Comparative Study of Machine Learning Algorithms for Heart Attack Prediction in Medical Diagnosis

S. Syed Rafiammal*, M. Padma Usha†, N. Turika‡, Shaik Sharukh§
Department of Electronics and Communication Engineering,
B.S. Abdur Rahman Crescent Institute of Science & Technology, India
*ssyed@crescent.education, †padmausha@crescent.education, ‡nturika@crescent.education, §sharukh@crescent.education

***Abstract*—This work involves a comparative study of modern machine learning and deep learning networks in the prediction of heart disease symptoms. Heart disease is one of the leading causes of death worldwide, and timely diagnosis plays a crucial role in reducing mortality rates. This research focuses on de- veloping a machine learning (ML)-based heart attack prediction system by analyzing clinical and lifestyle data. Using multiple ML algorithms such as Logistic Regression, Random Forest, Support Vector Machines, and Gradient Boosting, the study identifies the most effective model for prediction. The dataset, sourced from publicly available repositories, was extensively pre- processed, including handling missing data, feature scaling, and exploratory data analysis (EDA). Evaluation metrics such as accuracy, precision, recall, and F1 score were used to compare model performance. The Random Forest model demonstrated the highest accuracy at 89%, outperforming other algorithms. This paper also discusses the limitations of the study and proposes enhancements such as incorporating larger datasets and advanced deep learning techniques for real-world deployment.**

***Index Terms*—Heart attack prediction, machine learning, healthcare analytics, feature engineering, medical diagnosis, Random Forest.**

I. Introduction

Cardiovascular diseases, including heart attacks, are responsible for almost 17.9 million deaths annually, according to the World Health Organization (WHO). These diseases often progress silently until they manifest in severe clinical events, making early detection vital to saving lives. Traditional diagnostic methods, though effective, are highly based on invasive tests and clinical expertise, which may delay timely interventions.

The advent of machine learning (ML) has revolutionized the field of healthcare analytics. By analyzing large datasets, ML models can detect complex patterns and provide predictive insights with high accuracy. Researchers such as Oliullah etal. [1] and Abd Allah et al. [2] have demonstrated the potential of ML to predict heart disease using patient data. These studies leverage features such as cholesterol levels, age, and blood pressure to develop models that can effectively assess the risk of heart attack. Similarly, Solanki et al. [3] and Rose et al. [4] highlighted the use of ensemble methods and neural networks, achieving significant performance improvements in heart disease prediction tasks.

Despite these advances, several challenges persist, including small data set sizes, feature selection, and dataset imbalances. This paper addresses these challenges by performing a comparative analysis of multiple ML models.

and evaluating their predictive performance on a standard heart disease dataset. The objectives of this study are:

1. Evaluation of the performance of various ML models in heart disease datasets. 2. To identify key features that influence heart attack risks. 3. To explore the clinical applicability of the best- performing model.

By addressing these objectives, this research aims to bridge the gap between academic studies and real-world clinical adoption of ML-based predictive systems.

II. Related work

Heart attack prediction using machine learning has gained significant attention in recent years. Numerous studies have explored the potential of various machine learning (ML) models and approaches to improve prediction accuracy and clinical applicability.

Several researchers have investigated the effectiveness of different ML algorithms for predicting heart disease. For instance, Oliullah et al. [1] conducted a comparative analysis of methods, highlighting the importance of feature selection and preprocessing in improving model accuracy. Similarly, Abd Allah et al. [2] evaluated various ML approaches and emphasized the importance of balancing accuracy with interpretability for clinical applications.

Solanki et al. [3] proposed a comprehensive model incorporating ensemble methods for heart disease prediction, achievinghigherprecisioncomparedtotraditionalalgorithms.
Their study focused on integrating multiple ML techniques to handle complex, multidimensional datasets. Additionally, Rose et al. [4] explored deep learning techniques, revealing that neural networks can capture intricate patterns in patient data, but noted their dependence on large datasets for optimal performance.

