World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Comparative analysis of machine learning algorithms for ECG-based heart attack prediction: A study using Bangladeshi patient data

Md Alif Sheakh [1], Mst. Sazia Tahosin [1], Lima Akter [2,*], Israt Jahan [3], Md Nakibul Islam [3], Md Rafiuddin Siddiky [4], Md Mahadi Hasan [3] and Sakibul Hasan [5]

[1] *Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh.*
[2] *Department of Computer Science and Engineering, Atish Dipankar University of Science and Technology, Dhaka, Bangladesh.*
[3] *Department of Information Technology, Washington University of Science and Technology, Virginia, USA.*
[4] *Department of Information Systems Technology, Wilmington University, Delaware, USA.*
[5] *Department of Civil Engineering, Chongqing University of Science and Technology, China.*

World Journal of Advanced Research and Reviews, 2024, 23(03), 2572–2584

Publication history: Received on 18 August 2024; revised on 23 September 2024; accepted on 26 September 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.23.3.2928

Abstract

This study aims to identify the most accurate machine learning algorithm for predicting heart attacks using demographic data, physiological measurements, and electrocardiogram (ECG) results. We utilized a dataset of 4,000 patient records, combining data from DMCH and Kaggle. Our methodology involved comprehensive data preprocessing, including ECG noise removal and feature selection using the Brouta algorithm. We implemented and compared six machine learning algorithms: Decision Tree, Random Forest, Logistic Regression, Support Vector Machine, XGBoost, and K-Nearest Neighbors. The results demonstrate that our proposed method can accurately predict heart attacks with high sensitivity and specificity. Among the tested algorithms, Random Forest achieved the highest accuracy of 87%, with well-balanced precision (0.86), recall (0.85), and F1-score (0.87). K-Nearest Neighbors and XGBoost also showed strong performance, with accuracies of 81% and 80% respectively. This study contributes to the field by utilizing a large, diverse dataset and providing a comprehensive comparison of multiple algorithms. Our findings suggest the potential for integrating machine learning, particularly Random Forest models, into clinical practice for early heart attack risk assessment, representing a significant step towards improving cardiovascular care through advanced data analysis techniques.

Keywords: Heart attack; Cardiovascular; Heart disease prediction; Random forest; Electrocardiogram prediction

1. Introduction

Cardiovascular diseases, particularly heart attacks, remain a leading cause of mortality and morbidity worldwide, posing a significant challenge to global health systems. In 2023, cardiovascular diseases were responsible for an estimated 18 million deaths globally, accounting for 35% of all deaths [1]. The impact is particularly pronounced in developing countries, where healthcare resources are often limited. In Bangladesh, for instance, the World Health Organization estimates that coronary heart disease accounted for 16% of all deaths in 2022, underscoring the urgent need for effective preventive strategies. The ability to predict and prevent heart attacks not only saves lives but also significantly reduces the economic burden on healthcare systems and improves overall quality of life for at-risk individuals [2].

* Corresponding author: Lima Akter

In recent years, the integration of machine learning techniques in healthcare has opened new avenues for early detection and prevention of heart attacks. These advanced computational methods offer the potential to analyze complex, multidimensional medical data with unprecedented accuracy and efficiency. By leveraging demographic information, physiological measurements, and electrocardiogram (ECG) results, machine learning algorithms can identify subtle patterns and risk factors that might elude traditional clinical assessments [3]. This approach holds promise for developing more accurate and personalized risk prediction models, enabling timely interventions and targeted preventive measures [4].

Heart attacks, medically termed myocardial infarctions, are serious cardiovascular events that occur when blood supply to a portion of the heart muscle is suddenly cut off. This interruption in blood flow is typically caused by a blockage in one or more of the coronary arteries, often due to the accumulation of fatty deposits called plaque. If not addressed promptly, this lack of oxygen-rich blood can result in significant damage or even death of heart muscle cells. [5]. However, the complex interplay of these factors and their relative importance in different populations is not fully understood, highlighting the need for more sophisticated analytical approaches. Machine learning algorithms have shown promising results in various medical fields, including cardiology. These algorithms can be broadly categorized into supervised learning (e.g., logistic regression, decision trees, random forests), unsupervised learning (e.g., clustering algorithms), and deep learning methods [6]. Each category has its strengths and limitations in handling different types of data and prediction tasks. Previous studies have explored the application of individual machine learning algorithms for heart attack prediction, but comprehensive comparisons across multiple algorithms using a diverse dataset are limited, especially in the context of developing countries.

This research aims to address this gap by conducting a comparative analysis of six machine learning algorithms for heart attack prediction: Decision Tree, Random Forest, Logistic Regression, Support Vector Machine, XGBoost, and K-Nearest Neighbors. Our study utilizes a unique dataset combining records from Dhaka Medical College Hospital and open-source repositories, providing a rich and diverse pool of patient information. This approach not only allows for a robust evaluation of algorithm performance but also contributes to the understanding of heart attack risk factors in the Bangladeshi population, an area that has been underrepresented in global cardiovascular research. The key contributions of this research are given below:

- The use of a large patient record dataset to train and evaluate the predictive models.
- The combination of demographic data and physiological data, such as electrocardiogram (ECG) results, improves the accuracy of the predictions.
- The comparison of traditional machine learning algorithms to determine the best baseline model for heart attack prediction.
- The demonstration of the high accuracy of the proposed method in predicting heart attacks, which suggests its potential for clinical use.

Other existing works on heart attack prediction have mainly focused on using traditional statistical methods [7] and demographic data. However, these methods may not fully capture the complexity of the physiological data associated with heart attacks. The proposed approach in this paper utilizes machine learning algorithms and combines demographic data with physiological data results to improve the accuracy of the predictions. This approach allows for a more comprehensive data analysis that may be absent from traditional methods. Additionally, the proposed approach compares machine learning algorithms to determine the best model for heart attack prediction. This allows for a more robust evaluation of the performance of the predictive models. This approach has been tested using a large dataset of patient records, which increases the generalizability of the findings. It can overcome the limitations of traditional methods and may have the potential to be integrated into clinical practice as a tool for early heart attack detection and prevention.

