Heart Disease Prediction Using Machine Learning Algorithm

Dr.R.M.S.Parvathi[1], Supreeta S[2], Sudeeson S[3], Udhaya Sankar P[4]
[1]HOD, Dept of CSE
[2, 3, 4]Dept of CSE
[1, 2, 3, 4] Sri Ramakrishna Institute of Technology

Abstract- *Heart attack disease is one of the leading causes of the death worldwide. In today's common modern life, deaths due to the heart disease had become one of major issues, that roughly one person lost his or her life per minute due to heart illness. Predicting the occurrence of disease at early stages is a major challenge nowadays. Machine learning when implemented in health care is capable of early and accurate detection of disease. In this work, the arising situations of Heart Disease illness are calculated. Datasets used have attributes of medical parameters. The datasets are been processed in python using ML Algorithm i.e., Random Forest Algorithm. This technique uses the past old patient records for getting prediction of new one at early stages preventing the loss of lives. In this work, reliable heart disease prediction system is implemented using strong Machine Learning algorithm which is the Random Forest algorithm. Which read patient record data set in the form of CSV file. After accessing dataset the operation is performed and effective heart attack level is produced. Advantages of proposed system are High performance and accuracy rate and it is very flexible and high rates of success are achieved.*

Keywords- Heart Prediction, Machine Learning, Supervised Learning, Random Forest.

I. INTRODUCTION

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects. In the section of data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint 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 reduces the complications. Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. 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 algorithm. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analyzing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

Background History:

Heart disease is considered one of the top preventable causes of death in India. Some genetic factors can contribute, but the disease is largely attributed to poor lifestyle habits. Among these are poor diet, lack of regular exercise, tobacco smoking, alcohol or drug abuse, and high stress. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint 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 reduces the complications.

Problem Statement:

Detection of heart diseases in human is a major challenge. Early detection of cardiac diseases can decrease the mortality rate and overall complications. There are instruments available which can predict heart disease but either they are expensive or not efficient to calculate the

chance of the heart disease that may occur in human. It is not possible to monitor patients regularly 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. The hidden patterns can be used for health diagnosis in medicinal data. From that we can reduce the morality rate and overall complication.

Scope of the Project:

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 algorithm. This model predicts the likelihood of patients getting heart disease and enables significant knowledge example relationship between medical factors related to heart disease and patterns to be established. 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.Heart disease prediction has the potential to benefit stakeholders such as the government and health insurance company itcan identify patients at risk of disease or health conditions. After training this model with large number of data set, this can be used by hospitals and it will help them to predict heart disease patients from their datasets.

Existing System:

In the existing systems, heart disease prediction system is developed using various algorithms but has some disadvantages. All the existing systems uses many algorithms and hence are complex and also with less accuracy. The existing systems are build on some machine learning algorithms like decision tree, naïve based algorithm, K nearest algorithm, logistic regression, support vector machine and linear regression. But all these algorithms when implemented separately gives very less accuracy. So to achieve a efficient model combinations of the above mentioned algorithms are used to build a hybrid model. But it ends up with a complex algorithm with reasonable accuracy. So to make these complex system simple we are planned to use Random Forest algorithm which is comparatively simple but results with higher accuracy.

Limitations

The major drawbacks and limitations in the existing models are listed below:

- Integrated algorithms are used to build hybrid models which is a very complex and expensive task.
- Two are more machine learning algorithms are combined together increasing the complexity of the system.
- Datasets are taken from a single area or region of concern. Since this machine learning model deals with medical use it should make sure that it works well for people from all over the world. So collecting dataset from a single region is not good approach.

II. SYSTEM SPECIFICATION

2.1 Dataset Description:

The dataset is publicly available on the Kaggle Website at 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 Separated Value) format which is further prepared to data frame as supported by pandas library in python. The chosen dataset has 4 databases concerning heart disease diagnosis. All attributes are numeric-valued. There are several missing values, so the dataset has to be pre-processed.

