# Heart disease classification using Random Forest

**Arpit Gupta, Ankush Shahu, Masud Ansari, Nilalohit Khadke, Prof. Ashwini Urade**

*Students, Department of Computer Science J D College of Engineering and Management Nagpur, India*
*Professor, Department of Computer Science J D College of Engineering and Management Nagpur, India*

***Abstract:** Cardiovascular disease is still the leading cause of death worldwide, and the early prediction of heart disease is of great importance. In this paper, we propose a supervised learning algorithm for early prediction of heart disease using old patient medical records and compare the results with a well-known supervised classifier – Random Forest. Patient record information is classified using a CNN (Cascade Neural Network) classifier. At the classification stage, 13 features are provided as input to the CNN classifier to determine heart disease risk. The proposed system will help doctors diagnose diseases more efficiently. The effectiveness of the classifier was tested on 303 patient records. The raw data comes from a combination of 4 databases: Cleveland, Hungary, Switzerland and VA Long Beach data from the UCI Machine Learning Repository. This result suggests that CNN classifiers can more effectively predict the likelihood of heart disease. The proposed method allowed the model to achieve an accuracy of 95.17% in predicting heart disease. Experimental results show that our algorithm improves the accuracy of heart disease diagnosis.*

***Keywords:** Random forests, heart disease prediction, Machine learning.*

## I. INTRODUCTION

The human heart is the main organ of the human body. Any type of disturbance in the normal functioning of the heart can be classified as heart disease. In today's modern world, heart disease is one of the leading causes of most deaths. Heart disease can be caused by an unhealthy lifestyle, smoking, drinking alcohol, and eating too much fat. According to the World Health Organization, more than 10 million people worldwide die of heart disease every year.

A healthy lifestyle and early detection are the only ways to prevent heart disease. The greatest challenge in healthcare today is to provide the highest quality of service and accurate and efficient diagnosis. Although heart disease has proven to be the leading cause of death worldwide in recent years, it is also a disease that can be effectively controlled and managed. Any precision in the management of diseases depends on the right timing of these diseases. The proposed work attempts to detect these heart diseases early enough to avoid catastrophic outcomes.

Many medical data records created by medical professionals are available for analysis. Data mining techniques are methods of extracting valuable and hidden information from large amounts of available data. Medical databases consist mostly of fragmentary information. Therefore, making decisions usingdiscrete data sets becomes a complex and difficult task. Machine learning (ML), as a subfield of data mining, can efficiently handle well-structured large-scale datasets. In medicine, machine learning can be used to diagnose, detect and predict various diseases.

The main purpose of this article is to provide a tool to help doctors detect heart disease at an early stage. This will help to effectively treat patients and avoid serious consequences. ML plays a very important role in detecting hidden discrete samples and thus provides data analysis. After analyzing the data, machine learning technology helps predict heart disease and make early diagnosis. This article presents an analysis of the performance of random forest techniques in the prediction of early heart disease.

## II. RELATED WORK

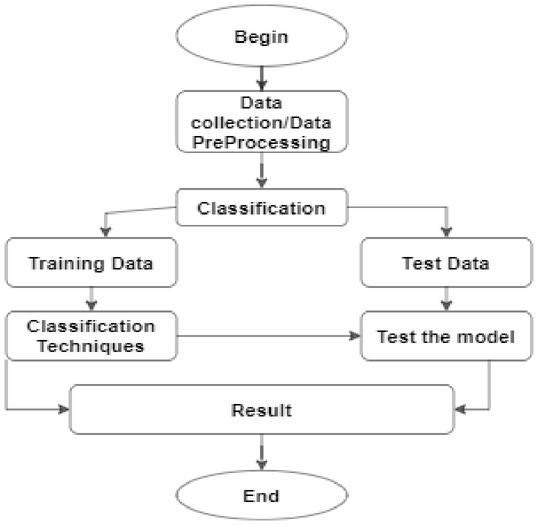
Background Research:

| Title | Publication, Year | Research Gap |
|---|---|---|
| Applying ML classifiers on ECG dataset for predicting heart disease | IEEE, 2021<br>Adiba Hossain Sabitri Sikder | Current SVM model has 85.49%<br>accuracy. In future, more analysis can be performed with the different combinations of algorithms to obtain a better heart disease prediction model. |
| Using machine learning to predict heart disease | CCLA USA, 2022<br>Nikhil Bora | The lowest accuracy was from Naïve Bayes of 79.83% against highest accuracy of 94.12% from Random Forest. |
| Latest trends of heart disease prediction using ML and image fusion | Elsevier, 2020<br>Manoj Diwakar P.Singh | Quality of dataset is an important factor, and thus hospitals should be encouraged to publish high quality datasets. |

