

Heart Ailment Prediction using Machine Learning Methods

Priya Shelke¹, Chaitali Shewale², Ratnmala Bhimanpallewar³, Tamkhade Jayashree⁴, Mrunali Gadekar⁵ and Abhigyan Hedau⁶

¹⁻⁶Vishwakarma Institute of Information Technology, Pune, India

Email: priya.shelke@viit.ac.in, {chaitali.shewale, ratnmala.bhimanpallewar, jayashree.tamkhade, mrunali.22010831, abhigyan.22010904}@viit.ac.in

Abstract—The heart is the coordinating centre of the major endocrine glandular structure of the body, which produces hormones that profoundly affect the operations of the body, and diagnosing cardiovascular disease is a difficult but critical task. By extracting knowledge and information about the disease from patient data, data mining is a more practical technique to help doctors detect disorders. We use a variety of machine learning methods here, including logistic regression and support vector classifiers (SVC), K-nearest neighbours Classifier (KNN), Decision Tree Classifier, Random Forest Classifier and Gradient Boosting Classifier. These algorithms are applied to patient's data containing 13 different factors to build a system that predicts heart disease in less time with more accuracy.

Index Terms— Logistic Regression, Support Vector Classifier, K-Nearest Neighbour, Decision Tree, Random Forest and Gradient Boosting.

I. INTRODUCTION

Rheumatic heart disease is linked to around 2% of cardiovascular disease-related fatalities worldwide. The terms "cardiovascular disease" and "heart disease" are sometimes used interchangeably. Heart attacks, chest pain (angina), strokes, and other illnesses caused by restricted or obstructed blood vessels are together referred to as cardiovascular disease. Symptoms consist of Angina, or chest pain from the heart muscle due to insufficient oxygen and nutrient-rich arterial blood flow, which is a typical sign of heart disease. You get chest discomfort as a result of angina. Around their breastbone, some people feel tightness or a squeezing sensation. The neck, shoulder blades, upper arms, upper abdomen, and upper back may all experience pain radiating from the lower back. The most important organ in the human body, the heart controls blood flow throughout the body. Other body parts may suffer if there is any kind of heart function impairment. Heart disease is currently the biggest cause of death among people. According to estimates from the World Health Organization, almost 12 million people die from heart disease each year (WHO). The WHO estimates almost the death rate would increase to 23.6 million by 2030 [6].

Dizziness, ankle swelling, shortness of breath, slow heartbeats, fainting, lightheadedness, pain in the neck, jaw, throat, dullness, weakness, or coldness in your body parts, and irregular heartbeats are all signs of this illness. Heart disease can be prevented if detected earlier. More accurate diagnoses in less time. Providing the best standard services and early, correct diagnosis is the industry's key problem. The extensive application of machine learning, which also produces favorable results with the highest accuracy for medical diagnostics, can

have a positive impact on the healthcare sector. Finding the best algorithm for heart disease prediction is the study's goal for data mining. Algorithms like Random Forest, Logistic Regression, K-Nearest Neighbours Classifier, Support Vector Classifiers, Decision Tree Classifier, and Gradient Boosting Classifier are used in the classification and construction of a model to diagnose heart disease in patients. [11]. A dataset is used to apply algorithms, and the accuracy levels of the results are compared. The complex task of decision making using discrete data is easily handled by machine learning. By identifying hidden patterns, machine learning (ML) analyses the provided data. As a result, a tool is created that enables medical professionals to diagnose patients quickly, treat them effectively, and prevent negative outcomes. [1] [4] [14]. The field of machine learning is growing at a fast pace in different industries such as healthcare, transportation, finance, agriculture, cybersecurity, marketing, etc. We can reduce manual error with computer analysis and also, increase the accuracy and efficiency of a system. The need for near 100% accuracy and reduction of human error is most in the healthcare industry.

II. LITERATURE SURVEY

A project is created to predict the possibility of getting heart disease in patients. It is predicted in terms of percentage using decision Tree as well as K-nearest Neighbour algorithm, taking into consideration vital factors which include pressure level, gender, age, cholesterol, chest pain, rest blood pressure, fasting blood sugar, chest pain type, electrocardiographic result [1].