The data was collected from the four following locations:
1. Cleveland Clinic Foundation (cleveland.data)
2. Hungarian Institute of Cardiology, Budapest (hungarian.data)
3. V.A. Medical Center, Long Beach, CA (long-beach-va.data)
4. University Hospital, Zurich, Switzerland (switzerland.data)

Feature	Data type
Age	Numeric
Sex	Nominal
Chest Pain	Nominal
Resting BP	Numeric
Cholesterol	Numeric
Fasting BP	Nominal
Resting ECG	Nominal
Max heart rate	Numeric
Exercise Angina	Nominal
Old peak	Numeric
ST Slope	Nominal
Target	Nominal

2.3 Methodology and Flowchart:

First stage of the process is the collection of data and selecting the important attributes. Then the required data is preprocessed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data.

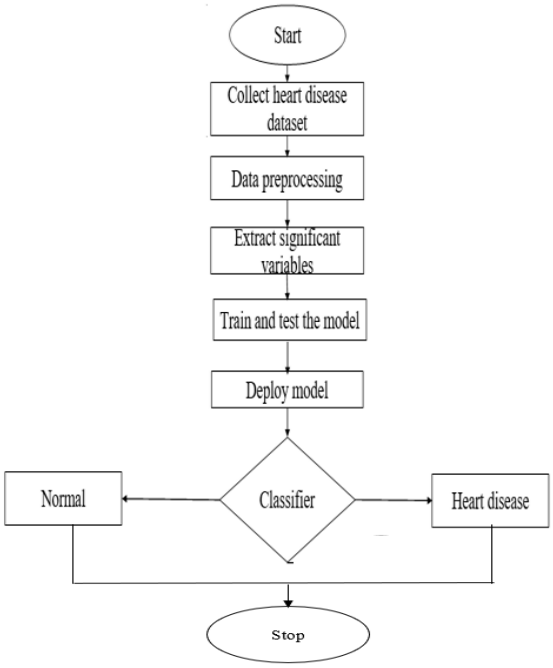


Fig 2.1 (Flowchart)

2.4 Data Cleaning and Pre-processing:

Data cleaning and pre-processing are the methods that eliminate noise from the data and put the raw data into a suitable form so that a machine learning algorithm may easily train on it. Several of the cells in the downloaded csv file are filled with null values. The column's mean values must be used to fill up these null values.

```
age                 0
sex                 2
chest_pain_type    10
resting_bp          5
cholesterol        50
fasting_blood_sugar 2
resting_ecg         3
max_heart_rate     48
exercise_angina     0
oldpeak            16
ST_slope            3
target              0
```

Fig 2.2 (Number of Null Valued Cell in each columns)

2.5 Exploratory Data Analysis:

Exploratory data analysis (EDA), which frequently makes use of data visualization techniques, is used to examine and summarize large data sets. It makes it simpler to find patterns, identify anomalies, test hypotheses, or verify assumptions by determining how to alter data sources to achieve the answers you need. EDA offers a deeper knowledge of data, variables, and the relationships between them and is generally used to examine what data might disclose beyond the formal modelling or hypothesis testing assignment. It can also assist in determining the suitability of the statistical methods you are contemplating using for data analysis.

EDA for the collected dataset is done and the output results are attached below.

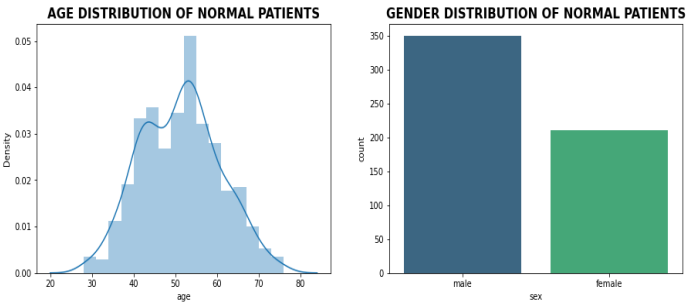


Fig 2.3 (Age and gender wise distribution of normal patient)