| | | |
|---|---|---|
| Heart Disease Prediction | IRJETS, 2022<br>Nakkina Rajjani | The scope is to check the availability of heart disease with fewer tasks and attributes that gives high accuracy and efficiency. |

| Title | Publication, Year | Research Gap |
|---|---|---|
| Random forest swarm optimization for heart diseases diagnosis | Elsevier, 2021<br>Shahrokh Asadi Michael Kattan | Many other multi-objective optimization methods appear in the literature such as non-dominated genetic algorithm ii which can be employed instead of MOPSO. |
| Machine Learning Models for prediction of co-occurrence of diabetes and cardiovascular diseases | Springer, 2022<br>Ahmad Abdalrada Jemal Abawajy | The model has high accuracy and in the future, it can be employed as a tool for web-based and mobile phone application, thus increasing its reach among people and healthcare providers. |
| Prediction of Heart Disease utilising SVM and ANN | IJEECS, 2021<br>Alaa Khaleel Faieq | SVM is employed currently. While in the future, other techniques can be applied to predict other heart diseases using the same data. |
| Improving the prediction of Heart Failure Patients' Survival using SMOTE and Data Mining Techniques | IEEE, 2021<br>Abid Ishaq, Muhammad Umer | To improve the performance of ML models, better features selection techniques can be devised. In this case, meta-heuristics can be used due to NP-hard nature of feature selection problems. |

## III. METHODOLOGY



**Proposed Model: Fig. 1 shows the entire process involved.**

### A. Data Collection and Preprocessing

The dataset used is the Cardiology dataset, which is a combination of 4 different databases, but only the UCI Cleveland dataset was used. The database contained a total of 76 traits, but all published tests only referenced a subset of 14 traits. Therefore, we use the UCI Cleveland processing dataset available on the Kaggle website for analysis. Table 1 below gives a full description of the 14 attributes used in the proposed work.

**TABLE I. FEATURES SELECTED FROM DATASET**

| Sl.No. | Attribute Description | inct Values of Attribute |
|---|---|---|
| 1. | *Age*- represent the age of a person | Multiplevalues between 29 & 71 |
| 2. | *Sex*- describe the gender of person (0- Feamle, 1-Male) | 0,1 |
| 3. | *CP*- represents the severity of chest pain patient is suffering. | 0,1,2,3 |
| 4. | *RestBP*-It represents the patient's BP. | Multiplevalues between 94& 200 |
| 5. | *Chol*-It shows the cholesterol level of the patient. | Multiplevalues between 126 & 564 |
| 6. | *FBS*-It represent the fasting blood sugar in the patient. | 0,1 |
| 7. | *Resting ECG*-It shows the result of ECG | 0,1,2 |
| 8. | *Heartbeat*- shows the max heart beat of patient | Multiple values from 71 to 202 |
| 9. | *Exang*- used to identify if there is an exercise induced angina. If yes=1 or else no=0 | 0,1 |
| 10. | OldPeak-describes patient's depression level. | Multiple values between 0 to 6.2. |
| 11. | Slope- describes patient condition during peak exercise. It is divided into three segments(Unsloping, Flat, Down sloping) | 1,2,3. |
| 12. | CA- Result of fluoroscopy. | 0,1,2,3 |
| 13. | Thal- test required for patient suffering from pain in chest or difficulty in breathing. There are 4 kinds of values which represent Thallium test. | 0,1,2,3 |
| | Target-It is the final column of the dataset. It is class or label Colum. It represents the number | 0,1 |

| | | |
|---|---|---|
| 14. | of classes in dataset. This dataset has binary classification i.e. two classes (0,1).In class "0" represent there is less possibility of heart disease whereas "1" represent high chances of heart disease.<br>The value "0" Or "1" depends on other 13 attribute. | |

Data exploration, also known as exploratory data analysis (EDA), is an essential step in the machine learning process. It involves analyzing and understanding data sets to better understand the data and identify patterns, relationships, and anomalies. While exploring data, we use a variety of statistical and visualization techniques to summarize and describe data. These techniques include:

1) **Descriptive statistics**: the mean, median, mode, standard deviation, correlation and other statistics are used to summarize the data.

2) **Data Visualization**: Histograms, scatter plots, boxplots, heatmaps, and other visualizations are used to visually explore data and identify patterns and trends.

3) **Dimensionality reduction**: Use principal component analysis (PCA), t-SNE, and other techniques to reduce the dimensionality of the dataset and visualize it in a low-dimensional space.

4) **Outlier detection**: Identify and analyze outliers to determine if they are true data points or false data points.

