HeartCare - A Heart Disease Prediction System based on Machine Learning

[1]Gaurav Singh, [2]Hemlata Pant, [3]Pratiksha Singh, [4]Gaurav Kumar, [5]Krishn Pratap Singh

[1,2,3,4,5]Department of CSE, Babu Banarasi Das National Institute of Technology and Management, Lucknow, India

Abstract : Heart-related diseases are one of the major causes of deaths across the globe. On an average, one person dies every minute across the world due to cardiovascular diseases. Prediction of heart disease is a difficult task and requires expertise. Therefore, an automated system capable of predicting heart diseases would enhance medical efficiency and reduce the overall cost of treatment. This paper aims to present the concept of the functional model of a heart disease prediction system. This system is highly efficient, built on Django platform and uses a machine learning model. It takes various medical inputs from the user and predicts whether the person has heart disease or not.

IndexTerms **- Machine Learning, Random Forest algorithm, Django, Framework, Heart Disease, Cardiovascular Disease, Cleveland Dataset, Heart Disease Prediction System**

I. INTRODUCTION

The research presented in this paper primarily focuses on various machine learning and data mining approaches for heart disease prediction. Human heart is one of the major organs of the human body. Its task is to circulate blood across the body. Cardiovascular diseases include a variety of disorders that have an effect on our heart. They include problems in blood vessels including coronary artery disease, heart rhythm issues (arrhythmias), and congenital heart abnormalities, among others.

The terms "heart disease" and "cardiovascular disease" are considered synonymous and refer to the symptoms related to heart. Cardiovascular abnormalities are characterized by restricted blood vessels caused by fat deposition, which can result in a heart attack, chest pain (angina), or stroke. Other disorders that affect the muscle, valves, rhythm, or efficiency of our hearts are also classified as heart disease.

CVDs are amongst the major causes of deaths across the globe. A large number of people die annually from CVDs as compared to any other cause. An estimated ~18 million people died from CVDs in the year 2016, which represents approximately 30% of all global deaths. Out of these 80-85% were caused due to heart attacks and strokes.

One of the major challenges that today's healthcare system faces is to provide best quality health services and diagnosis and that too at lower costs. While heart diseases are one of the major causes of death in the world, there have also been some cases where heart diseases have been cured and prevented effectively. The ability to treat an illness effectively is largely dependent on its early identification. The suggested research aims to detect heart illnesses early on so that adequate treatment can be administered.

Machine learning is one of the major subdivisions of artificial intelligence. This division primarily focuses on designing systems that can learn and make predictions based on the previous experience. The importance of machine learning and data science in health care units is due to their ability to process extremely huge datasets which is beyond the scope of human capability, and also convert analysis into clinical insights that can help physicians in planning and providing care, leading to better outcomes at much lower costs, and an enhanced way of treatment. The main aim of this research is to develop an application for the prediction of heart diseases. The system will be able to discover the pattern and hidden knowledge from a data set associated with heart diseases. Heart disease prediction system aims to harness the power and potential of machine learning and data science on clinical data set to provide assistance in the prediction of the heart diseases.

II. RELATED WORK

Various research works have been carried out to predict heart disease using the dataset of UCI Machine Learning repository. Several strategies have been used to achieve various levels of accuracy, which are detailed below:

Different machine learning algorithms have been studied by Chaitrali S. Dangare and Sulabha S. Apte et. al. [1] for the classification of heart disease. Researches have also been carried out to study Neural Networks, Decision Trees and Naive Bayes algorithms and their accuracy scores are also compared in which Neural Networks always performed better than others.
A study and comparison of the results of SVM, Decision Trees, Naive Bayes and Logistic regression algorithm has been done by Rishabh Magar, Rohan Memane, Suraj Raut et. al. [2]. Logistic Regression algorithm proved to be more efficient than the other three having an accuracy score of ~82%. Decision tree, Naive Bayes and SVM had accuracies of ~80%.