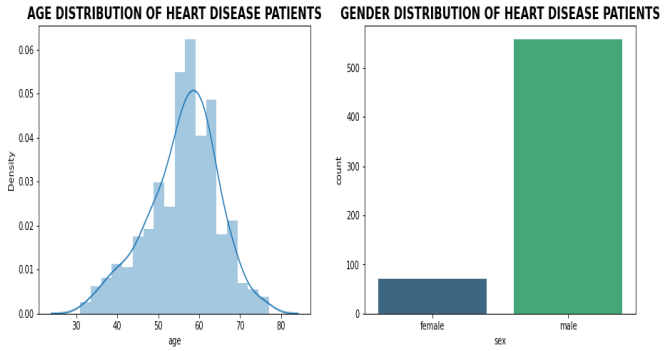
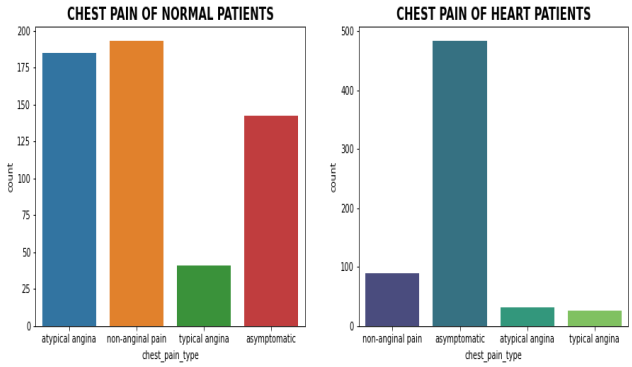


Fig 2.4 (Age and gender wise distribution of heart disease patients)

As we can see from above plot more patients accounts for heart disease in comparison to female whereas mean age for heart disease patients around 58 to 60 years.

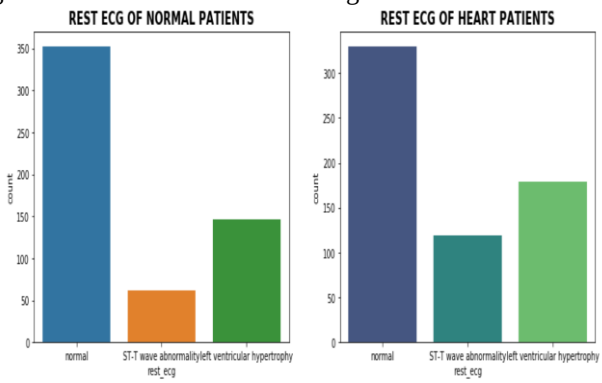


chest_pain_type	target 0	1
asymptomatic	25.31	76.91
atypical angina	32.98	4.94
non-anginal pain	34.4	14.17
typical angina	7.31	3.98

Fig 2.5(Chest pain distribution of normal and heart patients)

As we can see from above plot 76% of the chest pain type of the heart disease patients have asymptomatic chest pain.

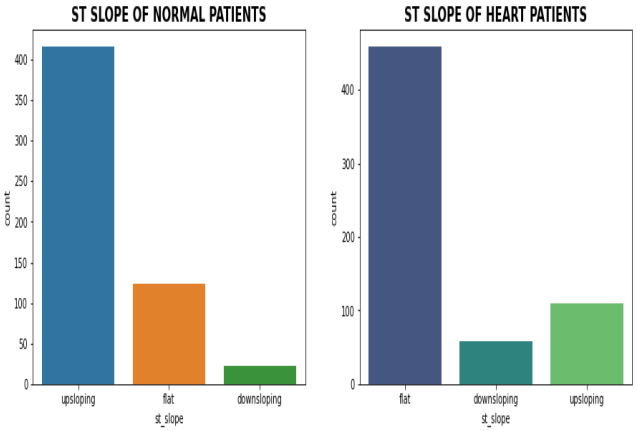
Asymptomatic heart attacks medically known as silent myocardial infarction (SMI) annually accounts for around 45-50% of morbidities due to cardiac ailments and even premature deaths in India. The incidences among middle aged people experiencing SMI is twice likely to develop in males than females. The symptoms of SMI being very mild in comparison to an actual heart attack; it is described as a silent killer. Unlike the symptoms in a normal heart attack which includes extreme chest pain, stabbing pain in the arms, neck & jaw, sudden shortness of breath, sweating and dizziness, the symptoms of SMI are very brief and hence confused with regular discomfort and most often ignored.