2. Literature review

Several research studies have explored various machine learning techniques for the early detection and prediction of heart disease. These techniques have proven instrumental in improving the accuracy and efficiency of diagnosing heart conditions. For instance, Pande et al. (2024) presented a heart disease prediction model using Random Forest, Gradient Boosting, and Deep Neural Networks (DNNs) with high accuracy, emphasizing ensemble learning methods to enhance predictions [8]. Their approach incorporated advanced data preparation and feature selection techniques, significantly improving prediction performance. Similarly, Golande and Kumar (2022) explored a combination of machine learning methods such as Naive Bayes and Support Vector Machines (SVM), achieving remarkable success in predicting heart diseases with high precision [9]. The application of Naive Bayes was also notable in the work of Shubham Mall (2024),

who demonstrated that it offered superior accuracy in comparison to other models like K-Nearest Neighbors (KNN) and Decision Trees [2].

Additionally, Ramalingam et al. (2018) utilized Principal Component Analysis (PCA) alongside Decision Trees to improve feature selection, which further enhanced the model's accuracy in detecting heart disease [10]. Another prominent study by Abubakar et al. (2024) demonstrated the use of Logistic Regression and Naive Bayes for heart disease classification, showing that the latter provided superior results with better accuracy and computational efficiency [11]. Despite these advancements, certain challenges persist, such as handling imbalanced datasets and interpreting complex models in clinical settings. Nevertheless, machine learning continues to offer robust solutions for early detection, enabling timely intervention and improved patient outcomes.

Table 1 Summarization of existing work

Ref	Work	Algorithms	Best Accuracy
[12]	Developed a prediction model by integrating advanced data preprocessing, feature selection, and ensemble learning for heart disease detection.	Decision Tree, Random Forest, Support Vector Machine, Gaussian Naive Baye	99.12%
[2]	Conducted a detailed comparison of several machine learning methods to identify the most effective approach for predicting heart disease.	Naive Bayes, Decision Tree	98.33%
[8]	Used a transformer-based self-attention mechanism to model complex patterns for accurate prediction of cardiovascular diseases.	Transformer Model	94.3%
[13]	Focused on improving heart disease detection through the use of hybrid classifiers and advanced feature selection techniques.	SVM, Naive Bayes, Random Forest	72.37%
[11]	A comprehensive exploration of various machine learning models, highlighting the best methods for classifying heart disease risks.	Random Forest, K-Nearest Neighbors	93.33%

3. Methodology

Our research methodology for heart attack prediction employs a comprehensive approach encompassing data collection, feature selection, model training, and evaluation. We begin by gathering patient information from various sources, continuously assessing data sufficiency to ensure a robust dataset. Once sufficient data is acquired, we move on to feature selection, utilizing the Brouta algorithm to identify the most relevant attributes. This step helps reduce dimensionality and improve model performance by focusing on the most informative features for our prediction task.

Following feature selection, we apply various machine learning algorithms to the prepared dataset, including LR[14], KNN [15], DT [16], SVM [17], RF [18], and XGB [19]. Each algorithm is used to create a predictive model. We then perform hyperparameter tuning to optimize the performance of each model, adjusting the parameters to find the configuration that yields the best results. The models are trained using the optimized parameters, and we implement an iterative process to refine their performance. If any errors or issues are detected during training, we return to the hyperparameter tuning stage for further optimization. This cycle continues until satisfactory performance is achieved across all models. Once the models are successfully trained, we evaluate their performance using a separate test dataset. We collect various metrics such as accuracy, precision, recall, and F1-score to comprehensively assess each model's effectiveness in predicting heart attacks. The final stage involves a thorough performance analysis of all the models, comparing their results across different metrics to identify the best-performing model for heart attack prediction. This systematic methodology ensures a rigorous and comprehensive evaluation of machine learning algorithms for cardiovascular risk assessment. By following this approach, we aim to develop more accurate and reliable tools for predicting heart attacks, potentially improving early detection and intervention strategies in clinical practice.