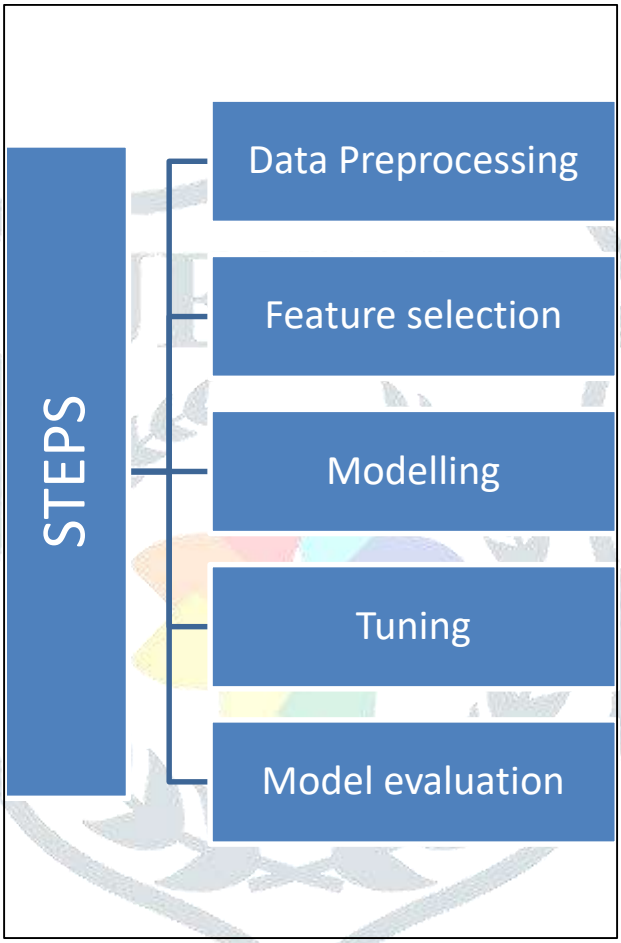
Study and use of the pre-processed dataset to carry out the tests and application of few machine learning algorithms has been done by Apurb Rajdan, Milan Sai and Dundigalla Ravi et. al. [3]. The confusion matrix has been used to calculate the performance measures. Confusion Matrix is a matrix that is used to describe the performance of the model. Here, Random Forest has been found to perform much better than other algorithms.

Random Forest, Decision Trees, and Naive Bayes are the three prominent algorithms that have been studied and analyzed by H. Benjamin Fredrick David and S. Antony Belcy et al. [4]. According to the research, Random Forest outperformed Decision Tree and Naive Bayes in terms of performance.

A comparative study of Decision Trees, Naive bayes and clustering techniques has been done by Jyoti Soni, Ujma Ansari et al. [5]. They primarily focused on the usage of different algorithms for prediction and using combinations of different target attributes for intelligent and effective prediction.

III. Proposed Model

The proposed work explores the Random Forest Classifier algorithm and assesses its performance in predicting heart disease. Random Forest is one of the easiest and simplest algorithms to implement. It can also be tuned to get even more accurate results. The user enters the input values from the patient's medical test report. The input data is fed into model via a backend server. Model predicts the probability of having heart disease as 0 or 1. Fig. 1 shows the entire process involved.



(Fig. 1: Steps for training model)

A. Data Collection and pre-processing

The dataset used is the clinical dataset of Cleveland Heart Clinic Foundation and is freely available on UCI machine learning repository. The original dataset consists of a total of 76 attributes but this paper uses a subset of only 14 important attributes. It has no missing values. Table 1 below provides a quick overview of the 14 attributes used in the proposed approach.