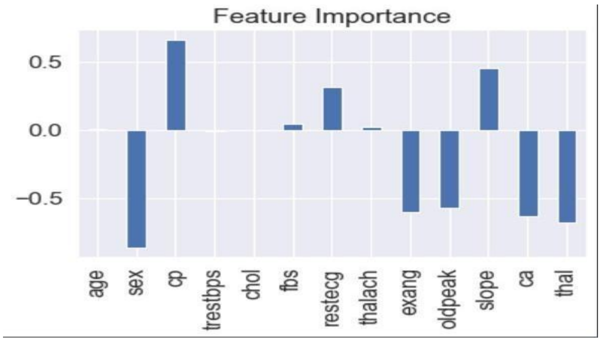


Figure 1: Feature Importance

## B. Classification:

The input dataset is split into 80% of the training dataset and the remaining 20% of the test dataset. Training dataset

is the dataset used to train the model. The test dataset is used to verify the performance of the trained model. Performance is calculated and analyzed based on different metrics used, such as accuracy, precision, retrievability, and F-score.

## C. Training data using random forest

Random Forest: Random Forest is a popular machine learning algorithm used for classification, regression, and other tasks. An ensemble learning method that combines multiple decision trees to make more accurate predictions. In random forests, a set of decision trees is created on a random subset of the original dataset. Each decision tree in the forest is built using a different subset of features and training examples. The process of building each tree is repeated until the specified number of trees have been created.

To predict the use of a random forest, we walk through each tree in the forest and make a classification or regression decision. The predictions from each tree are then combined to form the final prediction. The combined prediction is done by majority vote (in classification) or by average (in regression).

**The advantages of random forest algorithm are:**

**1) High accuracy:** Random forests are known for the high accuracy of their predictions.

**2) Robustness:** Random forests are less prone to overfitting than individual decision trees.

**3) Ease of use:** Random Forest does not require extensive data preparation and can handle missing data.

**4) Feature Importance:** Random forests provide a measure of feature importance that can help identify the most important features for prediction.

**Optimizing the accuracy of the model:** The initial predictions of the model are not always correct. To further improve accuracy and precision, we performed the following steps:

**1) Hyperparameter tuning**: Hyperparameter tuning is the process of selecting the best set of hyperparameters for a machine learning model. Hyperparameters are parameters that are not learned during training, but are set before training begins, such as the learning rate, the strength of regularization, or the number of hidden units in a neural network.

Hyperparameter settings are important because model performance depends on the hyperparameters chosen. We do this through trial and error by training models with different combinations of hyperparameters and evaluating their performance on the validation set.

We automate this using techniques such as grid search, random search, or Bayesian optimization.

Proper tuning of hyperparameters can significantly improve the performance of a machine learning model, while improper tuning can cause the model to perform poorly or even fail completely.

**2) Confusion matrix:** A confusion matrix is a table that summarizes the performance of a classification model on a set of test data whose true values are known. It is a way to visualize the performance of machine learning algorithms by comparing predicted and actual values. The confusion matrix is typically a 2x2 matrix for binary classification, with four possible outcomes:

True positive (TP): The model predicts a positive and the actual value is positive.

False Positive (FP): The model predicts a positive outcome, but the actual value is negative.

True Negative (TN): The model prediction is negative and the actual value is negative.

False Negative (FN): The model predicted negative, but the actual value was positive.

The confusion matrix can be used to calculate various evaluation metrics for classification models, such as accuracy, precision, recall, F1 score, etc. These metrics provide insight into model performance and help identify areas for improvement.

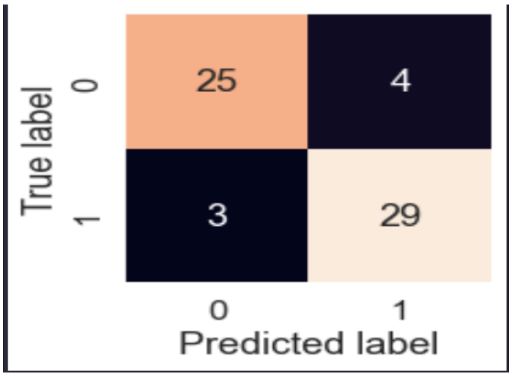


**Figure 2: Confusion matrix**

### D. Result and Analysis:

With the increasing number of deaths from heart disease, it is imperative to develop an efficient and accurate heart disease prediction system. The motivation for Study

was to find the most efficient ML algorithm for detecting heart disease. The random forest algorithm achieved 86% accuracy in predicting heart disease. In the future, the work could be improved by developing a web application based on the Random Forest algorithm and using a larger data set than used in this analysis, which would lead to better delivery of results and help health professionals. Heart disease can be predicted effectively. and efficiently.