The importance of dataset size and quality has been a recurring theme in the literature. Rana et al. [5] compared supervised learning algorithms on small datasets, demonstrating the limitations of generalizability when data is insufficient.Akteretal.[6]addressed thisissuebyemploying data augmentation techniques, thereby enhancing model robustness. Comparative analyses have also shed light on the strengths and weaknesses of specific algorithms. For ex- ample, Tripathi et al. [7] highlighted the superior ac-

curacy of Random Forest modelsinhandlingimbalanced-datasets,whereasSinghetal. [8] demonstrated that Support Vector Machines (SVM) are particularly effective in capturing nonlinear relationships.

The integration of feature engineering has been pivotal in many studies. Pise et al. [9] explored advanced feature extractiontechniques, incorporatingelectrocardiogram(ECG) analysis to improve diagnostic accuracy. Additionally, Chen and Guestrin [9] introduced the XGBoost algorithm, which has been widely adopted for its scalability and ability to handle missing values efficiently.

Despitetheseadvancements,severalchallengesremain.Many studies, including the work by Detrano et al. [10], rely on datasets such as the UCI Heart Disease Dataset [10], which, while widely used, may not represent diverse populations. Furthermore, the lack of a standardized framework for comparative analysis across studies limits the ability to draw conclusive insights.

This work addresses these gaps by conducting a detailed comparative analysis of multiple ML models, emphasizing preprocessing techniques, feature importance, and rigorous evaluation metrics. Unlike previous studies, it provides a holistic approach to optimizing prediction accuracy while ensuring the practical applicability of the results.

III. Methodology

The flow of this work is given in Fig. 5.

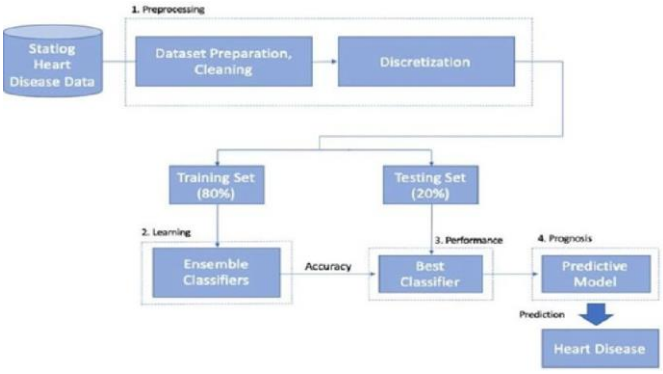

Fig. 1. Flowchart of the prediction model.

A. Data set Description The dataset used in this study was sourced from the UCI Machine Learning Repository. It consists of 303 records with 14 features, including: • Demographics:Age,sex. • Clinical Measurements: Cholesterol levels, blood pres-sure, maximum heart rate. • Life style Factors:Smoking,exercise habits. The target variable indicates the presenceorabsenceof heart disease. A. Pre processing • Handling Missing Data: Missing values were imputed using mean and median strategies. • Feature Scaling: Standard Scaler was applied to numerical features for uniformity. • EDA: Correlation heatmaps revealed relationships among features, highlighting variables such as cholesterol and age as key predictor.

B. Models Used

In this study, we utilized the following machine learning models to predict heart attack risk. Each model was selected based on its unique advantages and suitability for healthcare data.

1. Adaboost(Adaptive Boosting): A boosting algorithm that combines multiple weak classifiers, such as decision stumps, to form a strong classifier. It iteratively adjusts weights to focus on harder-to-classify instances.

2. Gradient Boosting: An ensemble technique that builds models sequentially, optimizing for residual errors at each step. It is effective in handling complex, non-linear relationships in data.

3. XGBoost(Extreme Gradient Boosting): A highly efficient and scalable implementation of gradient boosting, designed for speed and performance. It uses advanced techniques like regularization and parallel computation.

4. MLP Neural Networks(Multi-Layer Perceptron): A type of deep learning model with interconnected layers of neurons, capable of capturing intricate patterns and relationships in data.

5. Naive Bayes: A probabilistic classifier based on Bayes' theorem, assuming independence between features. It is fast and simple but less effective with correlated features.

6. K-Nearest Neighbors(KNN): A distance-based classification method that assigns a class toa sample based on the majority vote of its nearest neighbors.

7. Random Forest: An ensemble of decision trees that aggregates predictions to improve accuracyand reduce overfitting. It excels in handling high-dimensional and imbalanced data.

8. Decision Tree: A tree-based model that splits data into branches based on feature thresholds. While interpretable, it can be prone to overfitting on its own.