A study on the prediction of cardiovascular disease compared the accuracies of two algorithms and their hybrids. The study came to the conclusion that the Decision tree algorithm had a 79% accuracy rate, Random Forest algorithm had an 81% accuracy rate, and their Hybrid model had an 88% accuracy rate.[2]

Ekta Maini's work on developing a machine learning model for effective and early cardiovascular disease prediction. With the aid of different algorithms, India took into account eleven related factors of a subject and came to the conclusion that the accuracy of logistic regression is 90.8%, the specificity of KNN is 87.1%, and the specificity of the AB model is 93.1% [3].

Rati Goel et al gave a brief comparison between the efficiency of six different algorithms which include Support Vector Machine, Random Forest, Naïve Bayes, Decision Tree, Logistic regression, K-nearest Neighbor, for the purpose of finding the best suited algorithm to detect heart disease. The study came to a conclusion that the accuracies of each of the algorithms were as follows -Logistic Regression 77%, KNN 82%, SVM 86%, Naïve Bayes 68%, Decision Tree 83%, Random Forest 83%. According to the analysis, Support Vector Machine is the best algorithm for heart disease early prediction [4].

Santhana Krishnan used Decision Tree Classification and Naive Bayes Classification models for classification. After applying these two supervised-data mining algorithms to the dataset, it was discovered that the Naive Bayes classifier had an accuracy level of 87% in predicting heart disease patients and the Decision Tree Model had an accuracy level of 91%.[5].

Mohd Faisal Ansari studied how attributes affected the outcomes of a logistic regression technique model. He used a variety of models, including logistic (all attributes), logistic (most significant attributes), logistic (removing the least significant attribute), SVM, and logistic (removing the least significant attribute) (with PCA) gave 86% accuracy, recall 68% and specificity 69% with precision 77%, and a f1 score of 72%, the study's findings demonstrate that Logistic (with PCA) performed best. [6].

T Marikani went through various studies to find the best suited algorithm of heart disease prediction, the algorithms under scanner here are supervised learning algorithms like Decision tree, Naïve Bayes, Random Forest Tree, KNN and When finished, use a support vector machine. According to the study, the accuracy of various algorithms varied depending on the implementation tools and attributes used. [7]

V.V.Ramalingam carried a comprehensive study on the comparison of various methodologies for heart disease prediction which included algorithms and techniques such as Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest, K – Nearest Neighbour, Ensemble Model, the study concluded that Each of the above-mentioned algorithms have performed well in some cases but poorly in other cases. Different Models that were based on Naïve Bayes classifiers were quite quick and performed well. For most of the cases, SVM performed well. [8]

Pooja Anbuselvan evaluated the performance of a number of machine learning algorithms, including Naive Bayes, Logistic Regression, K-Nearest Neighbor, Decision Tree, Random Forest, and Support Vector Machine, and discovered that Random Forest and XGBoost are the most effective ones, each scoring 86.89% and 78.69%, respectively. The least accurate algorithm was K-Nearest Neighbor, which performed at 57.83% [9].

III. PROPOSED METHODOLOGY

The main goal of this paper is to estimate the likelihood that patients may develop heart disease, and data mining is crucial in achieving this goal. This research makes use of the 13-factor heart disease dataset. Gender, age, exercise-induced angina, resting blood pressure, cholesterol, fasting blood sugar, chest pain, thalassemia, results of resting electrocardiography, maximum heart rate reached, ST depression brought on by exercise in comparison to rest, slope, and number of major vessels are some of these factors. The program employs a classification technique.

A. Architecture Diagram

i. Data Import and Preprocessing

Fig. 1 shows 6 steps of preprocessing, which includes, import data, duplication removal, preprocessing, encoding, feature scaling and the preparation of training and testing dataset.