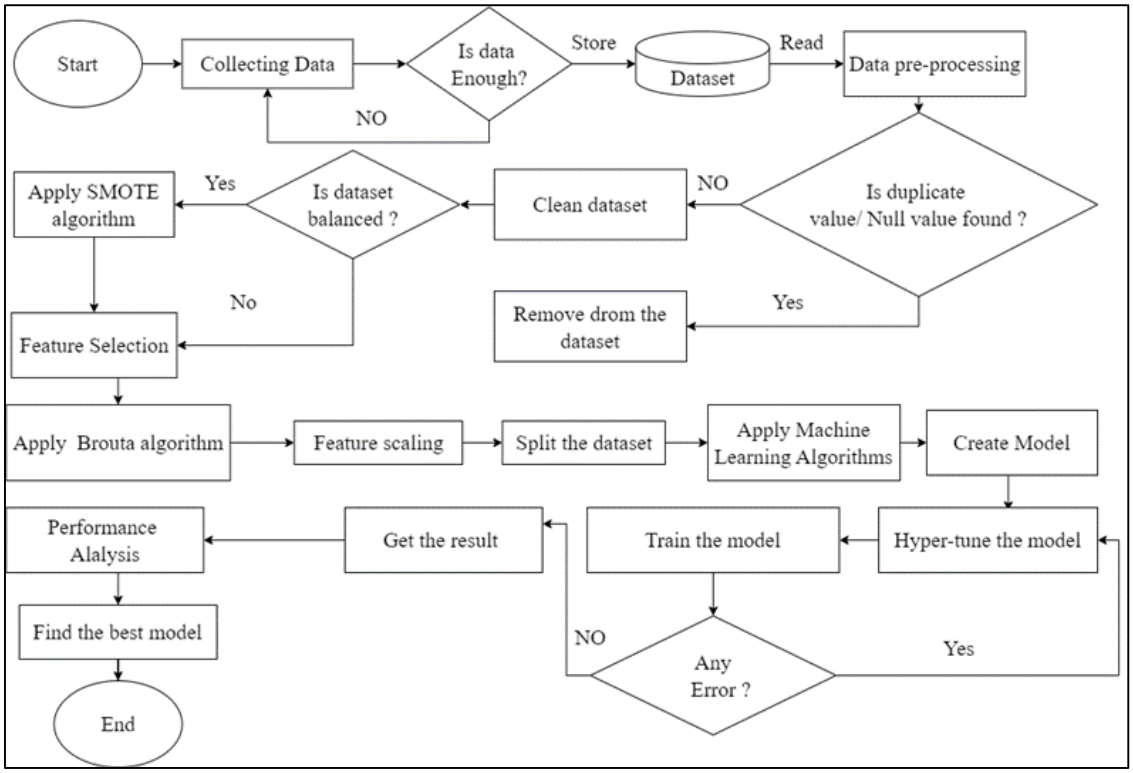

Figure 1 Proposed model flowchart with the working procedure

3.1. Data Description

This study utilizes a comprehensive dataset comprising 4,000 patient records, meticulously compiled from two primary sources. The majority of the data, consisting of 2,478 records, was manually collected from Dhaka Medical College Hospital, ensuring relevance to the local population. The remaining 1,522 records were sourced from open-source dataset repositories, particularly Kaggle, to enhance the diversity and robustness of our analysis. The dataset encompasses 15 distinct features, capturing a wide range of patient characteristics and medical indicators. Key demographic features include age and gender, while behavioral factors such as smoking status and average cigarette consumption per day are also recorded. Critical health indicators in the dataset comprise blood pressure medication usage, history of stroke, prevalence of hypertension, and diabetes status. Physiological measurements include total cholesterol levels, Body Mass Index (BMI), systolic and diastolic blood pressure, heart rate, and blood glucose levels. Notably, the dataset includes a target variable, 'TenYearCHD', which indicates the individual's 10-year risk of developing coronary heart disease. This comprehensive array of features allows for a nuanced analysis of factors contributing to heart attack risk, enabling our machine learning models to capture complex interactions and patterns within the data.

Table 2 shows the feature description of our dataset. It includes features ‘age’, representing the age of an individual in the format of year whose data is collected. The ‘male’ feature indicates gender, categorized as Male or Female. The ‘currentSmoker’ feature denotes that the person is taking cigarettes currently or not. The ‘cigsPerDay’ feature refers to the number of cigarettes is taking by the person. The ‘BPMeds’ feature represents that the person is taking blood pressure medication or not. The ‘prevalentStroke’ feature indicates that the person has suffered by the stroke in the past or not. The ‘prevalentHyp’ feature represents that the person has hypertension or not. The ‘diabetes’ feature represents the person has diabetes or not. The ‘totChol’ feature denotes the level of cholesterol. The ‘BMI’ feature measures Body Mass Index. The ‘sysBP’ feature signifies the level of systolic blood pressure. The ‘diaBP’ feature captures the value of diastolic blood pressure. The ‘heartRate’ feature indicates the value of heart rate. The ‘gulocose’ feature represents the value of gulocose which is present in the blood. Lastly, the ‘TenYearCHD’ represents thepossibility of having heart disease in the next 10 year.

Table 2 Feature description of the dataset

Features Name	Description
age	Age of a person in the format of the year
male	Whether the person is a male or not (binary format)
currentSmoker	Whether the person is a chain smoker or not (binary format)
cigsPerDay	Cigarettes are taken by the person per day
BPMeds	Whether the person is taking blood pressure medication or not (binary format)
prevalentStroke	Whether the person has an experience of stroke (binary format)
prevalentHyp	Whether an individual is suffering from hypertension or not (binary format)
diabetes	An individual has diabetes or not (binary format)
totChol	The level of cholesterol
BMI	The value of Body Mass Index
sysBP	The value of systolic blood pressure
diaBP	An individual's diastolic blood pressure
heartRate	The value of the heart rate in the continuous format
gulocose	The level of blood glucose of a person
TenYearCHD	The individual's 10-year risk of getting coronary heart disease

Figure 2. shows the ratio graph between "has disease" and "no disease" on different dimensions based on the collected dataset. This graph shows the ratio between male and female people, smokers and non-smokers person; people who have diabetic and non-diabetic people. This graph also shows the ratio between people who are under medication and who are not on medication. Finally, this graph shows the ratio between hypertensive and normotensive persons.

