International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Performance Comparison of Machine Learning Models for Heart Disease Prediction

Vijay Kumar Samyal[1], Satyam Singh[2]

[1]Associate Professor, Department of CSE, MIMIT MALOUT
[2]Student, Department of CSE, MIMIT MALOUT

Abstract

Heart-related diseases, also known as Cardiovascular Diseases (CVDs), are a major cause of death worldwide, and early prediction is essential for timely treatment. Machine learning techniques are increasingly used to analyze medical datasets and support clinical decision-making. In this study, three machine learning algorithms—Logistic Regression, Support Vector Machine (SVM), and Random Forest—are applied to the UCI Heart Disease dataset, which contains 14 clinical features. The dataset is split into an 80:20 ratio for training and testing, and feature scaling is performed where required. Model performance is evaluated using accuracy, precision, recall, F1-score, and confusion matrix. The results show that the Random Forest model achieves the highest performance due to its ability to capture non-linear patterns, while SVM also performs well, and Logistic Regression serves as a strong baseline. The findings highlight that machine learning models, especially ensemble methods, can effectively assist in early heart disease prediction.

Keywords: Health Disease Prediction, Machine Learning, Medical Diagnosis, Clinical Data Analysis

1. INTRODUCTION

Heart disease is one of the leading health challenges in the world today and continues to affect millions of people every year. The increasing number of cases is mainly due to changing lifestyles, stress, lack of physical activity, and unhealthy eating habits. Many heart conditions develop slowly and remain unnoticed until they become serious, which makes early detection extremely important. Accurate and timely prediction of heart disease can help patients receive treatment sooner, reduce complications, and potentially save lives.

With the growth of digital health data, machine learning has become a powerful tool for analyzing medical information and supporting clinical decisions. Machine learning models are capable of identifying hidden patterns in patient datasets and can provide fast and reliable predictions. These technologies can help doctors by acting as decision-support systems, reducing manual workload and improving diagnostic accuracy.

In this research, three commonly used machine learning algorithms—Logistic Regression, Support Vector Machine (SVM), and Random Forest—are applied to the UCI Heart Disease dataset. The dataset contains 14 important medical features related to heart health. The models are trained, tested, and compared using evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix. The goal of this study is to determine which algorithm performs best for heart disease prediction and to show how machine learning can be used as an effective tool for early diagnosis.

2. LITERATURE REVIEW

Many researchers have worked on using machine learning techniques to predict heart disease and improve medical diagnosis. Early studies mainly used traditional statistical methods such as Logistic Regression, which showed that simple linear models can provide a basic understanding of how different medical factors contribute to heart disease [1]. These models were effective but could not fully capture complex relationships in the data.

A variety of advanced techniques have also been introduced in past work. For example, some papers have used neuro-fuzzy systems to combine the learning ability of neural networks with the reasoning capability of fuzzy logic. Other studies have summarized commonly used heart disease prediction techniques and discussed their strengths, weaknesses, and computational complexities. Traditional classifiers like Naïve Bayes have also been tested for heart disease detection, showing that simple probabilistic models can sometimes produce competitive results [1] [2].

Researchers have also explored other machine learning models, such as K-Nearest Neighbors (KNN), Naïve Bayes, Decision Trees, and Neural Networks. While some of these methods provide good results, ensemble-based models, especially Random Forest, consistently achieve higher performance in heart disease prediction tasks [7]. Most studies agree that ML-based systems can assist doctors by providing fast and data-driven predictions, but they should be used as support tools rather than complete replacements for clinical judgment.

3. METHADOLOGY

This study followed a clear step-by-step process to build and compare three machine learning models for predicting heart disease. The UCI Heart Disease dataset, which contains 14 important medical features, was used for training and evaluation.

3.1. Data Understanding and Preprocessing

The dataset was first examined to check for missing values and to understand the distribution of each feature. Basic statistical summaries were generated, and a correlation heatmap was created to identify which features were more closely related to the target variable. This helped in understanding which factors may play a stronger role in prediction.

3.2. Train-Test-Split

The dataset was then divided into two parts:

- 80% for training the models
- 20% for testing their performance

This split ensures that the accuracy reported reflects how well the model performs on unseen data.

3.3 Feature Scaling

Feature scaling was applied to Logistic Regression and SVM because these models work better when all features are on a similar scale. Random Forest, being a tree-based algorithm, does not require scaling, so it was trained directly using the original values.