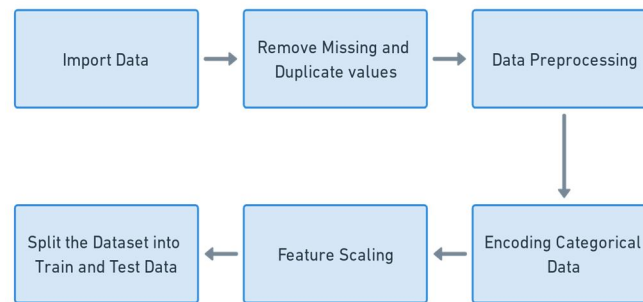


Figure 1. Preprocessing Steps

ii. Train Models

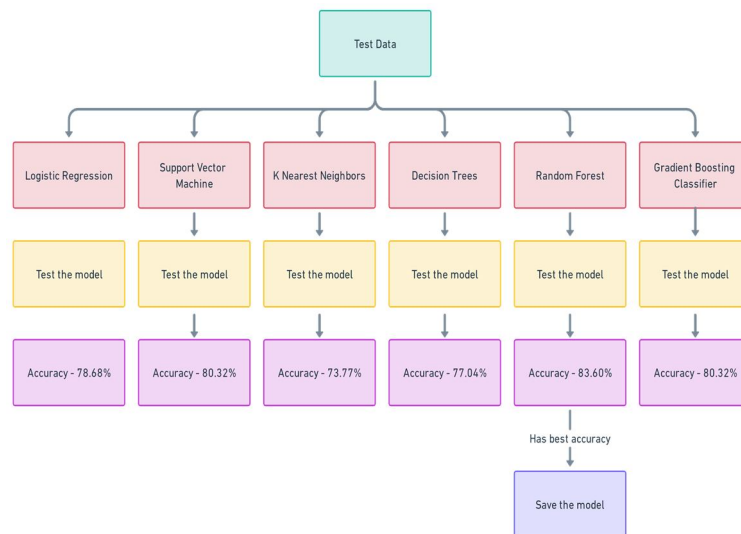


Figure 2. Training Steps

iii. Prediction

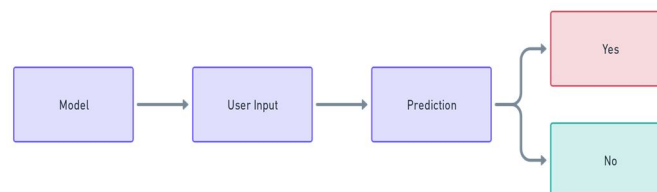


Figure 3. Prediction Steps

B. Data Source

The dataset used in the prediction process was obtained from the machine learning repository at the University of California, Irvine. The dataset consists of 1026 instances of data with the 13 medical factors that are appropriate for prediction. [15].

C. Steps

1. Preprocessing: Data is checked for null and duplicate values and is filtered. Then the data is encoded, feature scaling is carried out and lastly splitted into training and testing data. Data preprocessing is represented in fig. 1.
2. Training: The training data is fed to each of the ML algorithms and then tested using the test data as displayed in fig. 2. Accuracies are calculated and the model having best accuracy is saved as the model.
3. Prediction: Lastly, prediction is carried out on the saved model and output to the user is shown in the fig. 3.

IV. RESULT ANALYSIS

The dataset has two categories of data: training data and test data. Classification techniques including logistic regression, Gradient Boosting, SVC, Decision Tree, K-Nearest Neighbors, and Random Forest are applied after preprocessing the data. It demonstrates that K-Nearest Neighbors, Decision Tree, and Logistic Regression provide accuracy rates of 78.68%, 77%, and 73.77%, respectively. Random forest has an accuracy rate of about 83.60%. Both the Support Vector Machine and Gradient Boosting algorithms yield an accuracy of 80.327%. The experimental results of various algorithms performance with their accuracies is shown in table I and its graphically represented in fig. 4.

