International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

*Research Article*

# Optimizing Heart Disease Prediction Accuracy using Machine Learning Models

**Nidhi Singh[1*] and Bhanu Pratap Singh[2]**

Computer Science & Engineering, VITS Engineering College, Satna, M.P., India

Received 26 March 2025, Accepted 20 April 2025, Available online 23 April 2025, **Vol.15, No.2 (March/April 2025)**

***Abstract***

*Heart disease remains a major global health challenge and a leading cause of death. Early detection is critical, as untreated conditions can progress rapidly, leading to severe outcomes. Advances in modern medicine, such as electronic health records and connected medical devices, enable continuous health monitoring. However, analyzing the vast data generated by these technologies requires advanced data mining techniques to effectively classify health information and prioritize heart disease detection. Despite these tools, early diagnosis remains a significant hurdle for medical professionals. To address this, accurate and timely prediction systems are essential for saving lives. Our approach emphasizes thorough data preprocessing, including handling missing values, normalization, and encoding categorical features. We employ a range of machine learning algorithms, from traditional methods to advanced models, refining their performance through extensive experimentation and hyperparameter tuning. Feature selection is a critical component, enhancing model interpretability by identifying key predictors of heart disease risk.*

***Keywords:*** *Heart disease, Machine learning Classification, Ensemble Method, Cross-Validation.*

## Introduction

Heart disease is a widespread global health issue and the leading cause of death worldwide, as reported by the World Health Organization (WHO). Cardiovascular diseases (CVDs) are responsible for approximately 17.9 million deaths annually, accounting for 31% of all global fatalities. These diseases encompass a range of conditions affecting the heart and blood vessels, such as heart failure, hypertension, and coronary artery disease [1]. Identifying individuals at risk is essential for implementing preventive strategies and timely interventions to reduce their impact. Traditional risk assessments consider factors like blood pressure, cholesterol levels, age, gender, smoking habits, and family medical history. While these indicators are valuable, they often fail to capture the full complexity of an individual's risk profile. Furthermore, cardiovascular diseases disproportionately impact middle- and low-income populations, contributing to 75% of CVD-related deaths [2]. The situation is particularly concerning in countries like India, where rising cases have led to a surge in open-heart surgeries, highlighting the urgent need for effective prevention and management strategies.

*Corresponding author's ORCID ID: 0000-0000-0000-0000
DOI: https://doi.org/10.14741/ijcet/v.15.2.12

Early detection and intervention are crucial for mitigating the adverse effects of heart disease, which remains one of the leading causes of death worldwide, encompassing various cardiovascular conditions such as coronary heart disease, heart failure, and cardiac arrhythmias, due to their frequent occurrence and often asymptomatic nature in the early stages [3]. Integration of new technologies like machine learning, artificial intelligence, and wearable devices further enhances the accuracy and accessibility of heart disease prediction, allowing for continuous monitoring of physiological parameters, real-time data analysis, and personalized risk assessment, enabling individuals to actively engage in managing their cardiovascular health.

## Related work

Bo Jin, Chao Chi, *et al*. (2018) introduced a model titled "Predicting the Risk of Heart Failure with EHR Sequential Data Modelling," employing neural networks to analyse real-world electronic health records (EHR) related to congestive heart disease. The study emphasized the importance of maintaining the chronological order of medical records, utilizing one-hot encoding and word vectors to represent diagnostic events, and extending memory network models for predicting coronary failure events [4].

Senthilkumar Mohan, Chandrasekhar Tirumala, and their colleagues proposed an approach named

"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" in 2019. This method combined random forest and linear methods, utilizing specific attribute subsets from pre-processed cardiovascular disease datasets for analysis. The hybrid techniques applied were effective in diagnosing cardiovascular disease [5].

Mamatha Alex P. and Shaicy P. Shaji (2019) developed a paper titled "Prediction and Diagnosis of Heart Disease Patients Using Data Mining Techniques," utilizing methodologies such as Artificial Neural Networks, KNN, Random Forest, and Support Vector Machines. Among these techniques, Artificial Neural Networks exhibited higher accuracy in diagnosing heart disease compared to others [6].