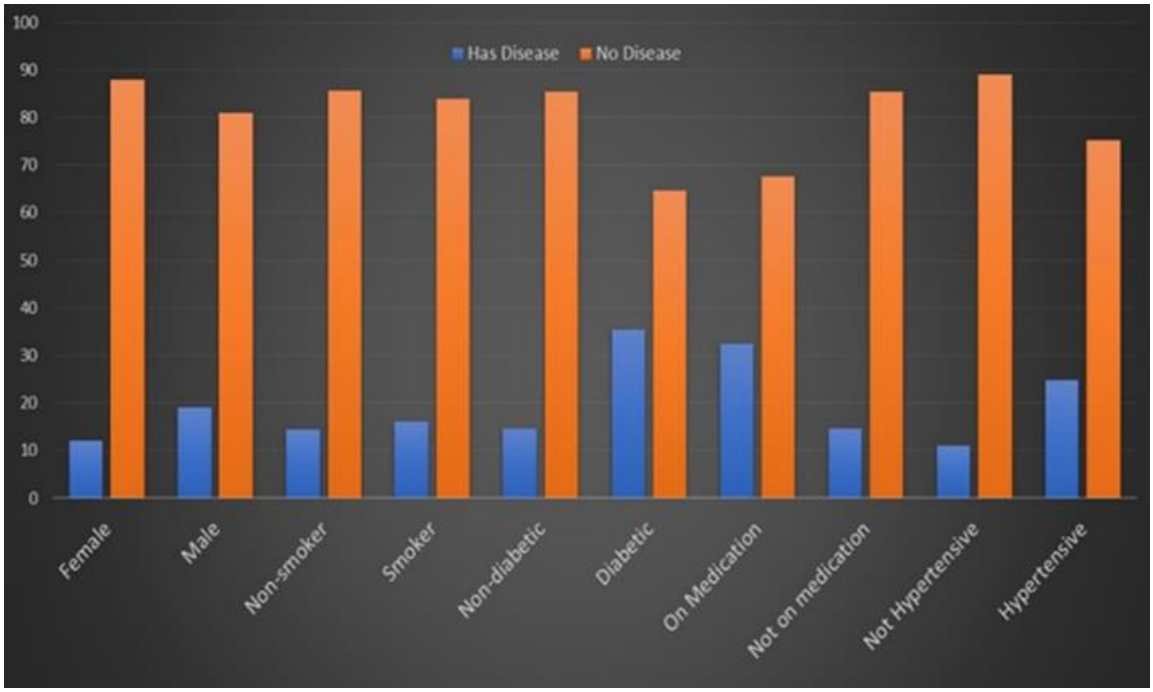

Figure 2 Data ratio based on people's current condition

3.2. Data Pre-Processing

Data preprocessing is a critical step in our methodology, ensuring the quality and reliability of our analysis. We begin by addressing missing values and duplicates in the dataset. Null values are removed using the dropna() function, maintaining data integrity and enhancing overall analysis quality. To handle duplicate records, which can skew sample weights and increase training time, we employ the drop_duplicates() function from pandas [20]. This efficiently identifies and removes duplicate entries, ensuring each record is unique and contributes meaningfully to the analysis. Following this initial cleaning, we perform noise removal on the ECG signals, a crucial step for improving the accuracy of our heart attack prediction models. Fig. 3. shows the steps followed during this data pre-processing.

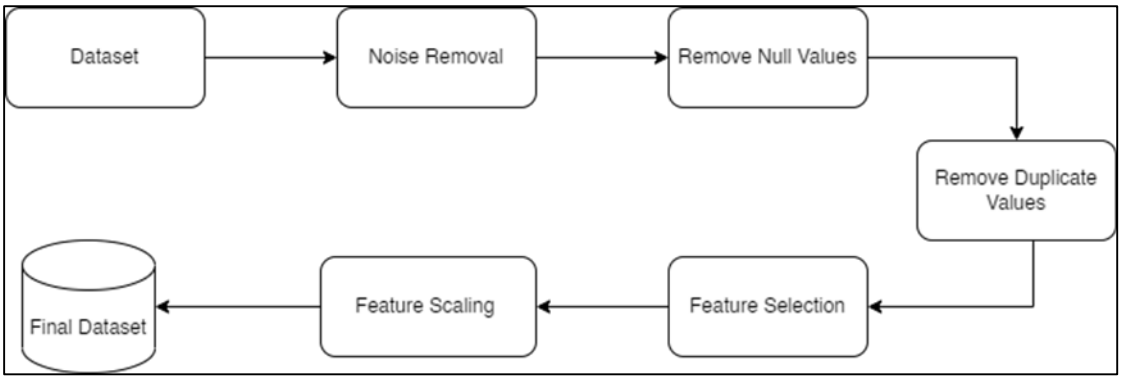

Figure 3 Data pre-processing steps of this work

3.3. Noise Removal

In the pre-processing step of ECG signals, noise removal is a critical aspect that significantly impacts the performance of machine learning models. Electrocardiograms (ECGs) were collected and utilized in this study. To combat baseline wander, we employed high-pass filtering methods, such as using a moving average or a high-pass Butterworth filter. Secondly, powerline interference was eliminated by adopting notch filters to suppress the frequencies associated with the power grid. To counter electromyographic (EMG) noise, we utilized adaptive filters to identify and attenuate muscle-related interference. Electrode motion artifacts were mitigated using wavelet denoising or template subtraction methods. These comprehensive noise removal strategies effectively enhanced the quality of ECG signals, enabling our machine-learning models to focus on extracting relevant diagnostic features for improved accuracy and performance.

3.3.1. Removing Null and Duplicate Values

Data cleaning is a crucial step in the data wrangling process. This work focuses on eliminating null values and duplicate records from the dataset to improve performance, accuracy, and reduce bias. Null values are removed using the dropna() function, which helps maintain data integrity and enhances the overall quality of the analysis. To address the issue of duplicate records, which can skew sample weights and increase training time, the drop_duplicates() function from pandas is employed. This function efficiently identifies and removes duplicate entries from the dataset, ensuring that each record is unique and contributes meaningfully to the analysis. By implementing these cleaning techniques, the dataset is refined and optimized for subsequent processing and modeling tasks.

3.3.2. Feature Selection

Feature selection is a crucial technique in data preprocessing that aims to identify and retain the most relevant attributes within a dataset while removing irrelevant ones. This process is instrumental in constructing a more efficient and accurate predictive model. By carefully selecting the best features, we can address several key challenges in machine learning. Feature selection helps mitigate the risk of overfitting, where a model becomes too closely tailored to the training data and fails to generalize well to new, unseen data. Additionally, it significantly reduces training time by focusing computational resources on the most informative attributes. Perhaps most importantly, effective feature selection can maintain or even improve model accuracy by eliminating noise and focusing on the most predictive variables. In this research, our feature selection procedure was specifically designed to optimize the dataset, streamlining it to include only the most impactful attributes. This approach allowed us to construct an improved predictive model that balances efficiency and accuracy.