3.4 Model Training

Three different models were trained:

- Logistic Regression, used as a baseline linear model
- SVM with RBF kernel, used to capture non-linear patterns
- Random Forest, chosen for its ability to handle complex relationships and reduce overfitting

All models were trained using the same training data so that their results could be compared fairly.

3.5 Evaluation Metrics

After training, each model was tested using the test set. Performance was measured using several metrics such as accuracy, precision, recall, F1-score, and confusion matrix. These metrics provide a balanced understanding of how reliably a model can predict heart disease.

Table 1: Model Overview for Individual Model

Model	**Working Principles**	**Advantage**	**Limitation**
Logistic Regression	Learns a linear relationship between features and the probability of heart disease	Simple fast and easy to interpret	Cannot Capture Complex Linear Pattern
SVM	Find Optimal separation Boundary	High Accuracy	Low Scalability
Random Forest	Builds multiple decision trees and combines their outputs for prediction	Very Accurate and Reduce Overfitting	Harder to interpret and slightly slower than simple models

4. DATASETS DESCRIPTION

The dataset used in this study is the UCI Heart Disease Dataset, which is one of the most widely used datasets for research related to heart disease prediction. It is a well-structured dataset that contains detailed medical information collected from patients, making it suitable for training and evaluating machine learning models. The dataset includes a total of 14 attributes, where 13 are input features and 1 is the target output.

The input features consist of clinically important measurements such as age, gender, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol level (chol), fasting blood sugar (fbs), resting ECG results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression (oldpeak), slope of peak exercise ST segment (slope), number of major vessels (ca), and thalassemia (thal). These attributes contain useful medical information that helps the model analyze the overall heart condition of the patient.

The output feature, known as target, indicates whether a person is likely to have heart disease (1 = presence of heart disease, 0 = absence of heart disease). This binary class label allows machine learning models to perform classification and differentiate between healthy and high-risk individuals.

One of the advantages of this dataset is that it does not contain missing values, which makes the preprocessing step easier and more reliable. Since the data is clean and well-organized, it helps in smoother model training and avoids complications related to imputation or data loss. Each feature captures a specific aspect of a patient’s heart health, and together, these features help the model learn patterns that are useful for identifying potential heart disease cases. Each sample has the following four features:

Table 2: Data Descriptions of Features

Feature Name	**Description**
Age	Age of Patient
Sex	Gender (1 = male, 0 = female)
cp	Chest Pain type
trestbps	Resting blood Pressure
chol	Serum cholesterol in mg/dl
fbs	Fasting blood sugar

restecg	Resting ECG results
Thalach	Maximum heart rate
exang	Exercise-induced angina
oldpeak	ST depression induced by exercise
slope	Slope of the peak exercise ST segment
Ca	Number of major vessels (0–3)
thal	Thalassemia type
target	1= Heart disease present, 0 = No heart disease

5. Experimental Results

Use the results print by the Python Code:

Table 3 Model Evaluation for Individual Models

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.76	0.74	0.85	0.79
SVM	0.76	0.74	0.85	0.79
Random Forest	0.80	0.78	0.87	0.82

5.1. Statistical Justification

The results show that although Logistic Regression and SVM perform similarly, the Random Forest model provides slightly better performance. This is mainly because Random Forest combines multiple decision trees, which helps it handle complex patterns in the medical data more effectively.

The model also achieved the highest recall (0.87), which means it correctly identifies more patients who actually have heart disease. In medical prediction tasks, missing a real case (false negative) can be risky, so a higher recall is an important advantage. The F1-score of Random Forest is also the highest among the three models, showing that it maintains a good balance between precision and recall.

Overall, the statistical metrics suggest that Random Forest is more reliable for heart disease prediction, while Logistic Regression and SVM still perform reasonably well as simpler baseline models.

6. Graph

6.0 Bar Chart

Accuracy Comparison The bar chart shows the overall accuracy of the three machine learning models—Logistic Regression, SVM, and Random Forest—used for heart disease prediction. It provides a clear comparison of how each model performs on the same dataset. By looking at the chart, we can easily identify which model gives the highest accuracy and how the others perform in comparison.

This visualization helps in quickly understanding the difference in performance between linear, kernel-based, and ensemble-based algorithms. From the chart, it is clear that Random Forest achieves the highest accuracy among the three models, followed by SVM, while Logistic Regression performs as a simple baseline model. This shows that ensemble methods like Random Forest are better at capturing complex patterns in the medical data.