TABLE I. ALGORITHMS AND THEIR ACCURACIES

Algorithm	Accuracy (%)
K Nearest Neighbor	73.77
Decision Tree	77.04
Logistic Regression	78.68
Gradient Boosting	80.32
Support Vector Machine	80.32
Random Forest	83.60

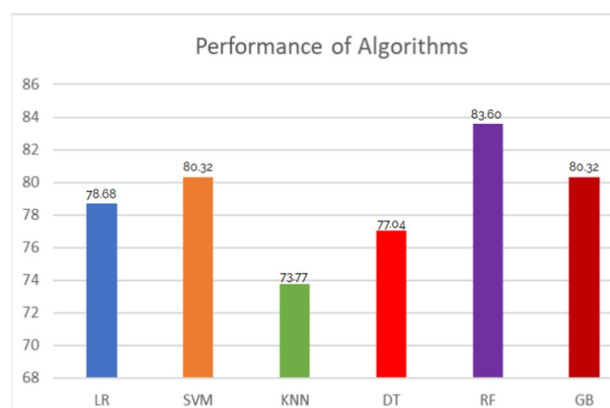


Figure 4. Results

However, amongst various classification algorithms, the random forest accuracy is at a higher side. A small part of the random forest is represented in fig. 5. The UI consists of a form as in fig. 6, where the user will be able to enter the values of the factors that we considered for training the model.

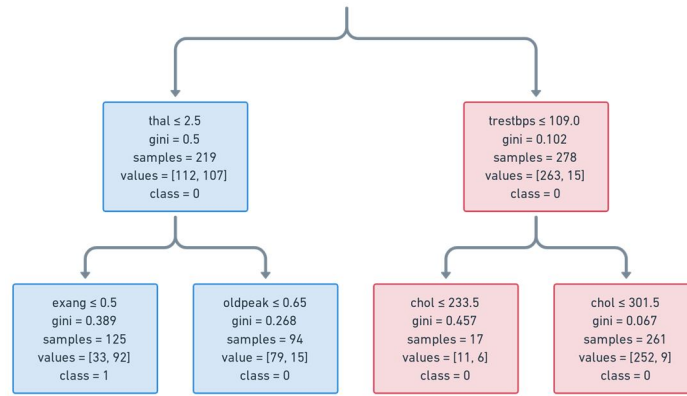


Figure 5. Small part of random forest

Figure 6. UI Screenshot 1

Figure 7. UI Screenshot 2

After the user has entered the values for the fields (fig. 7), if the model returns 1 for the values the user provided, then it will show “Possibility of Heart Disease”, else, it’ll show “No Heart Disease”.

V. MATHEMATICAL CALCULATIONS

Numerous decision trees are created by random forests (RF) during training. The final predictions—the method of the classes for grouping or the mean prediction for regression—are made by pooling predictions from all trees. They are referred to as "trooper methods" because they draw from a variety of outcomes before making a final decision. To predict the target value, decision trees determine how to divide the information into successively smaller subsets.

Scikit-learn gives data in their documentation on the recipes utilized for debasement basis. For order, it utilizes Gini impurity of course but offers Entropy (2) as another option.

$$Gini(D) = 1 - \sum_{i=1}^k p_i^2 \quad (1)$$

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i \quad (2)$$

The weighting of the node pollution by the likelihood of reaching that node determines the relevance of the feature. The number of tests that arrive at the hub, divided by the total number of tests, can be used to calculate the hub likelihood. The component's importance increases with value.

For every decision tree, Scikit-learn computes a hub's significance utilizing Gini significance, expecting just two child nodes nodes (binary tree):

Where, $n_i(j)$ is equal to node j 's significance, weighted number of tests arriving at node j , denoted as $w(j)$. $C(j)$ = Node j 's impurity value, $Right(j)$ = child node from right split on node j , $Left(j)$ = kid node from left split on node j , Gini index for each feature calculated from (1):

Gini Index for feature 0: 0.078939

Gini Index for feature 1: 0.091710

Gini Index for feature 2: 0.064395

Gini Index for feature 3: 0.080580

Gini Index for feature 4: 0.069999

Gini Index for feature 5: 0.073910

Gini Index for feature 6: 0.089164

Gini Index for feature 7: 0.087576

Gini Index for feature 8: 0.083251

Gini Index for feature 9: 0.058257

Gini Index for feature 10: 0.076762

Gini Index for feature 11: 0.079951

Gini Index for feature 12: 0.127049

These can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values:

$$normfi_i = \frac{fi_i}{\sum fi_j} \quad (3)$$

Average over all decision trees,

$$RFfi_i = \frac{\sum normfi_{ij}}{T} \quad (4)$$

$normfi(ij)$ = the normalized component significance for I in tree j , T is total number of trees represented in (3)

$RFfi(i)$ = the significance of the element I determined from all trees in the Random Forest model, shown in (4).