3.3.3. Brouta Algorithm

In our study, we employed the Brouta Algorithm for feature selection, a crucial step in optimizing our heart attack prediction model. This algorithm, known for its effectiveness in both regression and classification problems, operates by minimizing an objective function [21]

$$f(x) = \sum_{i=1}^{n} w_i \cdot (\hat{y}_i - y_i)^2 + \lambda \cdot \|\theta\|^2 \dots\dots\dots\dots (1)$$

where $\hat{y}_i$ and y_i are predicted and actual values respectively, w_i are weights, and λ is a regularization parameter. The algorithm iteratively updates model parameters θ using the gradient descent rule:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_\theta f(x) \dots\dots\dots\dots. (2)$$

where η is the learning rate and $\nabla_\theta f(x)$ is the gradient. This process continues until ||∇θf(x)||<ϵ, indicating convergence. The Brouta Algorithm's ability to consider multivariable relationships was crucial in exploring feature interdependencies in our dataset. It offered an enhanced version of random forest's variable selection technique, improving reliability and performance. By applying this algorithm, we identified the most significant features related to our outcome variable, mitigating overfitting risks and reducing model training time. This approach ensured our model maintained high accuracy while retaining only the most relevant attributes, resulting in more reliable predictions for heart attack risk. The Brouta Algorithm's sophisticated mathematical foundation, combined with its practical benefits in feature selection, significantly contributed to the efficiency and accuracy of our analysis, providing robust insights for our research in cardiovascular health prediction.

Algorithm 1: Pseudocode of the Brouta Algorithm

```
function BroutaAlgorithm(data, outcome_variable):
 selected_features ← ∅
 current_score ← 0
 while not stopping_criteria_met():
 for each feature in data.features:
 if feature ∉ selected_features:
 candidate_features ← selected_features ∪ {feature}
 model ← train_model(data[candidate_features], outcome_variable)
 score ← evaluate_model(model, data[candidate_features], outcome_variable)
 if score > current_score:
  selected_features ← candidate_features
  current_score ← score
 return selected_features
```

3.3.4. Data Balancing

A fundamental principle in machine learning is the importance of working with balanced datasets, or those as close to balance as possible. This approach ensures that each class receives equal consideration during model training. When machine learning algorithms are applied to imbalanced datasets, there's a significant risk of the minority class being overlooked or underrepresented in the model's predictions. To address this issue, this research employs the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is an effective method for balancing datasets by generating

synthetic samples for the minority class. It operates by interpolating between closely situated positive instances, creating new, synthetic data points that augment the underrepresented class. This technique not only helps in achieving a more balanced dataset but also contributes to reducing overfitting problems. By providing a more equitable representation of all classes, SMOTE enables the development of more robust and generalizable models, enhancing overall predictive performance across all classes.

When we first implement our model, we observed a notable discrepancy between the training and testing outcomes, which was a crucial aspect to address for ensuring the reliability and generalizability of our model. Such a discrepancy may indicate the presence of overfitting, where the model performs exceptionally well on the training data but fails to perform as effectively on unseen or testing data. To address these issues and control the substantial discrepancy between training and testing outcomes, we employed several strategies. First and foremost, the feature selection technique using the Brouta algorithm helped us remove irrelevant features, reducing the risk of overfitting due to noisy data.

3.4. Machine Learning Models

In the realm of medical science, machine learning has emerged as a pivotal tool, revolutionizing analysis, diagnosis, and prognosis. Its capacity to process extensive datasets and uncover intricate patterns renders it particularly valuable in cardiology, where timely detection and accurate prediction of cardiac events can be life-preserving. Our research harnesses this potential by employing a variety of machine learning algorithms to forecast heart attacks, with the aim of enhancing diagnostic precision and patient outcomes. Our study utilizes a diverse array of machine learning classifiers, each offering unique advantages:

- Logistic Regression: This foundational algorithm excels in binary classification, making it well-suited for heart attack risk assessment [22].
- K-Nearest Neighbors (KNN): An instance-based learning approach that categorizes new data points based on their proximity to existing data, potentially capturing localized patterns in cardiovascular risk factors [23].
- Decision Tree: This model creates a tree-like structure of decisions based on feature thresholds, providing interpretable rules for evaluating heart attack risk [16].
- Support Vector Machine (SVM): Renowned for its efficacy in high-dimensional spaces, SVM can delineate complex boundaries between high-risk and low-risk individuals [5].
- Random Forest: An ensemble technique that amalgamates multiple decision trees, often yielding robust and accurate predictions while mitigating overfitting [4].
- XGBoost: A sophisticated implementation of gradient boosting, recognized for its superior performance and efficiency in handling structured data [24].

By implementing and comparing these diverse algorithms, our research seeks to identify the most accurate model for heart attack prediction. This comparative analysis not only aids in determining the most effective predictive tool but also provides insights into the strengths and limitations of each algorithm within the context of cardiovascular risk evaluation.

The outcomes of this study have the potential to significantly advance the development of more precise and dependable heart attack prediction systems. Ultimately, this research aims to contribute to improved patient care and outcomes in the field of cardiology, leveraging the power of machine learning to enhance our ability to identify and mitigate cardiac risks.

4. Result and analysis

This study evaluated the performance of six distinct machine learning algorithms in predicting heart attacks. We carefully selected a diverse range of algorithms, encompassing both traditional and advanced techniques, to ensure a comprehensive assessment of predictive capabilities in this critical medical context. To rigorously evaluate each model's effectiveness, we employed four key performance metrics that are particularly relevant for classification tasks in healthcare: Accuracy, Precision, Recall, and F1-Score. Accuracy provides an overall measure of the model's correctness, while Precision indicates the reliability of positive predictions. Recall assesses the model's ability to identify all positive cases, which is crucial in medical diagnostics where missing a potential heart attack case could have severe consequences. The F1-Score offers a balanced measure of precision and recall, providing a single metric that captures both aspects of performance. This multi-faceted approach to evaluation allows for a nuanced understanding of each algorithm's strengths and limitations in the specific context of heart attack prediction. The comprehensive results of our analysis, detailing the performance of each algorithm across all four metrics, are presented in Table 3. This

detailed comparison enables a thorough examination of the relative merits of different machine learning approaches in addressing this vital healthcare challenge, potentially informing future developments in predictive cardiology.