rest_ecg	target 0	1
ST-T wave abnormality	11.05	18.95
left ventricular hypertrophy	26.02	28.5
normal	62.92	52.55

Fig 2.6(Distribution of rest ECG)

An electrocardiogram records the electrical signals in your heart. It's a common test used to detect heart problems and monitor the heart's status in many situations. Electrocardiograms — also called ECGs or EKGs. but ECG has limits. It measures heart rate and rhythm—but it doesn't necessarily show blockages in the arteries. Thats why in this dataset around 52% heart disease patients have normal ECG.



st_slope	target 0	1
downsloping	3.92	9.39
flat	21.93	73.09
upsloping	74.15	17.52

Fig 2.7(ST slope distribution of normal and heart patients)

The ST segment /heart rate slope (ST/HR slope), has been proposed as a more accurate ECG criterion for diagnosing significant coronary artery disease (CAD) in most of the research papers.

As we can see from above plot upsloping is positive sign as 74% of the normal patients have upslope where as 72.97% heart patients have flat sloping.

2.5.2 Correlation with Heat Map:

A heat map is a data visualization technique that shows magnitude of a phenomenon as color in two dimension. The variation in color may be due to hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space. There are two fundamentally different categories of heat map.

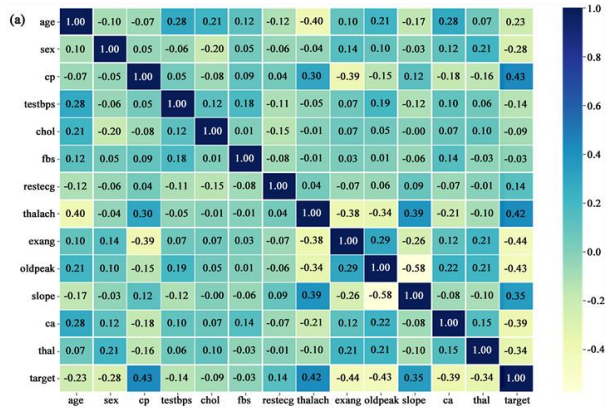


Fig 2.8 (Heat Map for the dataset)

2.6 Random Forest Algorithm:

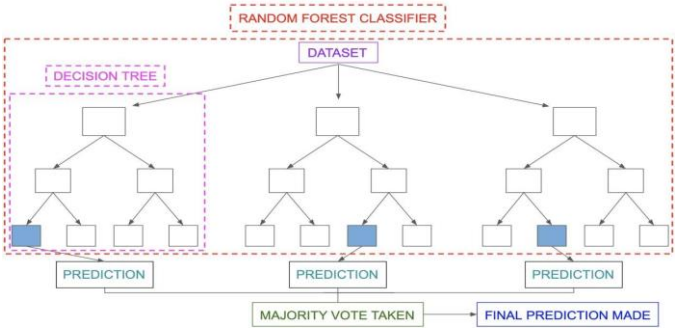


Fig 2.9 (Random Forest Classifier)

Random Forest is a classifier consisting of a collection of tree-structured classifiers {h(x, Θk) k=1,2, ….}, where the {Θk } are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.