Abhishek Tanja's (Year not provided) heart disease prediction system employed various supervised machine learning algorithms like J48, Naïve Bayes, and Multilayer Perception within WEKA Machine Learning software, utilizing 10-fold cross-validation. J48 demonstrated superior performance compared to Naïve Bayes and Neural Networks in certain scenarios [7]. Priti Chandra *et al*. (Year not provided) employed Computational Intelligence Techniques for Early Diagnosis of Heart Disease using WEKA and 10-fold cross-validation. They utilized the Naïve Bayes algorithm, achieving an accuracy of 86.29%. Although the accuracy was deemed good, it fell short of being satisfactory for automatic heart disease diagnosis [8].

Ashok Kumar Dived (Year not provided) evaluated the performance of various machine-learning techniques for predicting heart disease using tenfold cross-validation. Algorithms such as Naïve Bayes, Classification Tree, KNN, Logistic Regression, SVM, and ANN yielded accuracies of 83%, 77%, 80%, 85%, 82%, and 82%, respectively. Logistic Regression exhibited superior accuracy compared to other algorithms in predictive modelling [9].

Megha Shahi *et al*. (Year not provided) introduced a heart disease prediction system utilizing data mining techniques and WEKA software for automated diagnosis and quality assessment in healthcare centers. They employed algorithms like SVM, Naïve Bayes, Association Rule, KNN, ANN, and Decision Tree for analysis, noting that SVM's accuracy was approximately 85% in certain studies, outperforming other data mining algorithms [10].

**Table 1**

| Author | Purpose | Techniques used | Accuracy |
|---|---|---|---|
| Mamatha Alex P. and Shaicy P. Shaji (2019) | Prediction and diagnosis of patients with heart disease | Artificial neural network, ANN, random forest, and support vector machine | ANN has achieved the highest precision in diagnosing heart disease. |
| Abhishek Tanej (2013), 7 | A heart disease prediction system that uses data mining techniques and various supervised machine learning algorithms | J48 | 95.56% |
| | | | 92.42% |
| | | Multilayer perception | 94.85% |
| Priti Chandra *et al* (2015), 17 | Early Diagnosis of Heart Disease | Native Bayes | 86.29% |
| Ashok Kumar Dwivedi (2016), 13 | Evaluate the performance of various machine learning techniques for heart disease prediction using tenfold cross-validation | Native Bayes | 83% |
| | | Classification Tree | 77% |
| | | KNN | 80% |
| | | Logistic Regression | 85% |
| | | SVM | 82% |
| | | ANN | 84% |
| Megha Shahi*et al* (2017), 16 | Heart disease prediction system using data mining techniques. | SVM, Naïve Bayes, Association rule, KNN, ANN and Decision Tree | SVM has an effective and efficient accuracy of about 85% compared to other algorithms. |
| Syed Muhammad Saqlain Shahid al (2017), 19 | Heart disease diagnosis analysis base d on feature extraction using K-fold | SVM | 91.30% is the highest accuracy achieved. |

**Data Source:** The dataset sourced from IEEE Dataport.org contains extensive data on the human heart and its health status. It comprises 11 features and a target variable, with 6 nominal and 5 numeric attributes. The "target" field indicates the presence of heart disease, with 0 indicating no disease and 1 indicating its presence. Below are descriptions of the attributes and their relevance for research purposes.

**Table 2** Description of the Dataset

| | | |
|---|---|---|
| 1 | Age | Age of the patients in years (number) |
| 2 | Sex | Patient gender (male: 1, female: 0) (nominal) |
| 3 | Type of chest pain | type of chest pain experienced by the patient, divided into 1 typical, 2 typical angina pectoris, 3 non-angina, and 4 asymptomatic (nominal) |
| 4 | Resting basal points | resting blood pressure in mm/HG (numeric) |
| 5 | Cholesterol | Serum cholesterol in mg/dl (numeric) |
| 6 | Fasting blood sugar | Fasting blood sugar > 120 mg/dl means 1 for true and 0 for false (nominal) |
| 7 | Resting ECG | The result of the resting electrocardiogram is presented in 3 different values 0: Normal 1: ST-T wave abnormality 2: Left ventricular hypertrophy (nominal) |

| | | |
|---|---|---|
| 8 | maximum heart rate | maximum heart rate reached (numerical) |
| 9 | Exercise angina | 0 stands for no, and 1 stands for yes (nominal) |
| 10 | Old peak | ST segment depression due to physical activity compared to rest (numerical) |
| 11 | ST slope | The ST segment measured as the slope during peak load 0: Normal 1: Up sloping 2: Flat 3: Down sloping (nominal) |
| 12 | Target | This is the target variable that we need to predict. 1 means the patient is at risk for heart disease and 0 means the patient is healthy. |