Table 3 Classification report of six models

Algorithms	Accuracy	Precision	Recall	$F_1 - \text{Score}$
LR	69%	0.67	0.66	0.63
KNN	81%	0.80	0.83	0.75
DT	78%	0.75	0.79	0.78
SVM	70%	0.69	0.69	0.71
RF	87%	0.86	0.85	0.87
XGB	80%	0.79	0.81	0.79

RF emerged as the top performer across all metrics. It achieved the highest accuracy at 87%, demonstrating its superior ability to correctly classify both positive and negative cases. The RF model also exhibited the highest precision (0.86), recall (0.85), and F1-score (0.87), indicating a well-balanced performance in terms of minimizing both false positives and false negatives. KNN showed the second-best performance with an accuracy of 81%. It demonstrated particularly strong recall (0.83), suggesting its effectiveness in identifying true positive cases. However, its lower F1-score (0.75) compared to its accuracy indicates some imbalance between precision and recall. XGB and DT algorithms showed comparable performance, with accuracies of 80% and 78% respectively. XGB had a slightly better balance between precision and recall, as reflected in its F1-score of 0.79 compared to DT's 0.78. SVM and LR algorithms demonstrated the lowest performance among the six algorithms tested. SVM achieved 70% accuracy with balanced precision and recall (both 0.69), while LR showed slightly lower performance with 69% accuracy and lower precision (0.67) and recall (0.66).

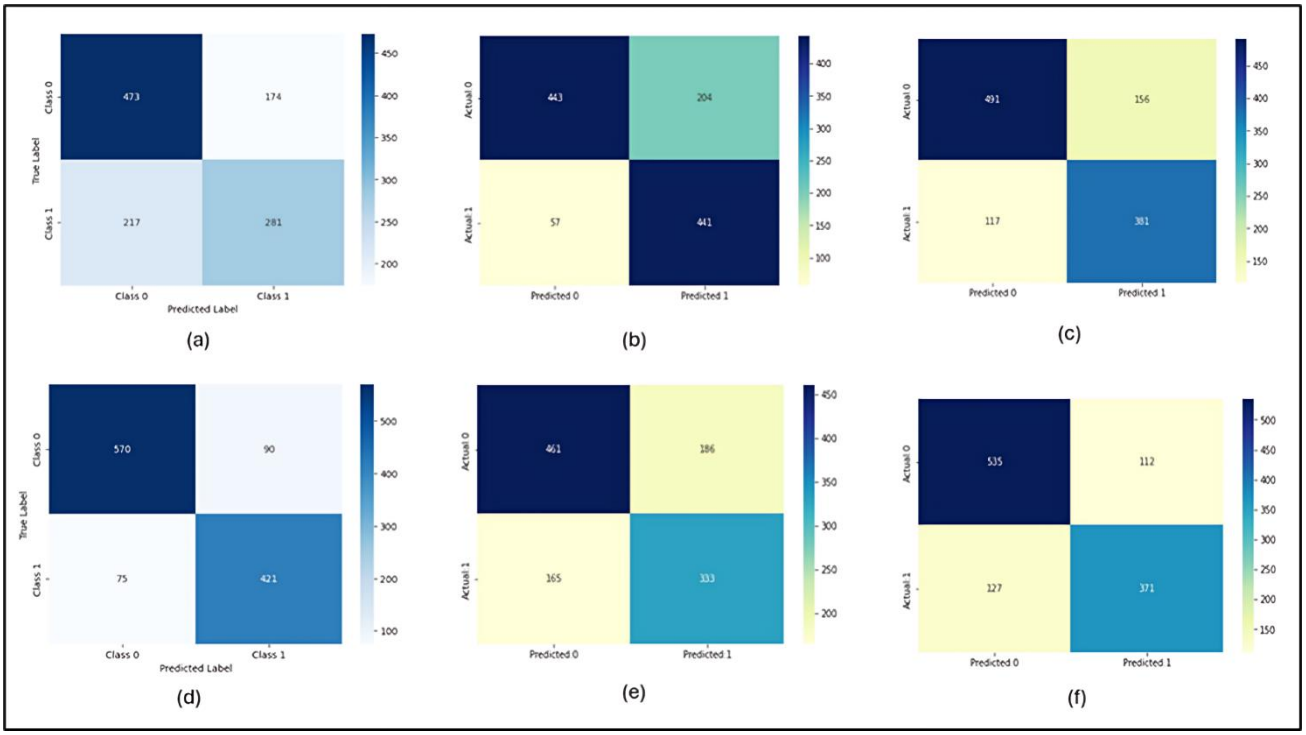

Figure 4 Confusion matrix of all algorithms test result

The superior performance of ensemble methods (RF and XGB) over single decision tree and linear models (LR and SVM) suggests that the relationship between the input features and heart attack risk is complex and nonlinear. The Random Forest algorithm's ability to handle this complexity, combined with its resistance to overfitting, likely contributed to its

top performance. These results underscore the potential of machine learning, particularly ensemble methods like Random Forest, in predicting heart attacks. The high accuracy and balanced performance of the RF model suggest its potential for clinical application in early heart attack risk assessment. However, the choice of algorithm in practice may depend on the specific requirements of the clinical setting, such as the need for model interpretability or computational efficiency. Figure 4 presents the confusion matrix of all algorithms implemented in this research.

5. Discussion

The comparative analysis of the six machine learning classifiers implemented in this study reveals several important insights into their performance for heart attack prediction using electrocardiogram results and other patient data. Random Forest (RF) algorithm remained the top performer, achieving an impressive 87% accuracy. The exceptional performance of this model can be attributed to its ensemble architecture, which integrates multiple decision-making components to create a robust and precise predictive system. This approach's strength lies in its ability to synthesize diverse perspectives, much like a panel of experts collaborating to reach a consensus. The model's capacity to navigate intricate relationships among various risk factors, coupled with its inherent resistance to overspecialization, likely played a crucial role in achieving such high accuracy in heart attack prediction. Furthermore, the model exhibited a well-balanced profile across key performance indicators. Its precision of 0.86 indicates a high rate of correctly identified positive cases, while the recall of 0.85 suggests a strong ability to detect a large proportion of actual heart attack risks. The F1-score of 0.87, which harmonizes precision and recall, underscores the model's overall effectiveness in balancing false positives and false negatives. This equilibrium is particularly crucial in medical diagnostics, where both missed cases and false alarms can have significant consequences.

