International Journal of Innovative Research in Engineering
Volume 7, Issue 2 (March - April 2026), PP: 336-340.
https://www.doi.org/10.59256/ijire.20260702041
www.theijire.com
ISSN No: 2582-8746

Heart Disease Prediction Using Machine Learning Techniques

Anamika Kumari [1], Adnan Mahmood[2]
[1,2] Department of Computer Science & Engineering, BIT Mesra, Patna Campus, Bihar, India.

How to cite this paper:
Anamika Kumari[1], Adnan Mahmood[2] ,"Heart Disease Prediction Using Machine Learning Techniques", IJIRE-V7I2-336-340.

Abstract: *Heart disease is one of the leading causes of death worldwide, making early prediction very important. This project develops a machine learning-based system to predict heart disease using patient data such as age, blood pressure, and cholesterol levels. Various algorithms including Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest are applied and compared. The dataset is preprocessed using scaling techniques and divided into training and testing sets. Model performance is evaluated using accuracy, confusion matrix, and ROC curve. A majority voting method is also used to improve prediction reliability. The system allows real-time prediction using user input. The results show that machine learning can help in early detection and support doctors in making better decisions.*

Key Words: *Heart Disease Prediction, Machine Learning, Logistic Regression, SVM, Random Forest, Classification.*

I.INTRODUCTION

Heart disease is one of the most serious health problems worldwide and is a major cause of death in both developed and developing countries. It includes conditions such as coronary artery disease, heart attack, and other cardiovascular disorders. Many factors like unhealthy lifestyle, lack of physical activity, smoking, high blood pressure, and high cholesterol contribute to the risk of heart disease. Early detection and proper treatment can significantly reduce the chances of severe complications and improve survival rates.

With the advancement of technology, machine learning has become an important tool in the healthcare sector. Machine learning algorithms can analyze large amounts of medical data and identify patterns that are difficult for humans to detect. This makes it possible to predict diseases at an early stage and assist doctors in making better decisions. In this project, machine learning techniques are used to build a prediction system for heart disease using patient data

The system uses different algorithms such as Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest to classify whether a person is likely to have heart disease or not. The dataset is processed and cleaned to improve accuracy, and different evaluation metrics are used to compare model performance. Additionally, the system provides a final prediction using a majority voting technique.

Overall, this project aims to develop an efficient and reliable tool for heart disease prediction, which can support healthcare professionals and contribute to better patient care.

Material And Methods

Study Design

This project is based on a machine learning approach for predicting heart disease using supervised learning techniques. The main objective is to develop a predictive system that can classify whether a patient is at risk of heart disease based on medical attributes. Multiple classification algorithms are applied and compared to identify the most accurate and reliable model. The system also incorporates a majority voting technique to improve prediction stability and performance.

Dataset Description

The dataset used in this study is the Heart Disease dataset, which contains patient health records with various medical attributes. These attributes include age, sex, chest pain type, blood pressure, cholesterol level, fasting blood sugar, electrocardiographic results, maximum heart rate, and other relevant features.

The target variable indicates whether a patient has heart disease (1) or not (0). The dataset is structured and suitable for classification tasks, allowing machine learning models to learn patterns and relationships between features and the target outcome.

Data Preprocessing

Data preprocessing is an important step to improve model performance and accuracy. The following steps were

performed:

- **Data Cleaning:** The dataset was checked for missing or inconsistent values.
- **Feature Selection:** Irrelevant features (if any) were removed to improve efficiency.
- **Feature Scaling:** StandardScaler was applied to normalize the feature values.
- **Data Transformation:** All features were converted into numerical format suitable for model training.

Train-Test Split

The dataset was divided into training and testing sets using an 80:20 ratio.

- **80% data** was used for training the models
- **20% data** was used for testing

This ensures that the model is evaluated on unseen data for better generalization.

