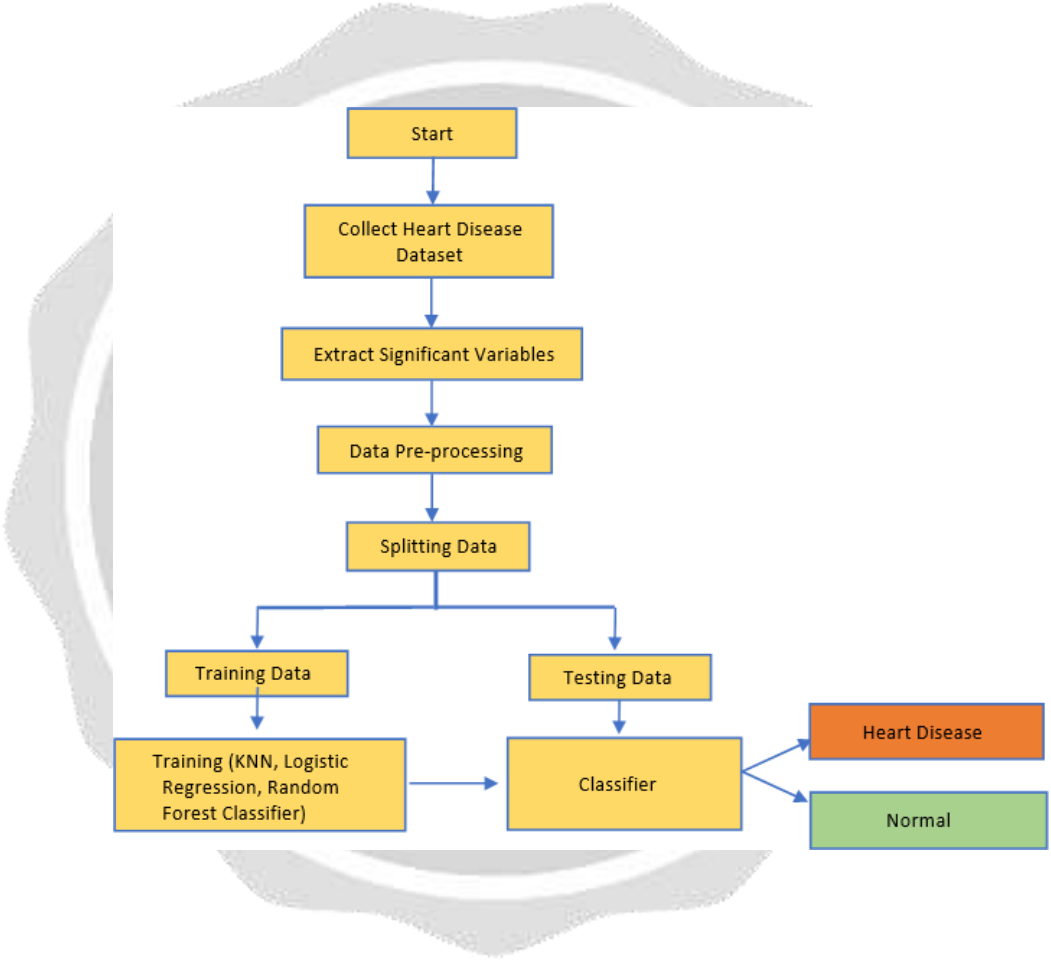
Figure 3: A general Framework for Heart Disease Classification

4. METHODOLOGY

This paper shows the analysis of various machine learning algorithms, the algorithms that are used in this paper are K nearest neighbors (KNN), Logistic Regression and Random Forest Classifiers which can be helpful for practitioners or medical analysts for accurately diagnose Heart Disease. This paperwork includes examining the journals, published paper and the data of cardiovascular disease of the recent times. Methodology gives a framework for the proposed model **[13].** The methodology is a process which includes steps that transform given data into recognized data patterns for the knowledge of the users. The

proposed methodology (Figure 1.) includes steps, where first step is referred as the collection of the data than in second stage it extracts significant values than the 3rd is the preprocessing stage where we explore the data. Data preprocessing deals with the missing values, cleaning of data and normalization depending on algorithms used **[15].** After pre-processing of data, classifier is used to classify the pre- processed data the classifier used in the proposed model are KNN, Logistic Regression, Random Forest Classifier. Finally, the proposed model is undertaken, where we evaluated our model on the basis of accuracy and performance using various performance metrics. Here in this model, an effective **Heart Disease Prediction System**

(EHDPS) has been developed using different classifiers. This model uses 13 medical parameters such as chest pain, fasting sugar, blood pressure, cholesterol, age, sex etc. for prediction **[17].**



A. *Data set information*

The main objective of this research is to develop a heart disease prediction system. The systemcan discover and extract hidden knowledge associated with diseases from a historical heart data set heart disease prediction systemaims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.

