International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Machine Learning Based Heart Disease Prediction System

Mrs. Hemapriya BC[1], Keerthana B G[2], Keerthi C R[3], Pallavi R[4]

[1]Assistant Professor, CSE Department, RL Jalapa Institute, RL Jalapa Instituteof Technology
[2,3,4] Student, CSE Department, RL Jalapa Institute, RL Jalapa Institute of Technology

ABSTRACT

Heart disease is one of the leading causes of death globally. In today's modern lifestyle, heart-related conditions have become a significant concern, with statistics showing that approximately one person dies every minute due to heart-related issues. Early detection of heart disease remains a major challenge in the medical field. However, machine learning in healthcare has proven effective in enabling early and accurate diagnosis of such conditions. In this work, a system is developed to predict the likelihood of heart disease based on various medical attributes. The dataset used consists of historical patient records with relevant health parameters. Using Python, the data is processed with the Random Forest algorithm—a robust and efficient machine learning technique. This approach allows the system to learn from past cases and predict the risk for new patients, thereby supporting early diagnosis and timely treatment. The prediction system reads patient data from a CSV file, processes the information, and outputs the estimated risk level of heart disease. The proposed method offers high accuracy, strong performance, and flexibility. As a result, it achieves a high success rate in predictions, making it a reliable tool for aiding in the prevention of heart disease and saving lives.

Keywords: ML: Machine Learning, Vector Quantisation, Questionnaire, CSV: Comma- Separated Values, Random Forest algorithm, Decision Trees.

1. INTRODUCTION

Heart disease significantly impacts the normal functioning of the heart and remains one of the major causes of death worldwide. According to a survey conducted by the World Health Organisation (WHO), around 10 million people are affected by heart-related conditions, leading to the loss of many lives annually. One of the key challenges faced by the healthcare sector today is the early prediction and diagnosis of heart disease after the onset of initial symptoms.

Medical records and patient histories generate massive amounts of data, but real-world datasets are often incomplete or inconsistent. In earlier times, accurately predicting heart disease at its early stages and providing timely treatment for every patient was difficult under such conditions [2].

Numerous researchers must build models capable of predicting heart disease in its early stages. However, no model has proven to be fully accurate, and each proposed system has its limitations. For example, Shen et al. introduced a system based on self- administered questionnaires, where users input their symptoms manually, and predictions were made based on those responses. This system was based on data collected using the Seattle Angina Questionnaire (SAQ).

In another study, Chen et al. proposed a method that used Vector Quantisation, a technique in artificial intelligence, for classification and prediction. Neural networks were trained using backpropagation to develop the prediction system. Although the system achieved about 80% accuracy during testing, it was time-consuming and less effective in real-world applications due to lower accuracy.

To overcome these limitations, we propose the use of the **Random Forest algorithm**, which is known for its efficiency and high accuracy. Machine learning has gained major importance in today’s world, especially in the healthcare industry. Among its many applications, prediction plays a crucial role.

In this work, we focus on predicting heart disease by analysing medical datasets using machine learning. The system processes patient data from historical records and new user inputs to assess the likelihood of heart disease, aiming to deliver faster, more reliable, and accurate results, ultimately helping reduce the risk of life loss.

2. RELATED STUDY

In existing systems, heart disease prediction has been developed using various machine learning algorithms. However, many of these approaches come with certain limitations. Different algorithms have been applied to improve the accuracy of prediction models, yet none have proven to be completely efficient in all scenarios. Machine learning techniques generally follow a series of common steps, which are outlined below:

A. Preprocessing

Most medical datasets contain missing or undefined values, commonly represented as Nan ("Not a Number"). These values cannot be directly processed by machine learning algorithms. Therefore, it is essential to handle such data during preprocessing. A common approach is to replace the Nan values with the mean of their respective columns, ensuring that the dataset is complete and suitable for analysis [1].

B. DataSplitting

The dataset is typically divided into two subsets: training and testing sets. The training set usually consists of 80% of the data, which is used to train the model. The remaining 20% is used as the testing set to In many heart disease prediction systems, various classification algorithms have been applied. Although each has its advantages, they also come with limitations. Below is a summary of key machine learning techniques used in existing systems:

1. Decision Tree

In a decision tree algorithm, the first step involves calculating the information gain of all attributes in the dataset. The attribute with the highest information gain is selected as the root node. This attribute serves as the starting point of the tree structure, and the model then recursively splits the dataset based on this feature to minimise entropy and improve classification accuracy [3].