Model Training

The following machine learning models were implemented:

- **Logistic Regression:** A statistical model used for binary classification.
- **Support Vector Machine (SVM):** Effective for high-dimensional data classification.
- **K-Nearest Neighbors (KNN):** A distance-based classification algorithm.
- **Random Forest:** An ensemble learning method using multiple decision trees.

All models were trained on the processed dataset and tested using the test data.

Evaluation Metrics

The performance of the models was evaluated using the following metrics:

- **Accuracy:** Measures overall correctness of the model
- **Precision:** Measures correct positive predictions
- **Recall:** Measures how many actual positives are correctly identified
- **F1-Score:** Harmonic mean of precision and recall Additional evaluation techniques include:
- **Confusion Matrix:** To visualize prediction performance
- **ROC Curve and AUC:** To measure classification performance
- **Learning Curve:** To analyze model performance with training size

Prediction System

A prediction system was developed using the trained models. The system allows users to input patient details such as age, sex, blood pressure, and other medical parameters. Based on the input, the system predicts whether the patient is at risk of heart disease or not.

Additionally, a **majority voting technique** is used to combine predictions from multiple models, improving accuracy and reliability. The system also provides a **risk score** indicating the severity of the condition.

II.RESULT

The machine learning models were evaluated based on accuracy, precision, recall, and F1-score. A comparative analysis of all models was performed.

The results show that:

- **Random Forest** achieved high accuracy due to its ensemble nature and ability to handle complex data patterns.
- **Logistic Regression** provided stable and interpretable results.
- **SVM** performed well but required proper scaling of data.
- **KNN** showed moderate performance and was sensitive to feature scaling.

The use of **majority voting** improved overall prediction performance by combining outputs from all models.

The confusion matrix and ROC curves indicate that the models can effectively distinguish between patients with and without heart disease. Overall, the system demonstrates good predictive capability and can be useful for early detection and decision support in healthcare.

Confusion Matrix Analysis

A confusion matrix is a useful technique used to evaluate the performance of classification models by comparing actual and predicted values. It provides detailed information about True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). This helps in understanding how well the model is performing and where it is making errors.

Confusion matrices were generated for all the models used in this project to analyze their classification performance.

Logistic Regression

Logistic Regression showed good performance with balanced classification results. It correctly predicted most of the positive and negative cases, although a few misclassifications were observed.

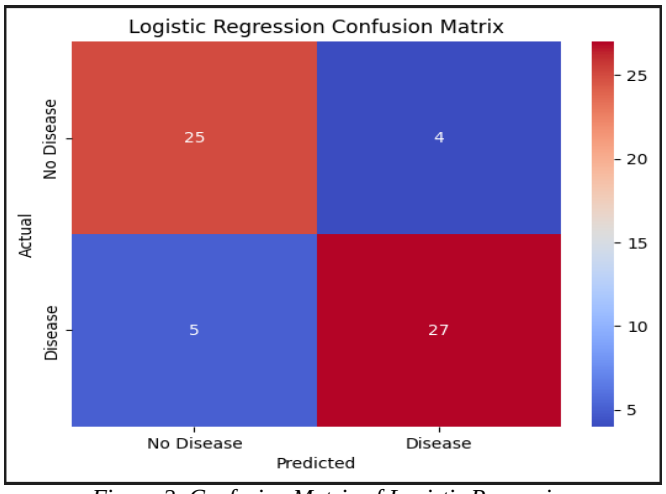

Figure 2: Confusion Matrix of Logistic Regression

Support Vector Machine (SVM)

SVM performed slightly better than Logistic Regression with fewer false predictions. It showed improved classification accuracy and better separation between classes.