## Methodology

The predictive model for heart disease classification follows a systematic methodology. It begins with data pre-processing tasks, handling missing values, encoding categorical variables, and removing duplicates. Outlier detection techniques like Z-score are then applied for data integrity. Feature selection is carried out to choose relevant attributes and define the target variable. Correlation analysis helps identify redundant variables. The data is split into training and test sets, maintaining class balance. Normalization techniques ensure consistent feature scaling. Cross-validation methods assess model generalization. Various classification algorithms are used, trained, and tuned for optimal performance. Model evaluation is based on metrics like accuracy, precision, recall, and AUC-ROC. Feature selection methods are employed to identify influential features. Models are re-evaluated based on selected parameters, completing the iterative refinement process.

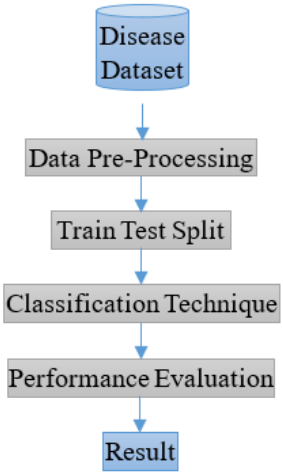


**Figure 1:** Propose Model

**Performance Metrics:** Performance metrics in machine learning assess the effectiveness of an algorithm across different criteria like accuracy, precision, convergence, and more.

**Confusion Metrics:** Confusion metrics aid in evaluating a model's performance by organizing classification outcomes, and distinguishing between actual and predicted values. It delineates true positives (correctly identified positive outcomes), false positives (incorrectly identified negative outcomes as positive), false negatives (mistakenly identified positive outcomes as negative), and true negatives (correctly identified negative outcomes).

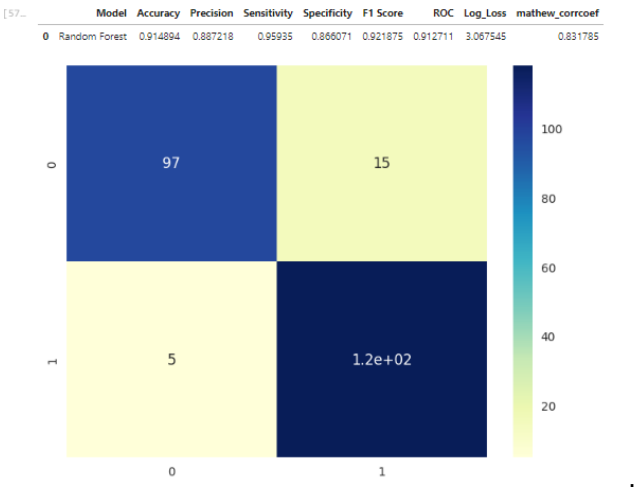


**Figure 2:** Confusion Metrics

To evaluate the model's performance, an accuracy score is used, calculated by ranking the true positives and true negatives and then dividing by the number of true positives, true negatives, negative positives, and negative negatives. this is the way

$$Accuracy = \frac{TP + TN}{TP = TN + FP + FN}$$

Accuracy indicates how close the predicted value is to the actual value. For example, if a person's actual height is 1.8 meters and the measured value or reported value is 5.9, this is the correct measurement.

After truth there is a certainty, that is, the degree of accuracy that accurately reflects judgments about evil and evil; This demonstrates its importance in correctly distinguishing and classifying samples. bad judgment. It is also called the true negative scale. this is the way

$$Specificity = \frac{TN}{TN + FP}$$

There is also sensitivity in estimating the actual number of cases well (or better). Sensitivity is also called memory. In other words, it is predicted that an unhealthy person will be unhealthy. this is the way

$$Sensitivity = \frac{TP}{TP + FN}$$

Accuracy is defined as the percentage of actual cases classified in good positive [5, 178]. It shows how close the predicted values are. Below you can read how to make the correct calculation.

$$Precision = \frac{TP}{TP + FP}$$

Note how close or close the predicted values are. For example, if the actual value of a person's height is 6 meters, the measured value is also reported. The value is 5.5 and if the height is measured as 5.4 or 5.6 the prediction is accurate but not very accurate.