2. K-Nearest Neighbours (KNN)

KNN is one of the simplest yet effective classification methods. It is particularly useful when probability distributions are hard to define. The algorithm works by using the training dataset to determine the location of the K-nearest neighbours to a target instance, calculated using Euclidean distance. Each data point is classified based on the majority label of its nearest neighbours. The algorithm continues this process for each instance in the testing set. In applications, the value of K (number of neighbours) can be varied to build multiple parallel models. Tools like WEKA help automate this by iterating through a range of K values to determine the optimal one. For best results, input and output variables are clearly defined, and the system builds models accordingly.

3. K-Means Clustering

K-Means is an unsupervised learning algorithm, which means it is used when class labels are unknown. The primary goal is to group the data into K clusters based on similarity. Each cluster has a centroid, and data points are grouped by their closeness to the centroid. Once the groups are formed, any new data point is assigned to a cluster based on its similarity to the existing centroids. Since this approach does not rely on labelled data, it is mainly used for exploratory data analysis and pattern recognition [3].

4. Adaptive Boosting (AdaBoost)

AdaBoost is a boosting technique used to improve the performance of weak classifiers. It is especially effective in binary classification problems. AdaBoost typically enhances the accuracy of decision trees, especially when using simple one- level trees known as decision stumps. By combining multiple weak classifiers, the final ensemble model achieves higher accuracy than any individual model. This technique is widely used for classification tasks where slightly improving existing models can yield significant benefits.

5. Other Classification Algorithms

In addition to the methods above, other algorithms like Naive Bayes, Neural Networks, and Support Vector Machines (SVM) have also been used for heart disease prediction [5]. These methods work as follows:

- They accept training and testing datasets as input.
- They utilise data mining environments such as WEKA for implementation.
- Cardiovascular disease (CVD) datasets are sourced from multiple databases and typically include binary attributes (True/False) indicating the presence of heart disease.
- These datasets are converted into ARFF format (Attribute-Relation File Format), which is compatible with the WEKA tool.
- User inputs are often managed using a SQL server, and inputs generated in Excel are also converted to ARFF format.
- Once a new instance is submitted, the system classifies it and returns a predicted class label, indicating the presence or absence of heart disease [4].

3. ENHANCED PREDICTION METHOD OF HEART DISEASE

First, train the model on a heart disease dataset (CSV file), preprocess the data, and save the trained model using **joblib**. In the Flask web app, create routes for uploading the CSV file, making predictions with the model, and displaying results. The frontend includes HTML forms for file uploads, while the backend handles data processing and prediction. Once tested locally, you can deploy the app to platforms like Heroku. Ensure proper error handling and security for the app.

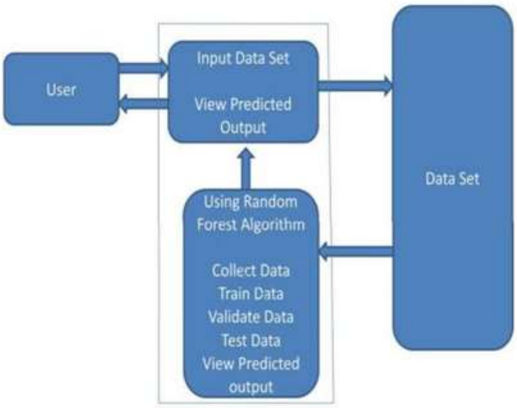


Fig-1: Architecture diagram.

The system operates using the Random Forest Algorithm, a robust supervised machine learning technique known for improving accuracy and preventing overfitting. It builds multiple decision trees on different subsets of the training data, and the final prediction is made by averaging the results from all trees. The more trees involved, the higher the accuracy. The process begins by selecting **K** data points from the training set, followed by building decision trees on these subsets. The number of trees

(**N**) is determined, and the process is repeated. For new data, each tree makes a prediction, and the class that receives the majority of votes is chosen as the final output. Typically, 70% of the data is used for training, while the remaining 30% is used for testing. The model's key advantages are its high performance, flexibility, and ability to achieve high accuracy. During decision tree construction, the dataset's attributes are categorised to ensure that the trees are built and predictions are made effectively.

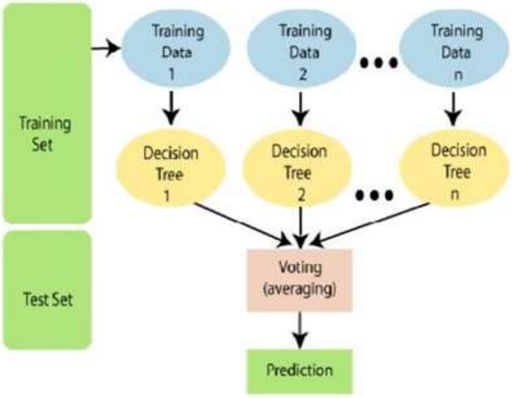


Fig-2: Procedure of random forest algorithm.