Table 1: Description of attributes

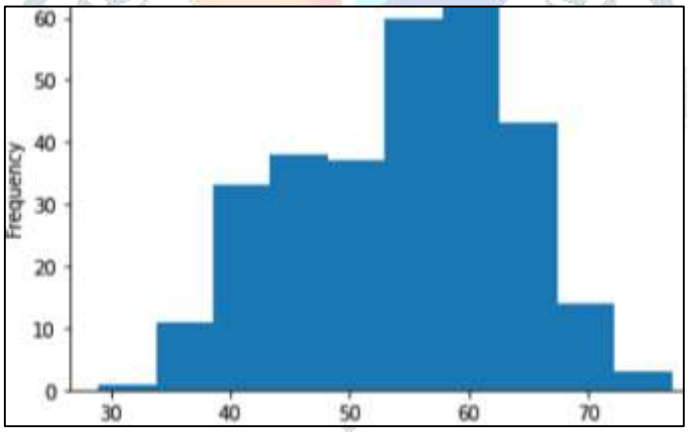
S. no.	Attribute	Description
1.	age	Age (years)
2.	sex	1 : male, 0 : female
3.	cp	Chest pain type
4.	trestbps	Resting blood pressure (mm Hg)

5.	chol	Serum cholesterol (mg/dl)
6.	fbs	Fasting blood sugar > 120 mg/dl (1 : true, 0 : false)
7.	restecg	Resting electrocardiographic results
8.	thalach	Max heart rate achieved
9.	exang	Exercise induced angina (1 : yes, 0 : no)
10.	oldpeakST	Depression induced by exercise relative to rest
11.	slope	Slope of peak exercise ST segment
12.	ca	Number of major vessels coloured by fluoroscopy (0-3)
13.	thal	Thalassemia
14.	target	Result (0 or 1)

B. Feature selection

As a part of this step, only 14 out of 76 attributes have been selected that form the basis for the training of machine learning model.

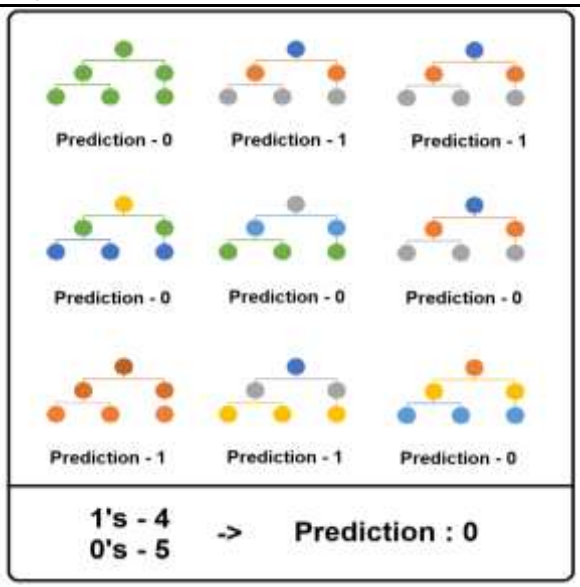
Fig. 2 analyses the age column of dataset using a histogram.



(Fig. 2: Distribution of age)

C. Model training

Random forest, as the name suggests, is a collection of a large number of independent decision trees that operate together. Every tree of the random forest gives a prediction for the class and the class which gets the majority votes becomes the prediction of the model. The basic concept that works behind random forest is really simple - the crowd's wisdom. In data science terms, the main reason behind random forest working so well is that – "A large number of relatively uncorrelated models (trees) that operate as a committee can outperform any of the individual models."



(Fig. 3: Random Forest Algorithm)

D. Tuning

It is the process of changing the settings or hyperparameters of trained model to improve the accuracy score while avoiding overfitting and underfitting condition. By trying different random states and heights of tree, accuracy score of random forest increases significantly.

E. Model evaluation

The performance of model is checked by using 5-fold cross-validation. Different parameters like accuracy, precision, recall, f1-score are observed on each fold of dataset using classification report.

