IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 7, July 2023, Pages 440-445

DOI: 10.17148/IARJSET.2023.10762

# Comparative Study of Various Machine Learning Algorithms for Heart Disease Prediction

**Aditi Sahal[1], H K Madhu[2]**

Student, Department of MCA, Bangalore Institute of Technology, Bangalore, India[1]

Associate Professor, Department of MCA, Bangalore Institute of Technology, Bangalore, India[2]

**Abstract**: The major cause of death worldwide in recent years has been heart disease. This issue is being raised globally as a result of changes in dietary habits, working cultures, and other aspects of daily living. The creation of a technique that can identify early indications and so save many lives is one method for treating and diagnosing this disease. Researchers can estimate the prevalence of heart disease in high-risk groups using machine learning (ML) techniques. For efficient prevention, management and treatment of diseases, it is crucial to create precise and trustworthy approaches for early illness prediction through automation. In previous publications, multiple experts have discussed efforts to create the optimum techniques for predicting heart disease. This study compares three commonly used heart disease prediction methods. These results can be utilised to assist in creating precise and effective models that aid physicians in lowering the count of heart disease fatalities. The cardiovascular systems of three ML algorithms—Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF)—are compared in this study.

**Keywords**: Logistic Regression, Support Vector Machine, Random Forest, Heart disease prediction.

## I. INTRODUCTION

Heart disease, further known as cardiovascular disease, is a leading contributor of deaths globally and affects an abundance of people every year. It includes several disorders that have a bearing on the heart and blood vessels, inclusive of arrhythmias, heart failure, valvular heart disease and coronary artery disease. Smoking, high blood pressure, and cholesterol, obesity, diabetes, sedentary lifestyle, family history, as well as older age are all risk factors for heart disease. Effective management and the avoidance of serious consequences rely on timely detection and prompt intervention. A subfield of AI (artificial intelligence) called as machine learning (ML), which has recently become a potent tool in the healthcare field, has the prospect to detect and diagnose cardiac disease. Large-scale patient data can be analysed by ML algorithms, which can spot patterns, connections, and risk factors that human specialists would miss. Utilising this technology, healthcare providers are able to anticipate outcomes more accurately, allowing for early interventions and better patient outcomes.

## II. LITERATURE SURVEY

The studies referred to highlight the significance of accurate diagnosis and risk assessment in preventing heart diseases and enhancing patient care through advanced automation and AI technologies.

Aritra Chakraborty et al. [1] in their study "Comparative Study of Myocardial Infarction Detection from ECG Data Using Machine Learning" investigate current research in the domain of ECG-based myocardial infarction diagnosis. The evaluation evaluates numerous machine learning algorithms and approaches used for this goal, with an emphasis on accuracy and efficiency. It also illustrates the obstacles and limits of present techniques, as well as prospective areas for development. Aritra's research intends to provide helpful information to optimize the accuracy and reliability of ECG-based myocardial infarction identification through the use of machine learning algorithms.

The methods utilised by Ashok Kumar Dwivedi [2], included Logistic Regression, Naive Bayes, SVM, Naive Bayes, Classification Tree, KNN, and ANN. Logistic regression achieved the accuracy rate having the highest percentage among these.

Data mining procedures were engaged by Muthuvel et al. [3] to forecast heart disorders. The medical professional was helped by this study to enhance decision-making in relation to a particular parameter. It obtained an accuracy of 86.3% in testing and that of 87.3% in training by training and testing a specific parameter.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

ISO 3297:2007 Certified

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 7, July 2023, Pages 440-445

DOI: 10.17148/IARJSET.2023.10762

Kishore et al. [4] proposed "Heart Attack Prediction Using Deep Learning" where to forecast the probable elements of heart associated illnesses of the patient, Recurrent Neural System is employed. This model employs deep learning and also data mining to present the most accurate model with the fewest errors. For other heart attack risk assessment models, this study serves as a reliable reference model.