Random Forest is highly effective for heart disease prediction because it can handle high-dimensional data and automatically select relevant features. It is robust against outliers and noisy data often found in heart disease datasets. Random Forest's ability to aggregate predictions from multiple decision trees reduces bias effectively. This is beneficial because decision trees have low bias but high variance, and the combination in Random Forest achieves a balance of reduced bias and reasonable variability. Overall, Random Forest is a powerful and reliable approach for heart disease prediction, offering enhanced performance.

VI. CONCLUSION

Our research focuses on using various machine learning techniques to predict heart disease, and we assess the efficacy of these algorithms by presenting a variety of signs that can be used to determine whether a patient has heart disease or not. The research demonstrates how several machine learning algorithms function in the foretelling of a cardiovascular disease. Using Python programming, the classification procedures employed in the study were carried out. According to the results above, the Random Forest Classifier is the best-performing machine learning technique out of all the strategies examined. It has an accuracy rate of 83.60 percent. The average accuracy predicted is 78.94%. K-Nearest Neighbors is the least accurate algorithm with accuracy 73.77%. In order to predict cardiac illness earlier and lower the death rate, machine learning can be utilized efficiently in this way.

FUTURE SCOPE

Advanced technology like deep learning can be applied to increase the correctness of the system up to 100%. With the implementation of better ML systems in the healthcare sector, we can briefly reduce the human error factor and also, increase the accuracy of prediction of various diseases such as heart disease, liver disease, diabetes, tumor predictions, etc.

REFERENCES

- [1] Jothi, K. A., Subburam, S., Umadevi, V., & Hemavathy, K. C. (2021). WITHDRAWN: Heart disease prediction system using machine learning. Materials Today: Proceedings.

- [2] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021). Heart Disease Prediction using Hybrid machine Learning Model. Heart Disease Prediction Using Hybrid Machine Learning Model.
- [3] Maini, E., Venkateswarlu, B., Maini, B., & Marwaha, D. (2021). Machine learning-based heart disease prediction system for Indian population: An exploratory study done in South India. *Medical Journal, Armed Forces India*, 77(3), 302–311.
- [4] Goel, Rati, Heart Disease Prediction Using Various Algorithms of Machine Learning (July 12, 2021). Proceedings of the International Conference on Innovative Computing & Communication (ICICC)2021.
- [5] J, S. K., & Geetha, S. (2019). Prediction of Heart Disease Using Machine Learning Algorithms.
- [6] Ansari, M.F., Alankar, B., Kaur, H. (2021). A Prediction of Heart Disease Using Machine Learning Algorithms. In: Chen, J.IZ., Tavares, J.M.R.S., Shakya, S., Iliyasu, A.M. (eds) *Image Processing and Capsule Networks. ICIPCN 2020. Advances in Intelligent Systems and Computing*, vol 1200. Springer, Cham.
- [7] Marikani, T., & Shyamala, K. (2017). Prediction of Heart Disease using Supervised Learning Algorithms. *International Journal of Computer Applications*, 165(5), 41–44.
- [8] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, 7(2.8), 684.
- [9] (n.d.-a). Heart Disease Prediction using Machine Learning Techniques. *International Journal of Engineering Research & Technology (IJERT)*, Volume 09(Issue 11 (November 2020)).
- [10] Heart Attack Prediction Using Machine Learning Algorithms, *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) ICEI – 2022, ICEI – 2022 (Volume 10) (Issue 11)*.
- [11] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. *IOP Conference Series: Materials Science and Engineering*, 1022(1), 012072.
- [12] (n.d.-a). Heart Disease Prediction using Machine Learning. *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT), NCETER – 2021 (Volume 09 – Issue 11)*.
- [13] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access*, 7, 81542–81554.
- [14] Hassan, C. H. C., Khan, M. S., & Shah, M. A. (2018). Comparison of Machine Learning Algorithms in Data classification.
- [15] <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>