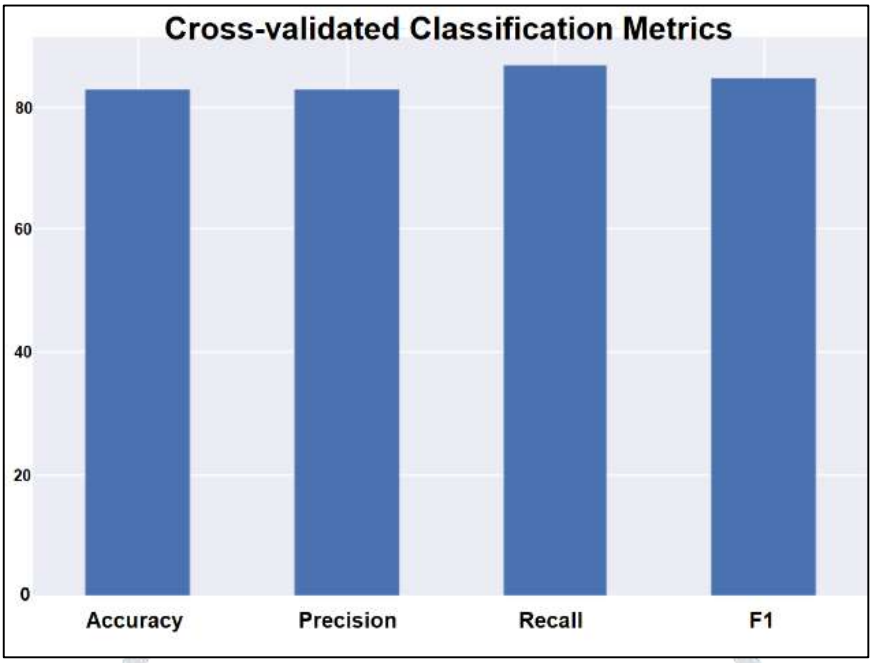
Fig. 3 describes the classification report of the model.

	precision	recall	f1-score	support
0	0.92	0.85	0.88	27
1	0.89	0.94	0.91	34
accuracy			0.90	61
macro avg	0.90	0.90	0.90	61
weighted avg	0.90	0.90	0.90	61

(Fig. 4: Classification report)

The model achieved an accuracy score of 90% which is outstanding.

Fig. 5 is the comparison of various performance metrics.

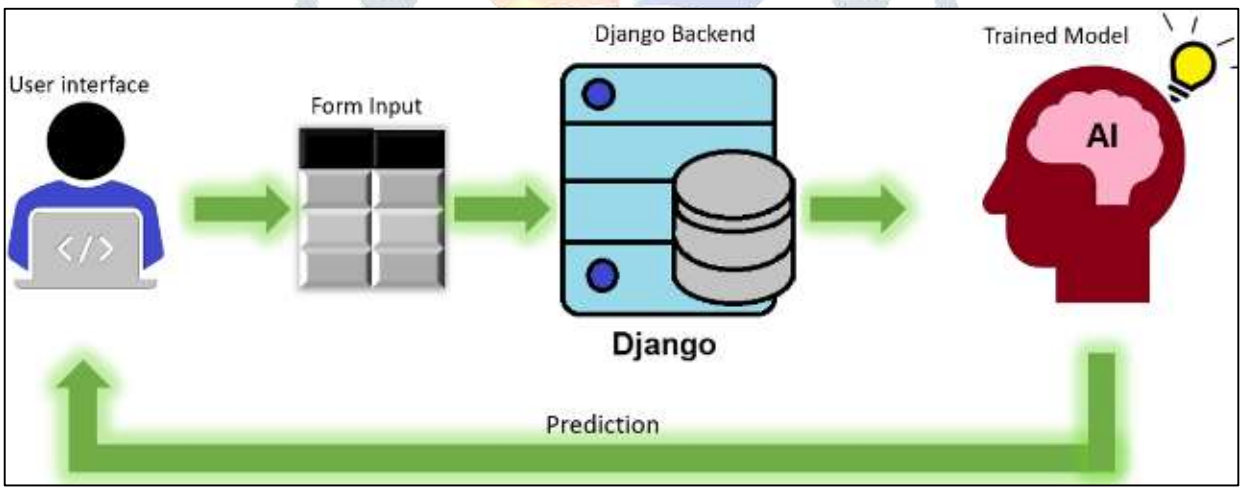


(Fig. 5: Comparison of accuracy, precision, recall and f1 score)

IV. ARCHITECTURE

Fig. 6 depicts the structure of proposed model.