Notably, the K-Nearest Neighbors (KNN) algorithm showed significant improvement in the new results, achieving 81% accuracy and emerging as the second-best performer. This is a substantial increase from the previously reported 78%. KNN's strong performance, particularly its high recall (0.83), suggests its effectiveness in identifying potential heart attack cases, which is crucial in a medical screening context. The XGBoost (XGB) algorithm maintained its strong performance with 80% accuracy. XGB's gradient boosting approach, which iteratively improves upon weak learners, proved effective in capturing the nuances of the heart attack prediction task. The Decision Tree (DT) algorithm also showed competitive results, with an accuracy of 78%, demonstrating its ability to capture non-linear relationships in the data. The Support Vector Machine (SVM) and Logistic Regression (LR) models achieved lower accuracies of 70% and 69% respectively. While still providing valuable predictions, their relatively lower performance suggests that the relationships in the data may be too complex for these algorithms to fully capture, especially compared to ensemble methods like RF and XGB.

It's worth noting that accuracy alone does not provide a complete picture of model performance. The precision, recall, and F1-scores offer additional insights. For instance, KNN's high recall (0.83) indicates its strength in identifying true positive cases, although its lower F1-score (0.75) suggests some trade-off between precision and recall.

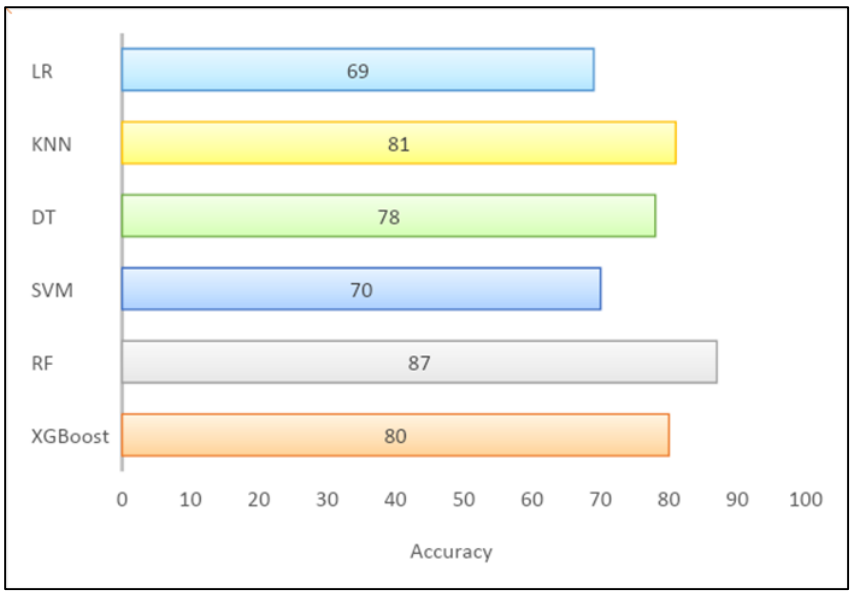

Figure 5 Comparative analysis of six classifiers based on test accuracy

It's worth noting that accuracy alone does not provide a complete picture of model performance. The precision, recall, and F1-scores reported in Table 3 offer additional insights. The RF model, for instance, not only achieved the highest accuracy but also demonstrated well-balanced precision (0.86), recall (0.85), and F1-score (0.87). This balance indicates that RF is effective at correctly identifying both positive and negative cases of heart attack risk, which is crucial in a medical context where both false positives and false negatives can have significant consequences. The superior performance of our approach can be attributed to several factors:

- Feature Selection: The use of the Brouta algorithm for feature selection likely contributed to the improved performance. By identifying the most relevant features and eliminating noise, we provided the models with a more informative and streamlined dataset.
- Data Preprocessing: Our comprehensive approach to data preprocessing, including noise removal from ECG signals and addressing class imbalance using SMOTE, likely enhanced the quality of the input data for the models.
- Diverse Algorithm Selection: By implementing and comparing a range of algorithms, from simple linear models to complex ensemble methods, we were able to identify the most suitable approach for this specific prediction task.
- Dataset Quality: The combination of manually collected data from Dhaka Medical College Hospital with open-source data may have provided a rich and diverse dataset, allowing the models to learn from a wide range of cases.

While our results are promising, it's important to acknowledge some limitations. The performance of machine learning models can vary depending on the specific characteristics of the dataset used. Therefore, further validation on diverse datasets from different populations would be beneficial to establish the generalizability of these results. Furthermore, while accuracy and other metrics provide valuable insights into model performance, the interpretability of the model predictions is crucial in a medical context. Future work could focus on developing methods to explain the predictions of the best-performing models, particularly for complex algorithms like Random Forest and XGBoost. Additionally, investigating the significant improvement in KNN's performance could provide valuable insights into feature importance and data characteristics.

6. Conclusion

This research has made significant strides in applying machine learning techniques for early-stage heart attack prediction, a critical area given the global prevalence of cardiovascular diseases. Leveraging data from Dhaka Medical College Hospital and implementing six distinct machine learning algorithms, our study yielded promising results that could revolutionize early detection strategies. The cornerstone of our methodology was a meticulous pre-processing approach, encompassing noise removal, elimination of null and duplicate values, feature selection, and scaling. These techniques significantly enhanced our dataset quality, contributing to the overall accuracy of our findings. Among the six algorithms applied, the Random Forest algorithm emerged as the top performer, achieving an impressive 87% accuracy. This result not only showcases Random Forest's potential in heart attack prediction but also underscores the importance of comparing multiple algorithms to identify the most effective approach.