"Comprehensive Analysis on Detecting Chronic Kidney Disease" by Mirza Muntasir Nishat et al. [5] leverages machine learning algorithms to predict CKD with 99.75% accuracy using random forest. The study highlights the significance of kidneys and showcases the applicability of supervised machine learning in bioinformatics for early-stage detection.

In the paper "Efficient Medical Diagnosis of Human Heart Diseases," Ahmad et al. [6] achieved 100% and 99.03% accuracy on various datasets using Extreme Gradient Boosting Classifier with GridSearchCV. The study aims to improve heart disease diagnosis and considers real-world relevance.

Ramesh et al.'s [7] study proposes the information gain-based method to accurately predict heart attacks, identifying influential factors such as sex, maximum heart rate, angina, and fasting blood sugar. The Support Vector Machine and Random Forest achieved 88% accuracy when implementing IGFS.

Researchers are also exploring feature importance at the algorithmic level. For example, Alam et al. suggested using Random Forest algorithm for feature importance analysis on ten datasets, including three heart disease datasets [8].

Ayon et al. [9] compared seven computational intelligence techniques for coronary heart disease prediction. Deep neural network achieved 98.15% accuracy on Statlog dataset, while SVM achieved 97.36% on Cleveland dataset. Study suggests exploring additional techniques and plans to create real-time website for accessible healthcare solutions.

Asif et al. [10] compared machine learning strategies. Hard and soft voting ensemble classifiers achieved best accuracy (92%). Adaboost showed promise with highest accuracy (0.938) and specificity (0.926). They introduced RSCV, a computationally efficient approach, supporting coronary disease management and healthcare diagnosis.

## III. METHODOLOGY

### *A. Data Collection*

The dataset utilized in this study is accessible from the UCI Repository. It originated in 1988 and comprises a fusion of four separate datasets obtained from Long Beach V, Switzerland, Hungary, and Cleveland. In total, the dataset encompasses 75 attributes, excluding the predicted attribute. However, all the research studies mentioned here employed a subset of 14 attributes. The "target" field represents the predicted attribute, resulting in a total of 76 columns in the dataset. The dataset contains 303 instances, and a thorough description of the attributes is provided in the accompanying Fig 1.

| S. No. | *Attribute* | *Description* | *Values* |
|---|---|---|---|
| 1 | Age | Age of patient's in years | 20 - 77 |
| 2 | Sex | Gender of patient (1: Male, 0: Female) | 0, 1 |
| 3 | Cp | Chest pain type | 1, 2, 3, 4 |
| 4 | Trestbps | Resting blood pressure in mm Hg | 94 - 200 |
| 5 | Chol | Serum cholesterol in mg/dl | 126 - 264 |
| 6 | Fbs | Fasting blood in mg/dl sugar>120 then 1(true): else 0 (false) | 0, 1 |
| 7 | Resting | Resting electrocardiographic Result | 0, 1, 2 |
| 8 | Thalach | Maximum Heart Rate Achieved | 71 - 202 |
| 9 | Exang | Exercise Included Angina (1:Yes, 0:No) | 0, 1 |
| 10 | OldPeak | ST Depression Introduced by Exercise Relative to Rest | 1 - 3 |
| 11 | Slope | Slope of the Peak Exercise ST Segment (1:up-sloping, 2:flat, 3: down-sloping) | 1, 2, 3 |
| 12 | Ca | Number of Major Vessels | 0 - 3 |
| 13 | Thal | 3- Normal, 6-Fixed Defect, 7-Reversible Defect | 3, 6, 7 |
| 14 | Target | Presence heart disease (0: No, 1, 2, 3 4:Yes) | 0, 1, 2, 3, 4 |

Fig. 1 Dataset Description


This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 7, July 2023, Pages 440-445