B. *Training and testing*

The training phase extracts the features (independent variables) from the dataset and the testing phase (containing dependent variables) is used to determine how the appropriate model behaves for prediction. We have divided the dataset into two sections. These are the training and testing phases. We have split the dataset into 90% training and 10% testing phase. And we have taken the random state as 1. For initializing the fixed internal random number generator, we use the random state parameter which will decide the splitting of data into train and test indices. Setting a random state will guarantee a fixed value that the same sequence of random numbers will be generated each time the code is being run. Setting random state, a fixed value will guarantee that the same sequence of random numbers is generated each time we run the code. Then we scaled the data using Standardscattered and fitted the training and testing data using 'fit. transform'.

C. *Classification used*

- Logistic regression

  Logistic Regression is an analytical modeling technique. It is used for analyzing a dataset in which there are one or more independent variables that decide a result. Logistic Regression was imported with a random state of 0. And then the training model was fitted. The testing accuracy was 87.09%

- KNN Classifier

  K-nearest neighbor algorithm is utilized for grouping and used in pattern recognition. It is widely used in predictive analysis. On the arrival of new data, the K- NN algorithm

  [8] identifies existing data points that are nearest to it. From 'sklearn. neighbors', 'Neighbors Classifier' was imported with n_neighbors = 1. Then the training model was fitted. The testing accuracy was 70.96%

- Support vector machine

  Support Vector Machine or SVM is one of the popular Supervised Learning algorithms in machine learning. The benefits of the SVM algorithm is that it creates the best suitable line or decision boundary that can separate a n-dimensional space into classesso that we can easily verify and put the new added data points in the correct category in the future. From 'sklearn', 'svm' was imported and we kept the kernel as linear and gamma as auto and C = 2. And the training model was fitted. The testing accuracy was 90.32%.

- Random forest

  Random forest classifier is a powerful supervised classification tool. RF generates a forest of classification trees from a given dataset, rather than a single classification tree. Each of these trees produces a classification for a given set of attributes. From 'sklearn.ensemble', 'Random Forest Classifier' was imported. The n_estimators is kept at 10 and random state at 0. Then the training model was fitted. The testing accuracy was 90.32%.

- Decision Tree

  The testing accuracy was 90.32%. A Decision tree is a tree shape-like diagram, where the internal nodes represent a test on an attribute, each branch denotes the outcome of the test, each leaf node denotes a class label. Decision Tree was imported where the random state was kept as 0 and then the training model was fitted. The testing accuracy was 83.87%. 6. Results Amongst all classification techniques, testing accuracy was best in the case of the random forest and SVM approach with an accuracy of 90.32%
  .

5. Results & Discussions

From these results we can see that although most of the researchers are using different algorithms such as SVC, Decision tree for the detection of patients diagnosed with Heart disease, KNN, Random Forest Classifier and Logistic regression yield a better result to out rule them **[23].** The algorithms that we used are more accurate, saves a lot of money i.e.it is cost efficient and faster than the algorithms that the previous researchers used. Moreover, the maximum accuracy obtained by KNN and Logistic Regression are equal to 88.5% which is greater or almost equal to accuracies obtained from previous researches.
So, we summarize

that our accuracy is improved due to the increased medical attributes that we used from the dataset we took. Our project also tells us that Logistic Regression and KNN outperforms Random Forest Classifier in the prediction of the patient diagnosed with a heart Disease. This proves that KNN and Logistic Regression are better in diagnosis of a heart disease. The following 'figure 4', 'figure 5', 'figure 6', 'figure 7' shows a plot of the number of patients that are been segregated and predicted by the classifier depending upon the age group, Resting Blood Pressure, Sex, Chest Pain:

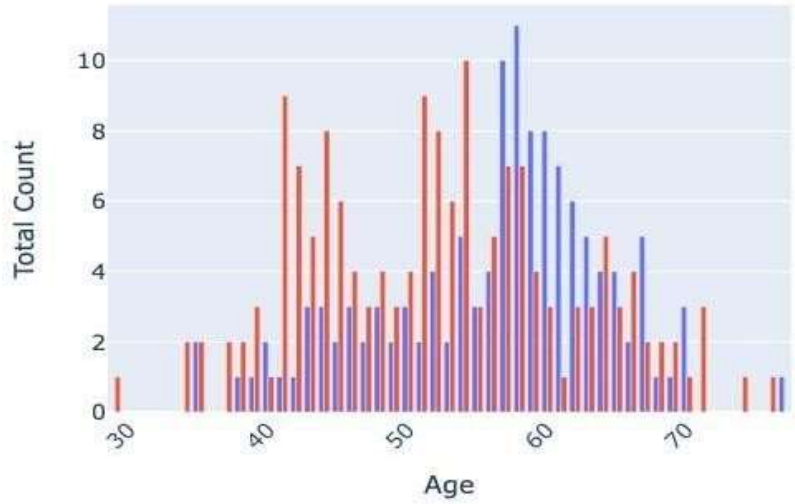


Figure 4. Shows the Risk of Heart Attack on the basis of their age.