The user enters data through a form interface. This data is fed into the machine learning model. The model uses Random Forest Classifier algorithm to get prediction. It predicts whether the person has heart disease or not. The output is the displayed at the front end in a very user-friendly manner.



(Fig. 6: Architecture)

V. CONCLUSION

Our purpose is to maximize our ability to predict the presence of cardiac problems. Thirteen input attributes from the Cleveland dataset are used to get more accurate results than ever before.

From the results, it has been very clear that Random Forest classifier gives highly accurate results and that too with simplicity. This application would help the end users to get a prediction whether they have heart disease or not. Since heart disease is one of the major cause of deaths not only in India but across the world, this application will have a profound impact on the healthcare systems.

The suggested system is a graphical user interface (GUI)-based, user-friendly, scalable, dependable, and expandable system. By delivering early diagnosis, the proposed operating methodology can also help to reduce treatment costs. The model can also be used as a teaching tool for medical trainees and will be available to cardiologists as a soft screening and diagnostic tool. Physicians can also use it to make initial patient diagnoses.

VI. FUTURE WORK

In order to improve the scalability, dependability, and effectiveness of this prediction system, numerous upgrades might be investigated. We can enhance this system with an even more reliable accuracy score in future since we proposed a fundamental system.

Handling other labels of the dataset in the prediction process can also increase the model's performance greatly, and this could be another interesting research topic. Because the heart database has a high complexity, identifying significant features is difficult and should be explored as a future study issue.

This system can also be further expanded. More data needs to be collected from various healthcare units in order to get highly accurate results. Various other data mining techniques can be used for predication like Clustering, Neural Networks etc. to improve the accuracy of prediction.

Data mining techniques can be used to mine extremely large amount of data available from the healthcare industry database that has not been utilized till date.

VII. ACKNOWLEDGEMENT

First and foremost, we are thankful to the BBDITM, Lucknow, department of CSE and to Dr. Diwakar Yagyasen, Head of Department, Computer Science and Engineering, BBDITM, Lucknow.

A special word of gratitude to Ms. Hemlata Pant, Associate Professor, Computer Science and Engineering Department, BBDITM, Lucknow for her continuous guidance and support for our project work.

REFERENCES

[1] Chaitrali S. Dangare and Sulabha S. Apte et al. "Improved Study of Heart Disease prediction System using Data Mining Classification Techniques", International Journal of Computer Applications, June, 2012.

[2] Rishabh Magar, Rohan Memane, Suraj Raut et al. "Heart Disease Prediction using Machine Learning", JETIR, volume 7.

[3] Apurb Rajdan, Milan Sai, Dundigalla Ravi et al. "Heart disease prediction using machine learning", International Journal of Engineering Research and Technology.

[4] H. Benjamin Fredrick David and S. Antony Belcy et al. "Heart disease prediction using data mining techniques", ICTACT journal, 2018.

[5] Jyoti Soni, Ujma Ansari et al. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications, March, 2011.

[6] Hanen Bouali and Jalel Akaichi et al. "Comparative study of Different classification techniques", 2014, 13th International Conference on Machine Learning and Applications.

[7] R. Kavitha and E. Kannan et al. "An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining", 2016

[8] Houda Mezrigui, Foued Theljani and Kaouther Laabidi et al. "Decision Support System for Medical Diagnosis using a Kernel-Based Approach", ICCAD'17

[9] M. A. Jabbar , B.L Deekshatulu and Priti Chandra et al. "Alternating decision trees for early diagnosis of heart disease", Proceedings of I4C 2014.

[10] Heart Disease Data Set [Database] (http://archive.ics.uci.edu/ml/datasets/Heart+Disease CHDD)