DOI: 10.17148/IARJSET.2023.10762

*B. Algorithms*

*1) Logistic Regression:*

Logistic Regression is recognized as one of the most straightforward and machine learning classification algorithms. It is frequently utilized for binary classification tasks in various applications. This algorithm operates on categorical dependent variables, producing discrete or binary outputs of either 0 or 1. The logistic regression model utilizes the sigmoid function as a cost function. The sigmoid function maps the predicted real values to probabilistic values ranging between the range of 0 and 1. Logistic Sigmoid Function is as follows:

$$P(x) = 1/(1 + e^{-x})$$

Here, the probability estimation function P(x) is used to provide a value ranging from 0 to 1. X stands for the probability function's input (prediction value of the algorithm). The mathematical constant e represents Euler's number, approximately equal to 2.71828, as illustrated in the equation above.

To forecast the existence of cardiac disease, logistic regression is employed. Initially, the logistic regression model is trained using a specified splitting condition. It is then tested using test data in order to obtain maximum accuracy and examine the model's behaviour. The algorithm divides the output into two categories: 1 and 0, which respectively indicate the presence or lack of heart illness.

*2) Support Vector Machine:*

SVM, a well-known algorithm for predicting coronary disease, is another option. Vapnik and Cortes came up with the idea, and it has been effectively used to solve a number of gender classification issues. A separating hyperplane is chosen by SVM, a linear classifier, to reduce anticipated classification error regarding test patterns that are still to be seen. It has the capability to recognize patterns and divide them into two groups. Based on the greatest distance to the nearest point in the training set, SVM trains a model to ascertain which class a test image belongs to. Even with limitations to single-pose (frontal) detection, SVM takes a significant quantity of training data to establish an appropriate decision border, and its computing cost is expensive. SVM is a learning algorithm that looks for the best separating hyperplane in order to minimize the expected classification error for patterns that have not yet been observed. For linearly non-separable data, SVM maps the input to a high-dimensional feature space where separation is possible using a hyperplane. This projection into high-dimensional feature space is efficiently performed using kernels. The objective of SVM is to identify the best separating hyperplane, represented by the equation $\omega\, Tx + b = 0$ that maximizes the distance segregating the two classes. Here, the vector ω represents the normal vector to the hyperplane, which is perpendicular to the hyperplane. The parameter b in the equation represents the offset or distance of the hyperplane from the origin along the normal vector ω.

*3) Random Forest:*

Random Forest (RF) serves as a machine learning approach applied to both regression and classification tasks. It employs ensemble learning, combining multiple decision trees to provide solutions for complex problems. A random forest algorithm consists of numerous decision trees, and the forest is trained through a process termed as bagging or bootstrap aggregating which is an ensemble meta-algorithm used to enhance the accuracy of machine learning algorithms. The random forest algorithm generates predictions by aggregating the predictions of individual decision trees. The outcome from each tree is averaged out to produce the final prediction. The accuracy of the result is improved by increasing the random forest's tree count. By lowering dataset overfitting and boosting prediction accuracy, random forest solves the drawbacks of a single decision tree approach. It allows the generation of predictions without requiring extensive configuration in packages like scikit-learn.

*C. Performance Evaluation Metrics*

In order to analyse the efficacy of machine learning models, performance evaluation is essential. The performance of these models is measured and described using a variety of ways. The following evaluation statistics were employed in this study:

1) True Positive (TP): Measure of the quantity of instances of heart disease the model properly identified as positive (true).

2) True Negative (TN): Measure of the quantity of instances of heart illness the model accurately identified as false positives (negative).

3) False Negative (FN): Measure of the quantity of instances of heart disease the model misclassified as negative (false).

4) False Positive (FP): The number of instances of heart disease that the model misclassified as positive (true).

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

ISO 3297:2007 Certified

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 7, July 2023, Pages 440-445

DOI: 10.17148/IARJSET.2023.10762

Other metrics were calculated to further evaluate the model's performance. Some of these metrics employed here include:

5) Sensitivity (Recall/True Positive Rate): This indicator counts the percentage of positive samples that were certainly positive but were mistakenly labelled as such.

6) Precision: The fraction of accurately classified positive instances to all positively classified instances is known as precision.