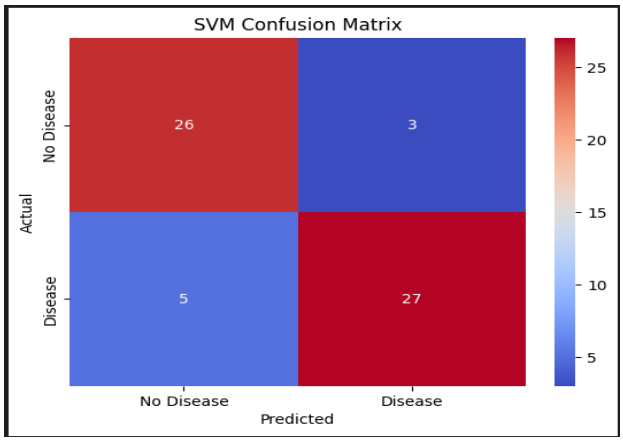

Figure 3: Confusion Matrix of SVM

K-Nearest Neighbors (KNN)

KNN achieved strong performance with fewer misclassifications. It showed high true positive and true negative values, indicating better prediction capability compared to other models.

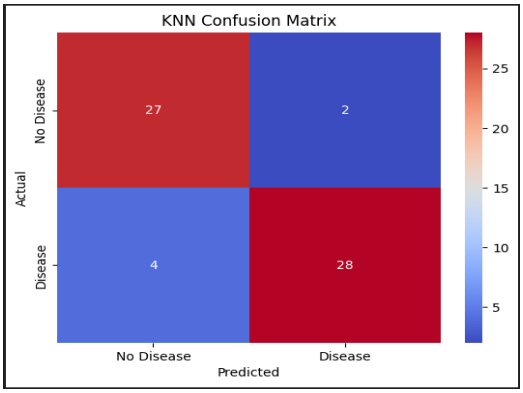

Figure 4: Confusion Matrix of KNN

Random Forest

Random Forest provided the best performance among all models. It had a well-balanced confusion matrix with high true positives and true negatives, and minimal false predictions. Its ensemble nature helped in improving accuracy and reducing errors.

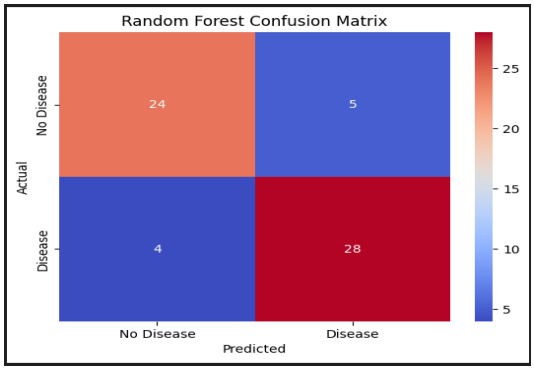

Figure 5: Confusion Matrix of Random Forest

ROC Curve Analysis

ROC (Receiver Operating Characteristic) curve is a tool to assess the performance of classification models by examining the trade-off between a True Positive Rate (TPR) and a False Positive Rate (FPR). It assists in knowing the extent to which the model is capable of differentiating between the various classes.

A model that lies nearer to the top-left corner is a pointer of better performance and a diagonal line is a representative of a random classifier.

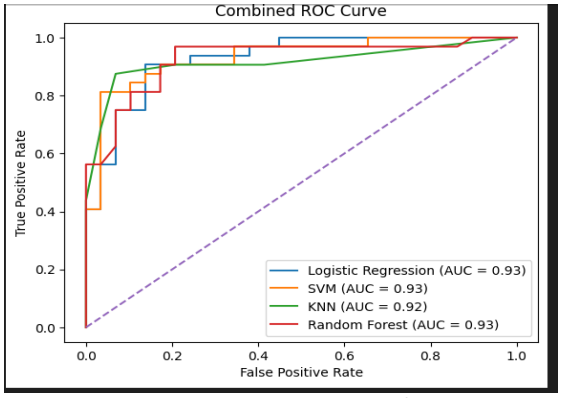

Figure 5: ROC Curve Comparison of All Models

Learning Curve Analysis

The learning curve is used to assess how a machine learning model's performance increases as training data volume increases. It aids in finding problems like underfitting and overfitting.

A successful model's ability to generalize to new data is demonstrated by a modest difference between training and validation scores.