An ensemble of decision trees is produced using Random Forest. Breiman used the randomization strategy, which integrates well with bagging or random subspace approaches, to produce diversity among base decision trees. Breiman used the Random Forest and performed the following procedures to create each individual tree: If there are N records in the training set, then N records from the original data are picked at random but with replacement. This is known as a bootstrap sample. This sample will serve as the tree's training set. There will be an increasing number m forest if there are M input variables. Every tree is developed to its full potential. Pruning is not done. In this way, multiple trees are induced in the forest; the number of trees is pre-decided by the parameter Ntree. The depth of the tree can be controlled by a parameter nodesize which is usually set to one.

Once the forest is trained and built as explained previously, to classify a new instance, it is run across all the trees generated in the forest. Each tree provides a category that applies to all the trees that have developed in the forest. A new instance is classified by each tree, and this classification is recorded as a vote. The classes with the most votes overall (majority voting) are proclaimed as the classifications for new instances after adding the votes from each tree.

About one-third of the original instances are lost throughout the forest-building process when a bootstrap sample set is created by sampling with replacement for each tree. OOB (Out Of Bag) data refers to this collection of instances. The process of estimating each individual tree's error in the forest using its own OOB data collection is known as OOB error estimation. The Random Forest method also has a built-in capability to determine the proximity and relevance of variables. Outliers and missing values are replaced using the proximities.

	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	Random Forest	0.932773	0.916667	0.960317	0.901786	0.937984	0.931052	2.423103	0.865792
1	KNN	0.752101	0.781513	0.738095	0.767857	0.759184	0.752976	8.935191	0.505076
2	SVC	0.861345	0.849624	0.896825	0.821429	0.872587	0.859127	4.997649	0.722024
3	Decison tree	0.869748	0.874016	0.880952	0.857143	0.877470	0.869048	4.694762	0.738488
4	Logistic Regression	0.848739	0.846154	0.873016	0.821429	0.859375	0.847222	5.451981	0.696223

Fig 2.10 (Comparing with KNN, SVC, LR, Decision tree)

.

III. CONCLUSION

Conclusion:

Random Forest algorithm is an efficient algorithm which is an ensemble learning method for regression and classification techniques. The algorithm constructs N of Decision trees and outputs the class that is the average of all decision trees output. So accuracy of prediction at early stages is achieved effectively. Processing of healthcare data i.e., data related to heart will help in early detection of heart disease or abnormal condition of heart which results in saving of long term deaths. Heart disease prediction is a major challenge in the present modern life. With this application if the patient/user is away from reach of doctor, he/she can make use of the application in prediction of disease just by entering the

report values. And can proceed further whether to consult a doctor or not.

Future Works:

In future this can extended by updating some features like, if the user is effected with heart disease all his family members will be notified with a message in early. And also the information should be passed to the nearest hospital. Another feature is there should be online doctor consultation with the nearest doctor available.

REFERENCES

[1] S. Mohan et al, "Effective heart disease prediction using hybrid machine learning techniques," Institute of electrical and electronics engineers(IEEE), Vol.7, July 2019.

[2] Jian Ping Le et al, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," Institute of electrical and electronics engineers(IEEE), Vol.8, June 2020.

[3] Sarria E et al, "HDPF: Heart Disease Prediction Framework Based on Hybrid Classifiers and Genetic Algorithm," Institute of electrical and electronics engineers(IEEE) , Vol.9, Oct 2021.

[4] Tsatsral Amarbayasgalan et al, "An Efficient Prediction Method for Coronary Heart Disease Risk Based on Two Deep Neural Networks Trained on Well-Ordered Training Datasets," Institute of electrical and electronics engineers(IEEE) , Vol.9, Oct 2021.

[5] Hira Fathima et al, "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV," Institute of electrical and electronics engineers(IEEE), Vol.10, April 2022.

[6] Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," 2022 Advances in Science and Engineering Technology International Conferences (ASET), 2022, pp. 1-6.

[7] Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 45.

[8] 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. 177-1812-457,.

[9] Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275-1278.