7) Accuracy: Taking into consideration both true positive and true negative predictions, accuracy is a metric of the proportion of accurate detections made by the model.

8) F1 Score: The F1 score is a statistic that combines recall (sensitivity) and precision (accuracy) into one number. It stands for the harmonic midpoint between recall and precision.

9) ROC curve: The efficacy of the model in binary classification is depicted graphically by the ROC (Receiver Operating Characteristic) curve. At various classification thresholds, it demonstrates the trade-off between sensitivity (the true positive rate) and the false positive rate. An indicator of the model's overall performance is the AUC-ROC (area under the ROC curve), with an elevated value indicating better performance. The AUCROC has a value ranging from 0 to 1, with a number closer to 1 suggesting a more effective algorithm performance.

Besides these, we also measure an average cross validation accuracy (after performing 10 fold cross validation on the models) to get a more robust measure of the performance of the models and their generalising ability.

## IV. RESULTS AND DISCUSSION

To draw a comparison amongst the three algorithms, we'll tabulate the performance evaluation metrics (accuracy, precision, recall and F1 score) for the correct prediction of heart disease to be able to visualise these parameters side-by-side and analyse.

TABLE I METRICS FOR THE PERFORMANCE OF THE MODELS

| Algorithms | Accuracy (%) | Precision (%) | Recall (%) | F1 score (%) | Avg. (10 Fold) Cross Validation Accuracy(%) |
|---|---|---|---|---|---|
| Logistic Regression | 89 | 88 | 91 | 89 | 82.84 |
| Support Vector Machine (Using GridSearchCV) | 90 | 91 | 91 | 91 | 83.49 |
| Random Forest | 90 | 96 | 84 | 90 | 83.46 |

With the highest measures for recall, accuracy and F1 score along with the average cross validation accuracy, among the criteria; SVM with hyper parameter tuning surpasses the other models. Random Forest tends to achieve the highest scores for recall. However, Logistic Regression (LR) exhibits the poorest performance across accuracy, precision, and F1 score.

The ROC curve pertaining to each classifier is displayed in Fig. 2. An indicator of the classifier's overall performance is the AUC (area under the curve), with a number closer to 1 indicating more successful performance. The graph's solid orange line, which possesses the largest area, is the ROC curve associated with Support Vector Machine. This shows that SVM surpassed the other classifiers in terms of performance.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

ISO 3297:2007 Certified

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 7, July 2023, Pages 440-445

DOI: 10.17148/IARJSET.2023.10762

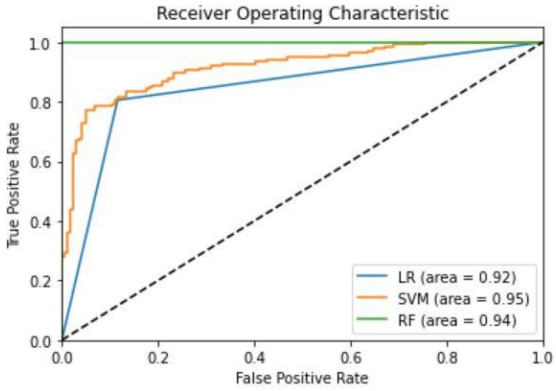


Fig. 2 ROC curves for the three models

Another bar graph is plotted to visualise the models' performance for the different metrics set and we can see again, clearly that SVM displays the best performance amongst the three.