F1 points: The F1 Score of is defined as a ratio that combines reality and memory and tries to find a balance between them. Here's how to calculate the F1 score.

$$F1\ Score = \frac{2 * Precision * Senstivity}{Precision + Sensitivity}$$

An F1 score indicates how well the class performs in recognition and memory.

So, it's good to have more value for both, and don't forget to have a good F1. Points The definition or recall value is lower because all F1 points are reduced.
The main evaluation metrics of this field are sensitivity, specificity, accuracy, F1 ratio, and ROC-AUC curve. Additionally, we use two other performance metrics that are reliable statistical measures: the Matthew Coefficient (MCC) and Log Loss.

The Matthews coefficient (MCC) is a statistical measure of how well the prediction performs for both values across all four categories of the confusion matrix (true positive, negative, true negative, and false positive). The number and size of negative features in the dataset.

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP + FP).(TP + FN).(TN + FP).(TN + FN)}}$$

Loss trees measure the performance of models where the expected input has a probability value between 0 and 1. Our machine-learning model aims to minimize value. The loss of a good model will be 0. Since it is said that it is possible to deviate from the real label, the loss of the tree increases. Therefore, estimating the probability of 0.012 when the actual observation signal is 1 would be inaccurate and result in a large algorithm.

**Result & Discussion**

These findings suggest that despite the widespread adoption of algorithms such as SVC and Decision Trees for diagnosing patients with heart disease, our utilization of KNN, Random Forest Classifier, and Logistic Regression outperforms them [12]. Not only do our selected algorithms demonstrate superior accuracy, but they also provide cost-efficiency and quicker processing compared to those utilized by earlier researchers. Furthermore, the maximum accuracies attained by KNN and Logistic Regression, reaching 88.5%, either match or surpass the accuracies reported in prior studies. After training and evaluating ten machine learning models and comparing their performances, it was observed that the Random Forest model utilizing the entropy criterion consistently outperformed the others, achieving an accuracy of 90.63%. Following this, we employed a majority vote feature selection technique, incorporating two filter-based, one wrapper-based, and three embedded feature selection methods. Post feature selection, the Random Forest model maintained its superiority, demonstrating an accuracy of 89.36%. Remarkably, this represented a minimal decrease of less than 1% in accuracy compared to its performance before feature selection.

The Random Forest model showed the best performance on cross-validation. Consequently, we'll evaluate the Random Forest model. Table 3 Show the result of best performance.

| Model | Accuracy | Precision | Sensitivity | Specificity | F1 Score | ROC | Log Loss |
|---|---|---|---|---|---|---|---|
| Random Forest | 0.914894 | 0.887218 | 0.95935 | 0.866071 | 0.921875 | 0.912711 | 3.067545 |

Figures 2, 3, 4, and 5 display plots illustrating the distribution of patients segregated and predicted by the classifier across different age groups, resting blood pressure levels, sex, and types of chest pain.

Number of Heart Disease Patients: 628

Number of Non-Heart Disease Patients: 561

Percentage of Heart Disease Patients: 52.817493692178296

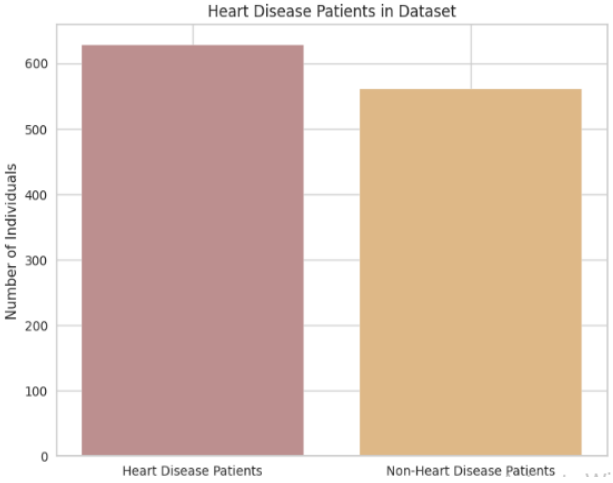


**Figure 3** Distributions of Age and Gender

As indicated in the figures above, the average dataset comprises 629 heart disease patients and 561 normal patients.