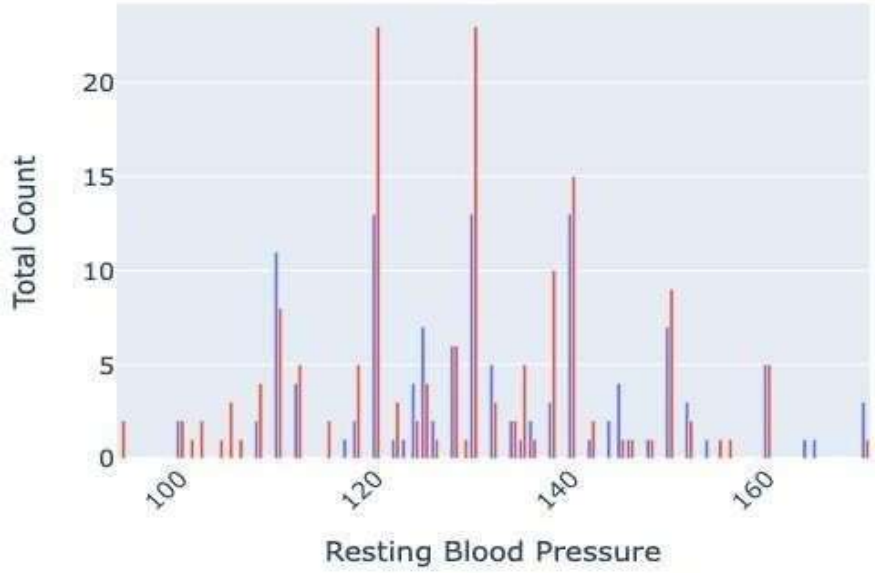
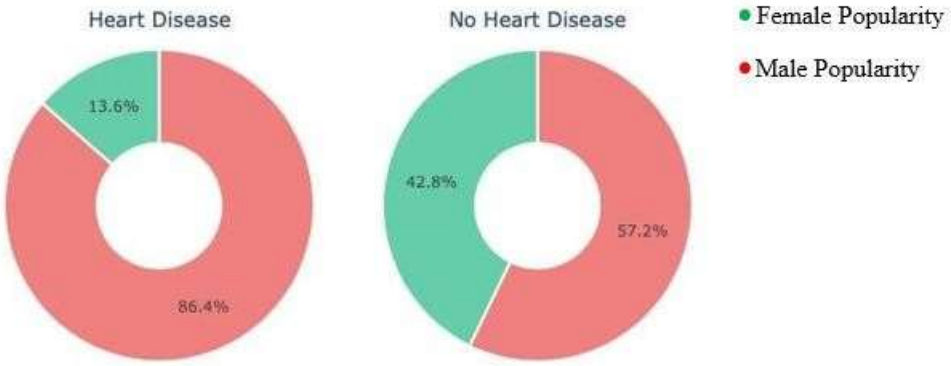


Figure 5. Shows the Risk of Heart Attack on the basis of their Resting Blood Pressure.



.**Figure 6.** Shows the patients having or not having Heart Disease on the basis of Sex

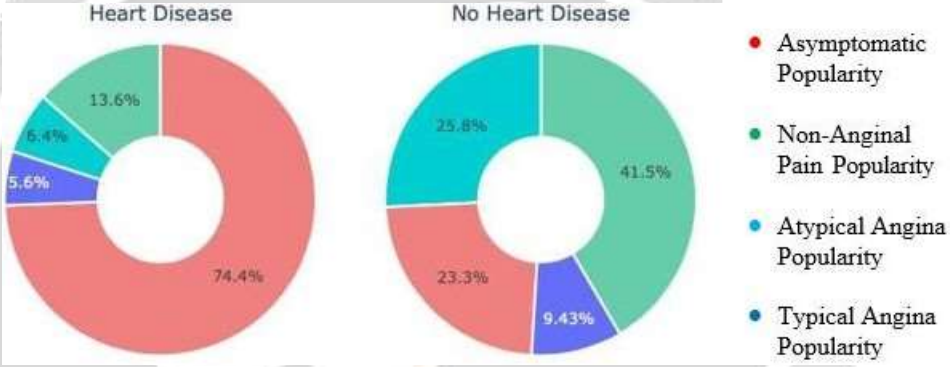


Figure 7. Shows the patients having or not having Heart Disease on the basis of type of Chest Pain