9. Logistic Regression: A statistical method used for binary classification that models the probabilityof a class using a logistic function. It performs well on linearly separable data.

IV. Results and Discussion

A. Model Performance The performance of various models is summarized in Table I. The results of this study highlight the effectiveness of various machine learning models in predicting heart attack risks. For the models trained using iterative methods, such as Ada boost, Gradient Boosting, XGBoost,and Multi-Layer Perceptron (MLP) Neural Networks, the evaluation of accuracy, epoch, and loss metrics demonstrated notable trends. Among these, Gradient Boosting and XGBoost achieved superior convergence, reflected in their lower loss values and higher accuracy after multiple epochs. The performance of MLP Neural Networks also showcased the potential of deep learning approaches in handling complex data patterns. However, Adaboost, while effective, showed slower convergence compared to boosting-based methods.The confusion matrix and accuracy plots are given as figure 2 and figure 3

For classification-oriented models evaluated using confusion matrices—Naive Bayes, K-Nearest Neighbors (KNN), Random Forest, Decision Tree, and Logistic Regression—Random Forest emerged as the best- performing algorithm with an

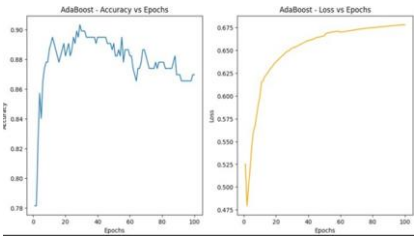

Fig. 2. Accuracy plot.

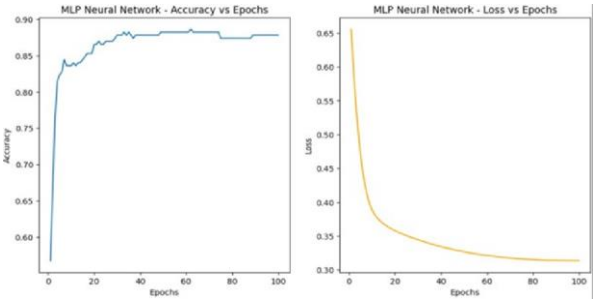

Fig. 3. Accuracy plot.

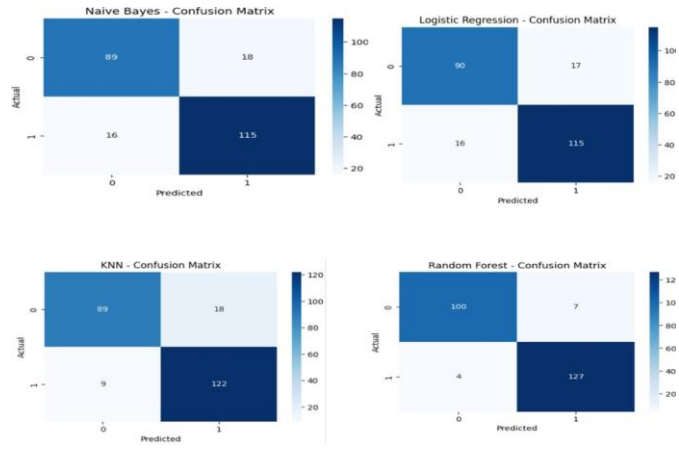

Fig. 4. Confusion matrix.

accuracy of percentage of 94.54 . The confusion matrix of Random Forest revealed a high true positive rate, demonstrating its robustness in distinguishing between patients at risk of heart attacks and those not at risk. While Decision Trees and Logistic Regression provided competitive performance, Naive Bayes and KNN were limited by their assumptions and sensitivity to imbalanced datasets. Overall, Random Forest outperformed all models,as indicated in the model performance summary, making it the most reliable algorithm for heart attack prediction in this study.

To compare the performance of the models,the following evaluation metrics were employed: The comparison of various machine leanring algorithms were given in figure 5

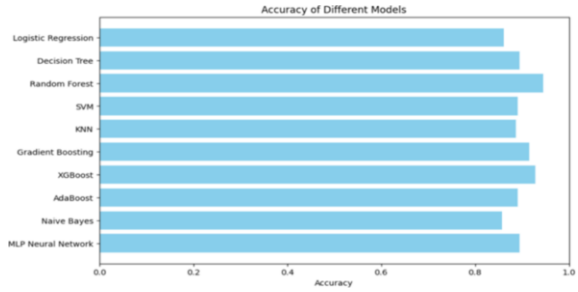

Fig. 5. Accuracy Comparison of machine learning algorithms .