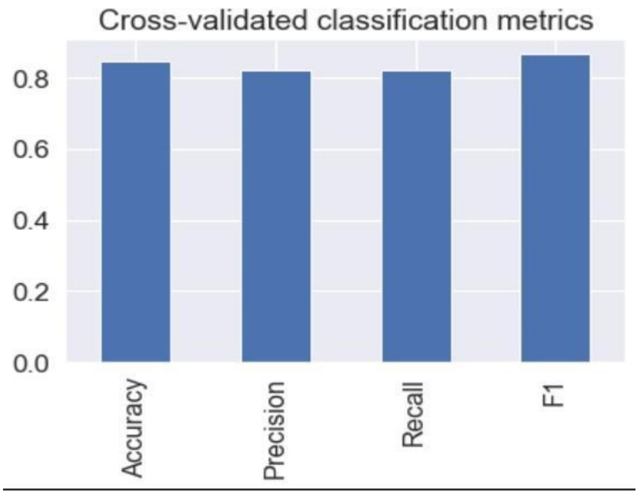
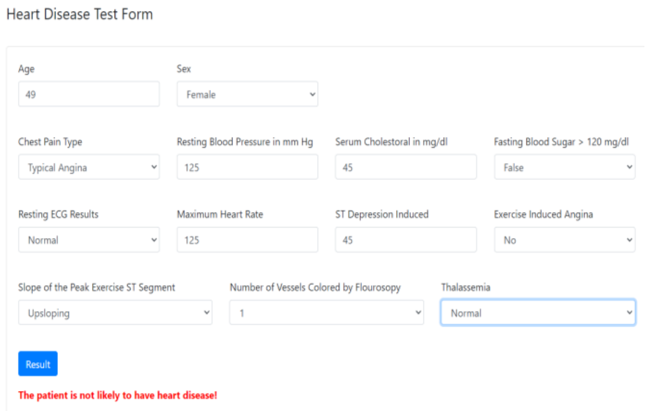


Figure 3: Cross-validated classification metrics



Heart Disease Test Form

Age: 49
Sex: Female
Chest Pain Type: Typical Angina
Resting Blood Pressure in mm Hg: 125
Serum Cholestoral in mg/dl: 45
Fasting Blood Sugar > 120 mg/dl: False
Resting ECG Results: Normal
Maximum Heart Rate: 125
ST Depression Induced: 45
Exercise Induced Angina: No
Slope of the Peak Exercise ST Segment: Upsloping
Number of Vessels Colored by Flourosopy: 1
Thalassemia: Normal

The patient is not likely to have heart disease!

Figure 4: User interface

## IV. FUTURE WORK

The quality of the data used to train a model has a significant impact on the final predictions of the model. Future work may involve improving data quality by cleaning and pre-processing the data more thoroughly or collecting more data. Additionally, collecting data from patients of all age groups can lead to significant improvements. More fine-tuning of hyperparameters can be performed to help improve model accuracy. Heart disease can also manifest in different ways, so classifiers are constructed for different outcomes (eg: heart attack, stroke, cardiac arrhythmia) may be more helpful. This can help develop more targeted interventions and improve overall health.

## V. CONCLUSION

In this paper, Random Forest data mining algorithm was implemented to predict heart disease. In the proposed work, we achieved a classification accuracy of 86.9% for predicting heart disease with a diagnosis rate of 93.3% using the random forest algorithm. As an extension of this work, different types of classifiers can be included in the analysis and further sensitivity analysis can be performed. This classifier can also be extended by applying the same data set analysis of other bioinformatics diseases and seeing the performance of these classifiers to classify and predict these diseases. Cloud computing technology can also be used for the proposed system to manage large volumes of patient data.

## VI. REFERENCES

1. J. Krishnan Santana; S. Geetha “Prediction of Heart Disease Using Machine Learning Algorithms”. 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)Publisher: IEEE

2. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). “Effective Heart Disease Prediction using Hybrid Machine Learning Techniques”. IEEE Access, 1–1. doi:10.1109/access.2019.2923707

3. Rajdhan Apurb, Agarwal Avi, Sai Milan, Ravi Dundigalla, Ghuli Poonam.” Heart Disease Prediction using Machine Learning” INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY

4. Abdullah AS. “A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier”, Proceedings on International Conference in Recent trends in Computational Methods, Communication and Controls (Icon3c); 2012. p. 22–5.

5. Kelwade JP. “Radial basis function Neural Network for Prediction of Cardiac Arrhythmias based on Heart rate”