While our results are promising, we acknowledge that model accuracy is inherently tied to dataset characteristics, presenting both a limitation and an opportunity for future research. Looking ahead, we aim to explore more complex deep learning algorithms, potentially uncovering more nuanced patterns in heart attack prediction. We plan to implement our methods with a significantly larger dataset to enhance model robustness and generalizability across diverse populations. Additionally, we intend to develop a software model based on our findings, serving as a valuable tool for healthcare professionals in early heart attack risk assessment. Future work will also involve comparing advanced deep learning algorithms against traditional machine learning classifiers. By continuing to refine our approaches and expand our research, we strive to develop accurate, reliable tools for early heart attack prediction, potentially saving countless lives through timely intervention and prevention strategies.

Compliance with ethical standards

Disclosure of conflict of interest

We declare no conflict of interest

References

[1] H. Takci, Improvement of heart attack prediction by the feature selection methods, TURK. J. OF ELECTR. ENG. COMPUT. SCI., vol. 26, pp. 1–10, 2018.

[2] S. Mall, Heart attack prediction using machine learning techniques, in 2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), IEEE, May 2024, pp. 1778–1783.

[3] K. Oliullah, A. Barros, and M. Whaiduzzaman, Analyzing the effectiveness of several machine learning methods for heart attack prediction, in Lecture Notes in Networks and Systems, in Lecture notes in networks and systems. , Singapore: Springer Nature Singapore, 2023, pp. 225–236.

[4] T. Islam, A. Kundu, T. Ahmed, and N. I. Khan, Analysis of arrhythmia classification on ECG dataset, in 2022 IEEE 7th International conference for Convergence in Technology (I2CT), IEEE, Apr. 2022. doi: 10.1109/i2ct54291.2022.9825052.

[5] S. H. Talukder, S. K. Mondal, M. Aljaidi, R. B. Sulaiman, and T. Islam, Heart disease risk assessment and prediction: A robust ensemble approach with extra tree classifier, in 2023 2nd International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI), IEEE, Dec. 2023. doi: 10.1109/eiceeai60672.2023.10590147.

[6] T. Islam, A. Vuyia, M. Hasan, and M. M. Rana, Cardiovascular disease prediction using machine learning approaches, in 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), IEEE, Apr. 2023, pp. 813–819.

[7] C. A. Alexander and L. Wang, Big data analytics in heart attack prediction, J. Nurs. Care, vol. 06, no. 02, 2017, doi: 10.4172/2167-1168.1000393.

[8] P. K. Pande, P. Khobragade, S. N. Ajani, and V. P. Uplanchiwar, Early detection and prediction of heart disease with machine learning techniques, in 2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET), IEEE, Jun. 2024, pp. 1–6.

[9] A. L. Golande and T. Pavankumar, Electrocardiogram-based heart disease prediction using hybrid deep feature engineering with sequential deep classifier, Multimed. Tools Appl., Apr. 2024, doi: 10.1007/s11042-024-19155-2.

[10] V. V. Ramalingam, A. Dandapath, and M. Karthik Raja, Heart disease prediction using machine learning techniques : a survey, Int. J. Eng. Technol., vol. 7, no. 2.8, p. 684, Mar. 2018.

[11] M. Abubakar, A. H. Maidabara, Y. M. Malgwi, and A. Mohammed, Web based heart disease Prediction Model using Machine Learning technique, Comput. sci. IT res. j., vol. 5, no. 2, pp. 518–527, Feb. 2024.

[12] M. A. Islam, M. Z. H. Majumder, M. S. Miah, and S. Jannaty, Precision healthcare: A deep dive into machine learning algorithms and feature selection strategies for accurate heart disease prediction, Comput. Biol. Med., vol. 176, p. 108432, Jun. 2024.

[13] M. Alshraideh, N. Alshraideh, A. Alshraideh, Y. Alkayed, Y. Al Trabsheh, and B. Alshraideh, Enhancing heart attack prediction with machine learning: A study at Jordan University Hospital, Appl. Comput. Intell. Soft Comput., vol. 2024, no. 1, Jan. 2024, doi: 10.1155/2024/5080332.

[14] M. P. LaValley, Logistic regression, Circulation, vol. 117, no. 18, pp. 2395–2399, May 2008.

[15] O. Kramer, K-Nearest Neighbors, in Dimensionality Reduction with Unsupervised Nearest Neighbors, in Intelligent systems reference library. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 13–23.

[16] B. de Ville, Decision trees, Wiley Interdiscip. Rev. Comput. Stat., vol. 5, no. 6, pp. 448–455, Nov. 2013.

[17] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, Support vector machines, IEEE Intell. Syst., vol. 13, no. 4, pp. 18–28, Jul. 1998.

[18] L. Breiman, Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[19] T. Chen and C. Guestrin, XGBoost, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, Aug. 2016. doi: 10.1145/2939672.2939785.

[20] T. Islam et al., Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI, Sci. Rep., vol. 14, no. 1, p. 8487, Apr. 2024.

[21] A. Slowik, Swarm Intelligence Algorithms (Two Volume Set). CRC Press, 2021.

[22] T. Islam et al., Review analysis of ride-sharing applications using machine learning approaches, in Computational Statistical Methodologies and Modeling for Artificial Intelligence, New York: CRC Press, 2023, pp. 99–122.

[23] A. Sheakh, T. A. Sazia, T. A. Islam, and R. J. Lima, Improving hepatits C diagnosis using machine learning techniques, in Artificial Intelligence for Intelligent Systems, Boca Raton: CRC Press, 2024, pp. 241–259.

[24] T. Islam, M. R. Sadik, M. F. R. Islam, T. R. Mona, T. Rahman, and M. M. R. Foysal, Early-stage diabetes risk prediction using supervised machine learning algorithms, in 2023 2nd International Conference on Futuristic Technologies (INCOFT), IEEE, Nov. 2023. doi: 10.1109/incoft60753.2023.10425305.