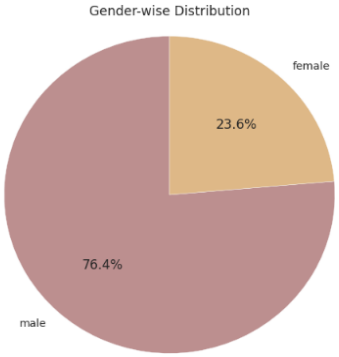


**Figure 4:** Distributions of Age and Gender

The graph below indicates a higher percentage of men compared to women in the dataset, while the histogram suggests an average patient age of 55.

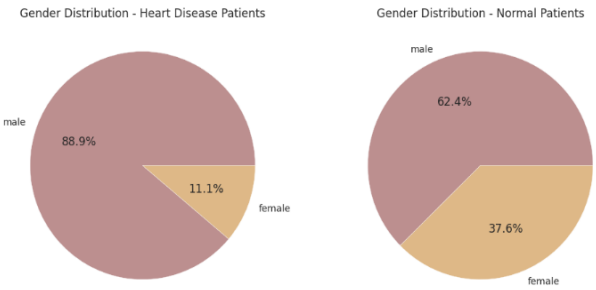


**Figure 5:** Heart disease-wise distribution of Age and Gender

In the graph provided, male patients exhibit a higher likelihood of heart disease compared to females, with heart patients typically having an average age ranging between 58 and 60 years.

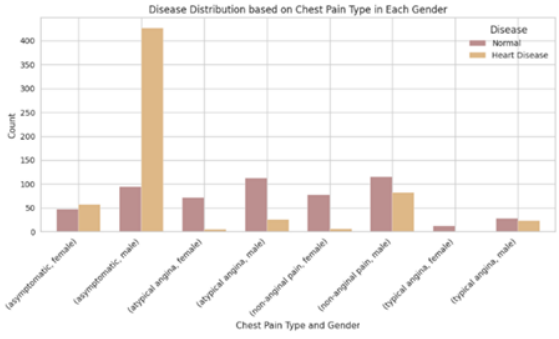


**Figure 6** Distribution of Chest Pain Type

About 76.91% of heart patients experiencing chest pain show no visible signs of discomfort. Silent myocardial infarction (SMI), accounting for 45-50% of heart disease-related deaths in India, manifests with mild symptoms, earning it the ominous nickname "silent killer" due to its subtle nature, often mistaken for chronic issues, leading to underestimation and neglect.

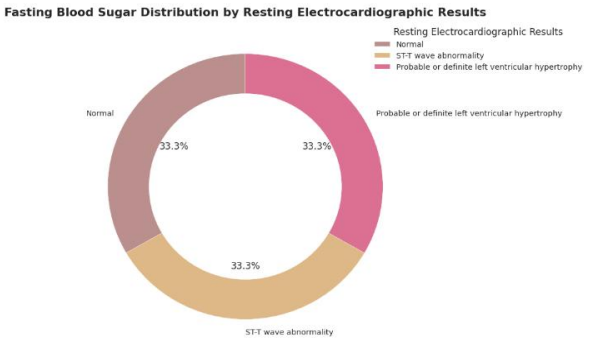


**Figure 7**: Distribution of resting electrocardiographic (ECG)

While Electrocardiogram (ECG) is adept at measuring heart rate and rhythm, it may not consistently identify blockages in the arteries. Consequently, approximately 52% of heart disease patients exhibit normal ECG results.

## Conclusion

In this study, we conducted 10-fold cross-validation to assess the performance of 13 machine learning algorithms with various hyperparameters. Following this, we trained and evaluated these models on the test set, ultimately identifying the top three performers. Notably, a stacked ensemble of powerful algorithms demonstrated superior performance over individual models. Specifically, the Random Forest model with entropy criterion exhibited the highest accuracy of 90.63%. Even after applying feature selection techniques, the Random Forest model remained the best performer, with a negligible decrease in accuracy. The most influential features identified through feature importance analysis were Cholesterol, Max Heart Rate achieved, and ST Depression.