Using a bar chart makes the comparison straightforward and easy to interpret. It summarizes the results visually and helps explain the performance of each model clearly in a research paper or presentation.

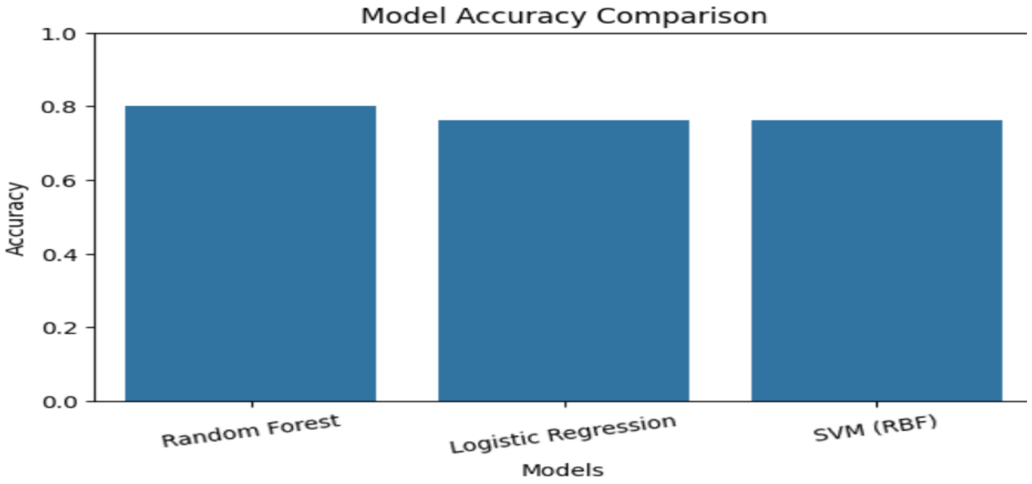

Fig 6.0: Model Accuracy Comparison

6.1 Confusion Matrices for each Model

6.1.0 Confusion Matrix – Logistic Regression

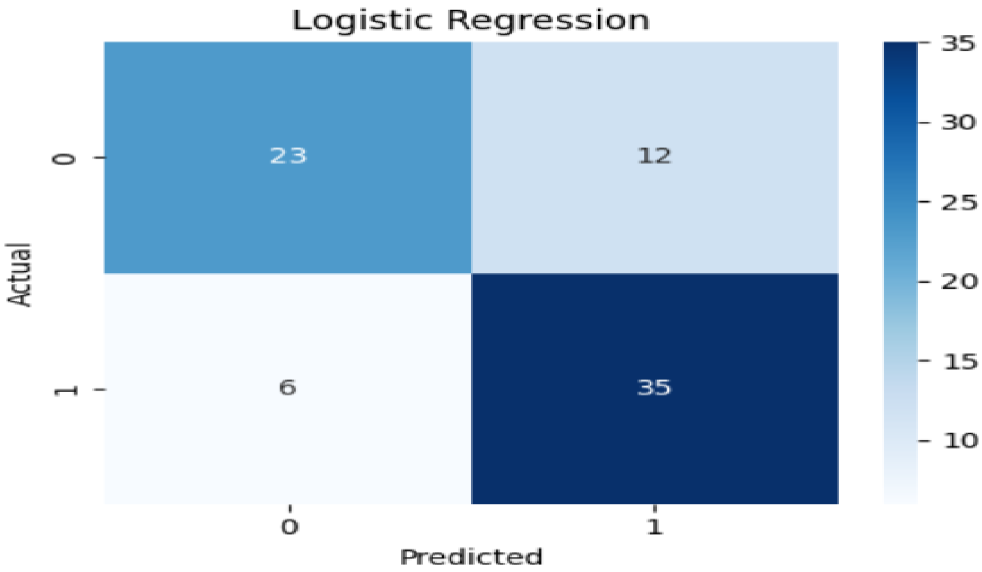

Fig 6.1: Confusion Matrix - Logistic Regression

6.1.1 Confusion Matrix – SVM(RBF)

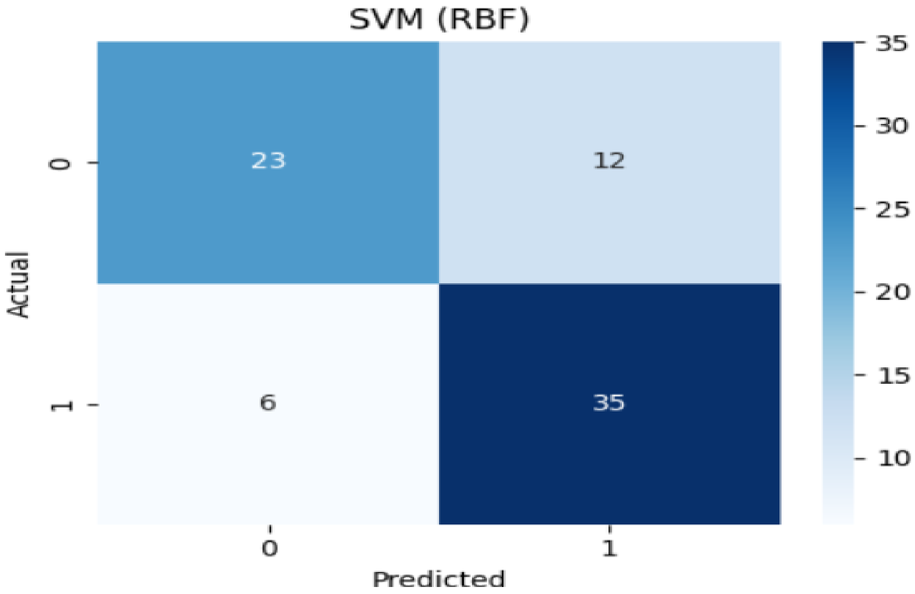

Fig 6.3: Confusion Matrix – SVM (RBF)

6.1.2 Confusion Matrix – Random Forest

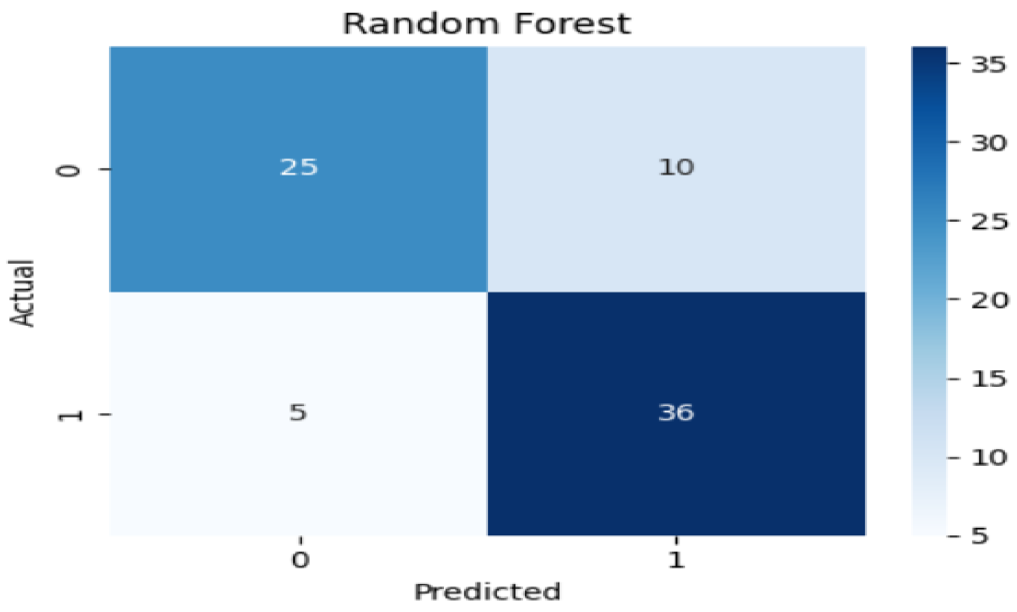

Fig 6.4: Confusion Matrix – Random Forest

7. Future Scope

There are several ways to improve this heart disease prediction system in the future. One of the most effective steps would be to use a larger and more diverse dataset. Adding data from different age groups, regions, and medical backgrounds can help the model learn better and make more accurate predictions.

Additional medical features such as family history, lifestyle habits, stress levels, and results from advanced diagnostic tests can also improve the model’s performance. These extra factors can give the model a clearer picture of the patient’s overall health.

In future work, more advanced machine learning models like Gradient Boosting, XGBoost, or even deep learning techniques can be explored. These models often perform better with larger datasets and can capture more complex patterns.