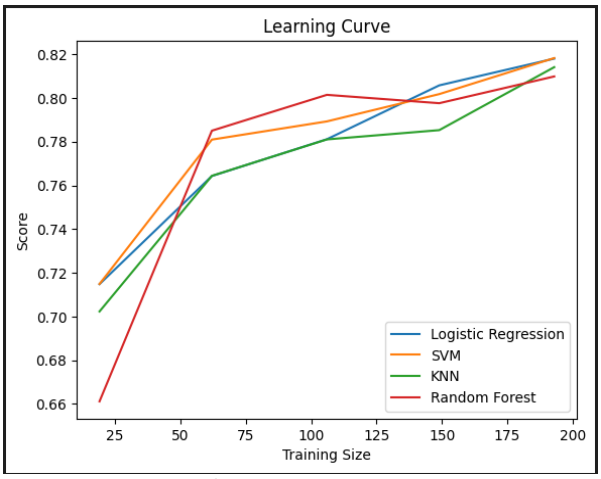

Figure 7: Learning Curve of XGBoost Model Based on the learning curve:

- The model effectively learns from the data, as evidenced by the excellent training score.

- As the training size grows, the validation score rises.
- There is not much of a difference between training and validation scores.

This suggests that there is little overfitting or underfitting and that the model is well-balanced. It also demonstrates how performance increases with additional data.

III. DISCUSSION

The results of this project show that machine learning techniques can effectively predict the risk of heart disease using patient data. Multiple algorithms such as Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest were implemented and compared to evaluate their performance. Each model has its own strengths; for example, Logistic Regression provides simple and interpretable results, while Random Forest offers higher accuracy by combining multiple decision trees.

The evaluation metrics, including accuracy, confusion matrix, and ROC curve, helped in understanding how well each model performs. It was observed that models with better generalization ability produced more reliable predictions on unseen data. The use of feature scaling improved the performance of distance-based algorithms like KNN and SVM.

The majority voting technique played an important role in improving prediction stability by combining outputs from different models. However, the project also has some limitations, such as dependency on dataset quality and size. More data and advanced models could further improve results. Overall, the system demonstrates the practical use of machine learning in healthcare for early detection and decision support.

IV. CONCLUSION

In this project, a machine learning-based system for heart disease prediction was successfully developed and analyzed. Different algorithms such as Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest were implemented to classify whether a person is at risk of heart disease. The dataset was carefully preprocessed using techniques like feature scaling and train-test splitting to improve model performance.

The results showed that all models were capable of predicting heart disease with good accuracy, but some models performed better than others depending on the data. A comparative analysis helped in identifying the most suitable algorithm. Additionally, the use of a majority voting technique improved the overall reliability of predictions. Evaluation metrics such as accuracy, confusion matrix, and ROC curve were used to measure model performance effectively.

The system also provides real-time prediction by taking user input, making it practical and user-friendly. Overall, this project demonstrates that machine learning can play a significant role in early detection of heart disease and can assist healthcare professionals in making faster and more accurate decisions, ultimately improving patient care and reducing risks.

References

1. Detrano, R., et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology.
2. Cleveland Heart Disease Dataset. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Heart+Disease
3. World Health Organization (WHO). (2023). Cardiovascular Diseases (CVDs). https://www.who.int
4. Kaur, H., & Kumari, V. (2017). Predictive modelling and analytics for diabetes using machine learning approach. Applied Computing and Informatics.
5. Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications.
6. Uddin, S., et al. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making.
7. Breiman, L. (2001). Random Forests. Machine Learning Journal.
8. Cortes, C., & Vapnik, V. (1995). Support Vector Networks. Machine Learning Journal.
9. Cover, T., & Hart, P. (1967). Nearest Neighbor Pattern Classification. IEEE Transactions.
10. Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. Wiley.
11. Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Elsevier.
12. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.
13. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.
14. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. New England Journal of Medicine.
15. Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases. International Journal of Advanced Computer Science.