## Acknowledgment

I would like to express my sincere gratitude to my research guide, Mr. Bhanu Pratap Singh, for his invaluable guidance, continuous support, and insightful suggestions throughout the course of this research. His expertise and encouragement have been instrumental in shaping this study.

## References

[1] "cardiovascular diseases (CVDs)." WHO, 2020, https://www.who.int/zh/news-room/factsheets/detail/cardiovascular-diseases-(cvds).
[2] Mijwil, M. M., Al-Mistarehi, A. H., and Mutur, D. S. conducted a literature review titled "The Practices of Artificial Intelligence Techniques and Their Value in Addressing the COVID-19 Pandemic" in the Mobile Forensics Journal in 2022, volume 4, issue 1, pages 11-30.
[3] The impact of maternal pre-pregnancy/early-pregnancy body mass index (BMI) and pregnancy smoking and alcohol on congenital heart diseases was investigated through a parental negative control study (Taylor *et al*., 2021).
[4] Bo Jin, Chao Che, Zhen Liu, Shillong Zhang, Xiaomeng Yin, And Xiaoping Wii, "Predicting the Risk of Heart Failure with EHR Sequential Data Modelling", IEEE Access 2018.

[5] "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," Senthilkumar Mohan, Chandrasekar Thirumalai, and Gautam Srivastava explore advanced methodologies for accurately predicting heart disease, as published in IEEE Access in 2019.
[6] Mamtha Alex P and Shaicy P Shaji, "Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique", International Conference on Communication and Signal Processing 2019.
[7] A. Taneja, ")riental journal OF Heart Disease Prediction System Using Data Mining Techniques," 2013.
[8] N. O. Fowler, "Diagnosis of Heart Disease," vol. V, no. March, pp. 1–7, 2012.
[9] Ashok Kumar Dwivedi explores the use of computational intelligence techniques in predicting diabetes mellitus in his article titled "Analysis of computational intelligence techniques for diabetes mellitus prediction," published in Neural Compute. Appl., vol. 13, no. 3, pp. 1–9, 2017.
[10] M. Shahi and R. Kaur Germ, "heart disease prediction system using data mining techniques," Orient. J. Compute. Sci. Technol., vol. 6, no. 4, pp. 457–466, 2013.
[11] S. M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, and S. A. Hussain researched heart disease diagnosis using parallel Probabilistic Principal Component Analysis to extract relevant features. A Stat. Mech. its Appl., vol. 482, pp. 796–807, 2017.
[12] Folsom, A. R., Princes, R. J., Kaye, S. A., and Solar, J. T. (1989) investigated the association between body fat distribution and the self-reported prevalence of hypertension, heart attack, and other heart diseases in older women.
[13] Gour, S., Panwar, P., Dwivedi, D., and Mali, C. (2022). 'Intelligent Sustainable Systems,' published by Springer in Singapore, the chapter titled 'A Machine Learning Approach for Heart Attack Prediction' spans pages 741 to 747."
[14] Y. Ali, R. Amir, and A.-M. Fardin conducted a study titled "Profile-based assessment of diseases affecting factors using fuzzy association rule mining approach: a case study in heart diseases," published in the Journal of Biomedical Informatics in 2021.
[15] Ghwanmeh, S., Mohammad, A., & Al-Ibrahim, A. (2013). "Innovative artificial neural network-based decision support system for heart disease diagnosis." Journal of Intelligent Learning Systems and Applications, 5(3), 353-396.
[16] A review on genetic algorithm models for hadoop mapreduce in big data Chandra Shekhar Gautam1, Pandey* (2019) International Journal of Recent Scientific Research ISSN:0976-3031, Vol.13, Issue-03(E), Pageno771-775, June 2022
[17] Clustering of Bigdata Using Genetic Algorithm in Hadoop MapReduce Chandra Shekhar Gautam, Mr. L N SONI, P Pandey European chemical bulletin Year 2022, issue 12,963-973