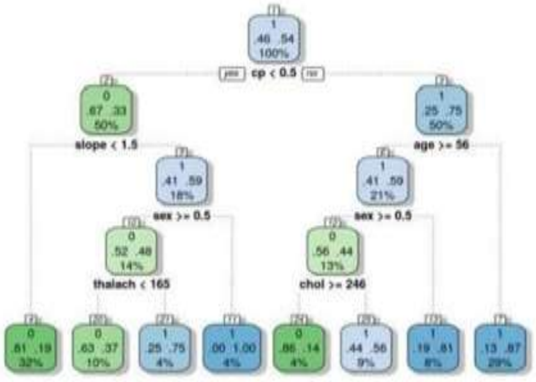
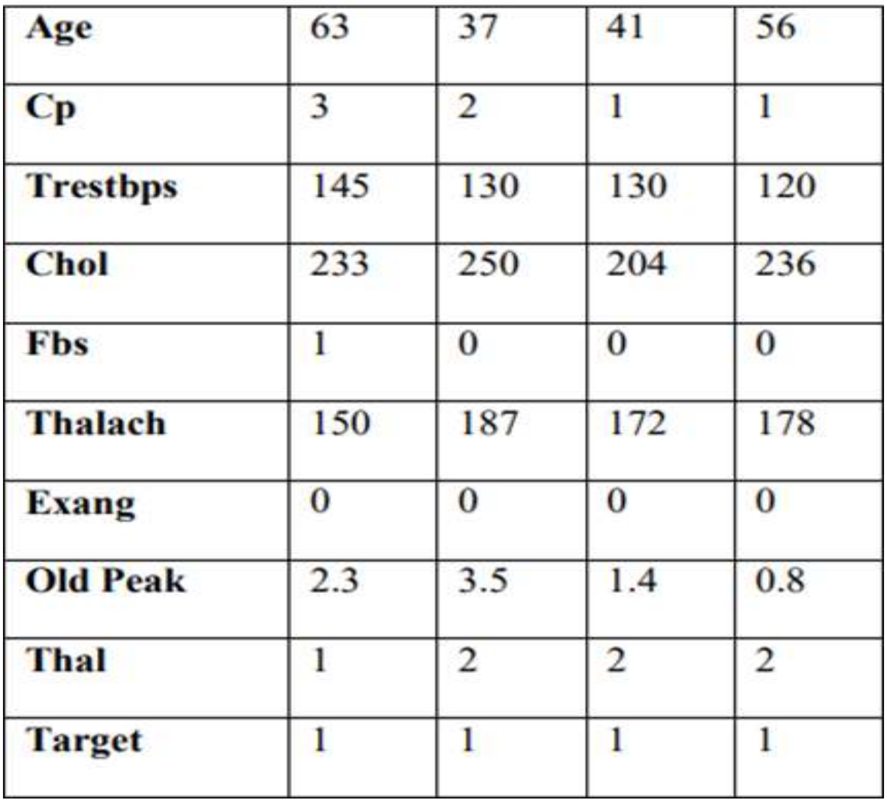


Fig-3 Building decision tree.

4. Result Analysis:

The primary objective of this project is to determine whether a person has heart disease or not and provide recommendations for further action. By utilising the Random Forest algorithm, we can achieve a high accuracy rate in predicting heart disease. This is due to the algorithm's ability to combine the results from multiple decision trees, leading to better performance and reliability. The dataset used for this model contains various attributes related to heart health. These attributes are processed and analysed to predict the likelihood of heart disease. Below is a sample of the dataset that was used for training and testing the model:

Table-1 Sample data set

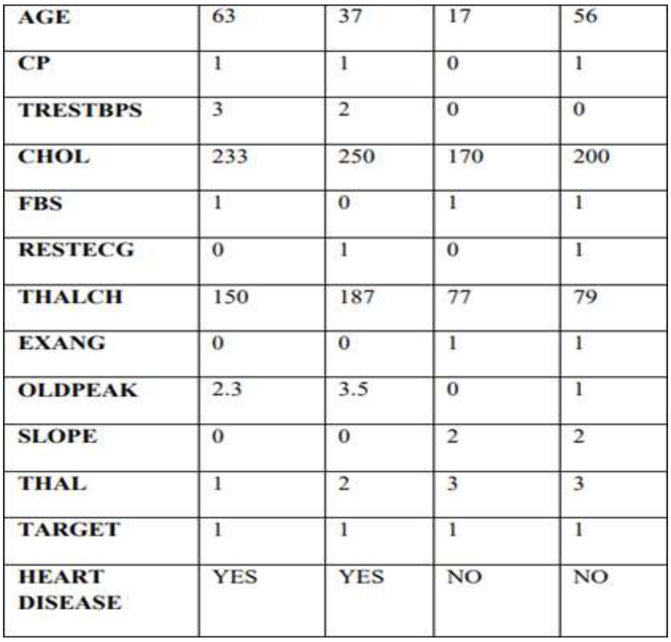


Age	63	37	41	56
Cp	3	2	1	1
Trestbps	145	130	130	120
Chol	233	250	204	236
Fbs	1	0	0	0
Thalach	150	187	172	178
Exang	0	0	0	0
Old Peak	2.3	3.5	1.4	0.8
Thal	1	2	2	2
Target	1	1	1	1