Another important direction is to focus on model explainability. Techniques that show which features influenced the prediction can help doctors trust the system more and make better decisions. Finally, this project can be developed into a simple web or mobile application so that users can enter their medical data and instantly check their heart disease risk.

In addition, integrating this prediction model with wearable health devices (such as fitness bands or smartwatches) could make real-time monitoring possible. This would allow users to track changes in heart-related parameters and get early warnings if their risk increases. With proper security measures, such a system can also store patient history and help doctors make better long-term decisions.

8. Conclusion

In this research, we aimed to predict heart disease using three different machine learning models: Logistic Regression, SVM, and Random Forest. The main goal was to understand how each model performs on the same dataset and identify which one gives the most accurate and reliable results. After completing the training and evaluation, we observed that each model has its own strengths, but Random Forest clearly performed the best among the three. This is because Random Forest is able to capture complex patterns in the data by combining multiple decision trees. SVM also showed very good performance, especially in handling non-linear relationships. Logistic Regression, although simpler, still provided meaningful results and served as a good baseline for comparison.

The results of this study show that machine learning can play a very important role in the early detection of heart disease. These models can analyze patient data faster than manual methods and can help doctors by providing additional insights. Even though these systems cannot replace medical experts, they can support them in making better and more informed decisions. Early prediction can help reduce the risk of severe heart conditions and improve the overall quality of healthcare.

This work can be extended in the future by using larger datasets, adding more medical features, or trying advanced models like XGBoost or deep learning. Improving explainability can also help doctors understand why the model predicts a certain result. With continuous improvements, these prediction systems can become even more helpful in real-world clinical environments and contribute to better patient care.

Overall, this research demonstrates how data-driven approaches can support healthcare systems and make early diagnosis more accessible. As more patient data becomes available in the future, such machine learning models can be further refined to achieve even better accuracy and reliability. With responsible use and continuous development, these predictive systems have the potential to reduce the burden of heart disease and assist doctors in delivering timely and effective treatment.

References

1. Dinesh D. Shah, Harshal A. Patel, K. K. Patel , “Heart disease prediction using machine learning techniques”, Springer (SN Computer Science journal) Heart Disease Prediction using Machine Learning Techniques | SN Computer Science
2. 2 Baban.U. Rindhe, Nikita Ahire, Rupali Patil, Shweta Gagare, Manisha Darade ,“Prediction of Heart Disease Using Machine Learning Algorithms”, International Journal of Advanced Engineering, Management and Science (IJAEMS) : 11_Prediction_of_Heart_Disease_Using_Machine_Learning_Algorithms-libre.pdf
3. 3 Diaa s AbdElminaam, Mostafa Radwan2, Nada Mohamed Abdelrahman, Hady Wael Kamal4, Abdelrahman Khaled Abdelmonem Elewa4, Adham Moataz Mohamed4, “ ML Heart Dis Prediction: Heart Disease Prediction using”, Journal of Computing and Communication : MLHeartDisPrediction: Heart Disease Prediction using Machine Learning

4. 4. Sonam Nikhar , A.M. Karandikar, " Heart Disease Prediction Using Machine Learning", International Journal of Advanced Engineering, Management and Science (IJAEMS) : HeartDiseasePredictionUsingMachineLearni ng.pdf
5. Praveen Kumar Reddy M, T Sunil Kumar Reddy, S.Balakrishnan, Syed Muzamil Basha, Ravi Kumar Poluru ,"Heart Disease Prediction Using Machine Learning Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-10, August 2019: J93400881019-libre.pdf
6. Heart Disease Prediction Using Machine Learning Algorithm: file:///C:/Users/MB511WS/Downloads/Heart_Disease_Prediction_using_Machine_L%20(3).pdf
7. Pooja Anbuselvan," Heart Disease Prediction Using Machine Learning Algorithm", International Journal of Engineering Research & Technology (IJERT): Heart disease prediction using machine learning techniques - IOPscience
8. Mr.Valle Harsha Vardhan, Mr.Uppala Rajesh Kumar, Ms.Vanumu ardhini, Ms. S abbileela V aralakshmi, Mr.A.Suraj Kumar, "Heart Disease Prediction Using Machine Learning Algorithm", International Journal of Engineering Research & Technology (IJERT): 2023-HEART DISEASE PREDICTION USING MACHINE LEARNING (1).pdf

10. APPENDIX

CODE