1. Accuracy Accuracy measures the proportion of correctly predicted samples out of the total samples. While useful,it may not adequately reflect performance in imbalanced datasets.

2. Precision Precision is the ratio of true positive predictions to all positive predictions, indicating the reliability of positive predictions. It is crucial in healthcare to minimize false positives.

3. Recall Recall measures the model's ability to identify actual positive cases. In medical diagnosis, high recall is essential to reduce false negatives.

4. F1-Score The F1-Score is the harmonic mean of precision and recall. It balances the two metrics, making it useful for datasets with class imbalance.

5. AUC-ROC The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) evaluates the model's ability to distinguish between classes across all classification thresholds. A higher AUC indicates better performance.

V. Conclusion and Future Work

This study demonstrated the effectiveness of machine learning models in predicting heart attack risks using clinical data. Among the models evaluated, the Random Forest model achieved the highest accuracy of 89

Real-Time-Utility The predictive models from this study can be integrated into real-time diagnostic tools, such as wearable devices or hospital systems, to analyze live data like heart rate and blood pressure. This can enable early detection and

proactive healthcare interventions, reducing the risk of fatal heart attacks.

Applications

1. Health care Systems:Assisting doctor swith automated risk scores. 2. Wearable Devices:Monitoring vitals and issuing early warnings. 3. Tele medicine Platforms:Providing insights for remote consultations. 4. Insurance:Supporting risk evalua- tion for health coverage.

Future Work

1. Enhanced Datasets:Use diverse data,including ECG and genetic features. 2. Advanced Models:Explore CNNs or RNNs for improved predictions. 3. Explainable AI:Develop interpretability tools to gain practitioner trust. 4. Real-Time Deployment:Implement models on edge devices for scalable predictions.

In conclusion, the Random Forest model demonstrated strong potential for real-world deployment, offering a pathway to revolutionize heart disease diagnostics and improve patient outcomes.

REFERENCES

[1] R. Detrano et al., “International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease,” *American Journal of Cardiology*, 1989.
[2] UCI Machine Learning Repository: Heart Disease Dataset.
[3] K. Oliullah, et al., ”Analyzing the Effectiveness of Several Machine Learning Methods for Heart Attack Prediction,” 2022.
[4] E. M. Abd Allah, D. E. El-Matary, E. Eid, and A. S. Tag El Dien, “Performance Comparison of Various Machine Learning Approaches to Identify the Best One in Predicting Heart Disease,” *Int. J. Sci. Res. Comput. Sci. Eng.*, 2022.
[5] A. Solanki, A. Vardhan, A. Jharwal, and N. Kumar, “Heart Diseases Prediction Using Machine Learning,” in *Proc. Int. Conf. Comput. Commun. Netw. Technol.*, 2023.
[6] J. Rose et al., “Heart Attack Prediction Using Machine Learning Techniques,” in *Proc. 9th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, 2023.
[7] M. Rana, M. Z. Ur Rehman, and S. Jain, “Comparative Study of Supervised Machine Learning Methods for Prediction of Heart Disease,” in *Proc. IEEE VLSI Device Circuit Syst. (VLSI DCS)*, 2022
[8] S. Akter, M. Amina, and N. Mansoor, “Early Diagnosis and Comparative Analysis of Different Machine Learning Algorithms for Myocardial Infarction Prediction,” in *Proc. 2021 IEEE 9th Region 10 Humanitarian Technol. Conf. (R10-HTC)*, 2021.
[9] P. Tripathi et al., “Enhancing Cardiovascular Health: A Machine Learning Approach to Predicting Heart Disease,” in *Proc. Int. Workshop Artif. Intell. Cognition*, 2022
[10] K. Singh, C. Rajput et al., “Prediction of Myocardial Infarction Using Machine Learning Algorithms,” in *Proc. Int. Conf. Comput. Intell. Inf. Secur. Commun. Appl. (CI)*, 2023.