The attributes mentioned in **Table 1** are sufficient to predict whether a person is affected by heart disease. Each attribute represents a specific aspect of heart functionality and helps in making an informed prediction. For instance, the **Chest pain type (Cp)** is categorised into four distinct values: **Typical Angina**: Chest pain that occurs with physical exertion or stress. **Atypical Angina**: Chest pain that doesn't follow the typical pattern and might occur even without physical exertion. **Non-anginal Pain**: Chest pain that is not related to heart disease and usually caused by other factors, like indigestion. **Asymptomatic**: No chest pain or symptoms are experienced by the person. These attributes provide critical insights into the patient's heart condition. The dataset, as illustrated in **Fig.4** and listed in **Table 1**, contains several other features such as age, blood pressure, cholesterol levels, and ECG results, all contributing to the overall prediction of heart disease risk. Each attribute reflects a specific functionality of the heart, allowing the Random Forest model to analyse and predict the likelihood of a person having heart disease with high accuracy.

- Trestbps- Level of blood pressure at resting mode.
- Chol-Serum cholesterol in mg/dl. • Fbs- Blood sugar levels on fasting (if>120mg/dl represented as 1, otherwise 0)
- Resting ECG- Results of electrocardiogram while at rest.
- Exang- Angina induced by exercise (0-No, 1-Yes) •
- Old peak- Exercise induced ST depression in comparison with the state of rest.

Table 2: Sample data set with results.



AGE	63	37	17	56
CP	1	1	0	1
TRESTBPS	3	2	0	0
CHOL	233	250	170	200
FBS	1	0	1	1
RESTECG	0	1	0	1
THALCH	150	187	77	79
EXANG	0	0	1	1
OLDPEAK	2.3	3.5	0	1
SLOPE	0	0	2	2
THAL	1	2	3	3
TARGET	1	1	1	1
HEART DISEASE	YES	YES	NO	NO

From this, we can say that the application, when implemented using random forest algorithm has more accuracy rate when compared to other algorithms.

5. Conclusion:

The **Random Forest algorithm** proves to be an efficient and reliable tool for both regression and classification tasks. By constructing multiple decision trees, it calculates the final output as the average of all the trees, leading to high accuracy in predictions. This approach allows for early-stage detection of heart disease with improved reliability. Processing healthcare data, specifically heart-related data, aids in the early detection of potential heart conditions, which can significantly reduce the risk of long-term health complications or even death. Heart disease prediction remains a crucial challenge in today's fast-paced life, and this application offers a valuable solution. Even for patients who are unable to visit a doctor, the app allows them to input their health data (like reports or test results) and receive a prediction about their heart condition. Based on the result, users can decide whether they need to seek medical consultation, ultimately improving proactive healthcare management.

6. Future Scope:

In the future, this heart disease prediction application can be expanded by adding several new features to enhance its functionality and impact. One such feature could be notifying the user's family members if a heart disease prediction is positive, allowing them to take timely action. Additionally, the application could send the user's information to the nearest hospital, ensuring that emergency services are ready if needed. Another important feature would be enabling **online doctor consultations** with the nearest available doctor. This would provide users with immediate professional advice and guidance, even if they are unable to visit a healthcare facility in person.

Furthermore, **Machine Learning** applications, such as the one used in this project, have enormous potential beyond heart disease prediction. These algorithms can be applied in various fields like **radiology**, **bioinformatics**, and **medical imaging** to improve the accuracy of disease diagnosis and treatment planning, ultimately contributing to more efficient and personalised healthcare solutions.

7. REFERENCES

1. Kaan Uyar and Ahmet İlhan, "Diagnosis of heart disease using genetic algorithm-based trained recurrent fuzzy neural networks" in B.V ICTASC, Elsevier, pp

2. Ashish Chhabbi, Lakhan Ahuja, Sahil Ahir and Y. K. Sharma, "Heart Disease Prediction Using Data Mining Techniques", © IJRAT Special Issue National Conference “NCPC-2016”, pp. 104-106, 19 March 2016.

3. Kirmani, M.M., Ansarullah, S.I.: Prediction of heart disease using decision tree, a data mining technique. IJCSN Int. J

4. Salam Ismaeel, Ali Miri et al., "Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis", IEEE Canada International Humanitarian Technology

5. N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian network classifiers,” Machine Learning,(1997)