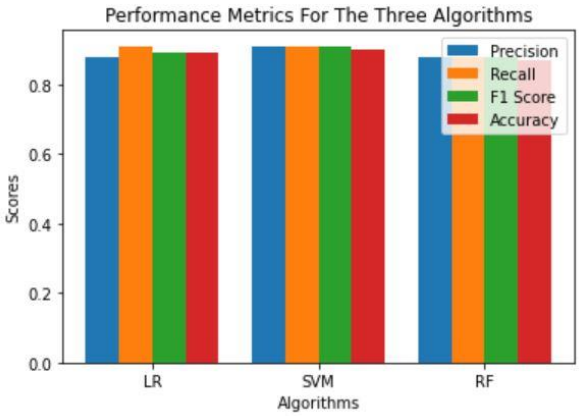


Fig. 3 Bar graph for the models' performance metrics

## V. CONCLUSION

This research compared Logistic Regression, Random Forest, and Support Vector Machine models for heart disease prediction. SVM achieved the highest accuracy of 90% and an impressive ROC-AUC score of 0.95. The study provides valuable insights for researchers and medical specialists. Possible future research directions include exploring ensemble methods like stacking, boosting, or bagging to enhance heart disease prediction models. In-depth feature engineering and selection techniques, along with investigating domain-specific knowledge, can improve model performance. Temporal analysis using longitudinal data and external validation on diverse datasets are essential steps as well. Implementing the best-performing model in healthcare systems requires addressing deployment challenges and ensuring compliance with medical regulations.

## REFERENCES

[1]. A. Chakraborty, S. Chatterjee, K. Majumder, R. Shaw, and A. Ghosh, "A Comparative Study of Myocardial Infarction Detection from ECG Data Using Machine Learning," 2021. doi: 10.1007/978-981-16-2164-2_21.

[2]. A. K. Dwivedi, "Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation," *Springer*, 17 September 2016.

[3]. M. Muthuvel, M. Marimuthu, M. Abinaya, K. Hariesh, K. Madhankumar, and V. Pavithra, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach," *International Journal of Computer Applications*, vol. 181, pp. 975-8887, 2018. doi: 10.5120/ijca2018917863.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

ISO 3297:2007 Certified

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 7, July 2023, Pages 440-445

DOI: 10.17148/IARJSET.2023.10762

[4]. A. Kishor, A. Kumar, K. Singh, M. Punia, and Y. Hambir, "Heart Attack Prediction Using Deep Learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 05, issue 04, pp. 4420, Apr 2018.

[5]. M. M. Nishat, F. Faisal, R. R. Dip, S. M. Nasrullah, R. Ahsan, F. Shikder, and M. A. Hoque, "A comprehensive analysis on detecting chronic kidney disease by employing machine learning algorithms," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 7, no. 29, pp. e1-e1, 2021.

[6]. G. N. Ahmad, H. Fatima, S. Ullah, and A. S. Saidi, "Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV," *IEEE Access*, vol. 10, pp. 80151-80173, 2022.

[7]. Gowtham Ramesh, Madhavi Karanam, P. Reddy, J. Somasekar, and Joseph Tan, "Improving the accuracy of heart attack risk prediction based on information gain feature selection technique," *Materials Today: Proceedings*, 2021. doi: 10.1016/j.matpr.2020.12.079

[8]. S. A. Suha and M. N. Islam, "Exploring the dominant features and data-driven detection of polycystic ovary syndrome through modified stacking ensemble machine learning technique," *Heliyon*, vol. 9, no. 3, e14518, 2023, ISSN 2405-8440.

[9]. S. I. Ayon, M. M. Islam, and M. R. Hossain, "Coronary artery heart disease prediction: a comparative study of computational intelligence techniques," *IETE Journal of Research*,, vol. 68, no. 4, pp. 2488-2507, 2022.

[10]. M. A. A. R. Asif, M. M. Nishat, F. Faisal, R. R. Dip, M. H. Udoy, M. F. Shikder, and R. Ahsan, "Performance Evaluation and Comparative Analysis of Different Machine Learning Algorithms in Predicting Cardiovascular Disease," *Engineering Letters*, vol. 29, no. 2, 2021.

[11]. Hidayatullah Arghandabi and Parvaneh Shams, "A Comparative Study of Machine Learning Algorithms for the Prediction of Heart Disease," *International Journal for Research in Applied Science and Engineering Technology*, vol. 8, pp. 677-683, 2020.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

ISO 3297:2007 Certified