6. CONCLUSION

A cardiovascular disease detection model has been developed using three ML classification modelling techniques. This project predicts people with cardiovascular disease by extracting the patient medical history that leads to a fatal heart disease from a dataset that includes patients' medical history such as chest pain, sugar level, blood pressure, etc. This Heart Disease detection system assists a patient based on his/her clinical information of them been diagnosed with a previous heart disease. The algorithms used in building the given model are Logistic regression, Random Forest Classifier and KNN **[22].** The accuracy of our model is 87.5%. Use of more training data ensures the higher chances of the model to accurately predict whether the given person has a heart disease or not **[9].** By using these, computer aided techniques we can predict the patient fast and better and the cost can be reduced very much. There are a number of medical

databases that we can work on as these Machine learning techniques are better and they can predict better than a human being which helps the patient as well as the doctors. Therefore, in conclusion this project helps us predict the patients who are diagnosed with heart diseases by cleaning the dataset and applying logistic regression and KNN to get an accuracy of an average of 87.5% on our model which is better than the previous models

having an accuracy of 85%. Also, it is concluded that accuracy of KNN is highest between the three algorithms that we have used i.e. 88.52%. 'Figure 6' shows 44% of people that are listed in the dataset are suffering from Heart Disease.

Figure 8. Shows the total number of patients having or not having Heart Disease.

Diagnoses of cardiac disease are the sternest of challenge in the medical profession. It is based on the thorough review by medical experts of the various clinical and medical data of the patient. Because of the advances in machine learning and IT, researchers and medical experts are interested in creating a highly accurate, efficient and supportive predictive framework for the prediction in heart disease. Data analysis and machine learning methods have been used to forecast heart disease events and have summarized. Determine each algorithm's prediction output and apply the method proposed for the area needed. To boost the exact performance of algorithms, using more specific methods of feature selection. If patients are diagnosed with the specific type of heart disease, there are many treatments methods

7. REFERENCES

[1] Mohan, Senthilkumar & Thirumalai, Chandra Segar & Srivastava, Gautam. (2019). Effective Heart Disease Prediction using Hybrid Machine Learning Techniques. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2923707.

[2] A. N. Repaka, S. D. Ravikanti and R. G. Franklin, "Design And Implementing Heart Disease Prediction Using Naives Bayesian," 2019 3rd International Conference on Trendsin Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 292-297.doi: 10.1109/ICOEI.2019.8862604.

[3] A. Ed-Daoudy and K. Maalmi, "Real-time machine learning for early detection of heart disease using big data approach," 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez, Morocco, 2019, pp. 1-5. doi: 10.1109/WITS.2019.8723839.

[4] Amin Ul Haq, J. P.Li ,M.H.Memon,Shah Nazir and Ruinan Sun," A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning
Algorithms", Wearable Technology and Mobile Applications for Healthcare, Volume 2018 |Article ID 3860146 | 21 pages | https://doi.org/10.1155/2018/3860146.

[5] M. S. Satu, F. Tasnim, T. Akter and S. Halder, "Exploring Significant Heart Disease Factors based on Semi Supervised Learning Algorithms," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, 2018, pp. 1- 4.doi: 10.1109/IC4ME2.2018.8465642.

[6] R. Sharmila, S. Chellammal, "A conceptual method to enhance the prediction of heart

diseases using the data techniques", International Journal of Computer Science and Engineering, May 2018.
[7] Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis theOccurrence of Heart Disease Using Data Mining Techniques",International Journal of Pure and Applied Mathematics, 2018.
[8]"Prediction of Heart Disease using Machine Learning Algorithms" Krishnan J Santhana and S Geetha ICIICT |Year :2019| Conference Paper | Publisher: IEEE.

[9]. "Prediction of Heart Disease using Machine Learning". Aditi Gavhane, Gouthami Kokkula, Isha Panday and Kailash Devadkar, Proceedings of the 2nd International conference on Electronics Communication and Aerospace Technology(ICECA) |Year :2018| Conference Paper | Publisher: IEEE.

[10]. "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" Senthil Kumar, Mohan Chandrasegar Thirumalai and Gautam Srivastva |Year :2019| Conference Paper | Publisher: IEEE.

[11]. "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, Himanshu Sharma and M A Rizvi |Year :2019| Conference Paper | Publisher: IEEE .

[12] "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools"International Journal of Scientific Research in Computer Science Engineering and Information Technology IJSRCSEIT , M. Nikhil Kumar K. V. S. Koushik and K. Deepak|Year:2019| Conference Paper | Publisher: IEEE.

[13]"Heart Diseases Prediction using Data Mining Techniques: A survey" Amandeep Kaurand Jyoti Arora International Journal of Advanced Research in Computer Science IJARCS
|Year :2019| Conference Paper

.