JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Comparative analysis of AI Models for Cardiovascular Disease Prediction

Soumendra Kumar Mishra[1], Manoranjan Dash[2], Saumendra Pattnaik[3], Suprava Ranjan Laha[4, *], Salankara Sarkar[5]

[1]Student, [2]Professor, [3]Assistant Professor, [4]Assistant Professor, [5]Student
[1]Department of Management Studies
[1] Sikkim Manipal University, India

Abstract : **Cardiovascular disease remains a leading cause of global mortality, responsible for around 12 million deaths annually. Early detection is essential for high-risk individuals to make necessary lifestyle changes, reducing the likelihood of severe complications. This study introduces a machine learning based system for predicting heart disease using several classification algorithms: Random Forest, Decision Tree, Support Vector Machine (SVM), AdaBoost, XGBoost, and Logistic Regression. A dataset containing 76 attributes, with 14 critical features such as age, cholesterol levels, blood pressure, and fasting blood sugar, was used to predict outcomes. The models were evaluated using accuracy, precision, recall, and F1 score. XGBoost achieved the highest accuracy at 82%, followed by SVM with 81% and Logistic Regression with 80%. These results demonstrate the potential of ML techniques to predict cardiovascular disease with high reliability, enabling healthcare professionals to make informed decisions and improve patient outcomes. This research contributes to public health by providing a tool to predict cardiovascular disease more accurately, allowing for earlier interventions. The system helps medical practitioners prioritize high-risk patients and promotes preventive care measures. By enhancing prediction accuracy, this ML approach could lower healthcare costs and improve the quality of life, reducing global cardiovascular mortality rates.**

IndexTerms **- Machine learning, Heart disease prediction, XGBoost, SVM, cardiovascular disease**

I. INTRODUCTION

Cardiovascular disease (CVD) continues to be one of the leading causes of mortality worldwide, with the World Health Organization (WHO) estimating that it accounts for approximately 12 million deaths annually. Heart attacks, in particular, are often termed "silent killers" due to their ability to strike without warning, often resulting in fatal outcomes before symptoms manifest. Early prediction and detection of heart disease are therefore essential, especially for high-risk individuals, as they allow for critical lifestyle adjustments that can significantly reduce complications and mortality rates. Machine learning has emerged as a powerful tool in the medical field, enabling the analysis and prediction of diseases by processing large datasets generated by healthcare institutions [1-3]. Various studies have highlighted ML techniques' effectiveness in improving heart disease diagnostic accuracy. Mishra et al. (2023) used visual analysis through machine learning algorithms to predict cardiac arrest, emphasizing the role of technology in health education and awareness [4]. Similarly, Whig et al. (2022) developed a real-time deep-learning model to detect cardiac arrest, achieving significant accuracy [5]. Other researchers, such as Moffat (2022), focused on comparing different models for predicting in-hospital cardiac arrest, demonstrating the advantages of machine learning in patient monitoring [6]. Javeed et al. (2022) conducted a systematic review of automated diagnostic systems for heart failure prediction, showcasing the potential of ML models across various data modalities [7]. Angraal et al. (2020) developed ML models for predicting mortality and hospitalization in patients with heart failure, reinforcing the relevance of ML in clinical risk assessment [8]. Despite the advancements, there still needs to be a gap in identifying the most suitable ML algorithms for predicting heart disease with high accuracy using multiple patient-specific features.

This work aims to address this gap by developing and comparing several machine learning models—Random Forest, Decision Tree, Support Vector Machine (SVM), AdaBoost, XGBoost, and Logistic Regression—on a dataset containing 14 key attributes relevant to heart disease prediction. The primary objective is to determine the most accurate model for predicting cardiovascular disease and to provide healthcare professionals with a reliable tool for early diagnosis. In this work, various ML algorithms are employed and evaluated for their ability to predict heart disease. A dataset of 76 attributes is analyzed, with 14 critical attributes used to classify the likelihood of a patient developing heart disease. The algorithms are evaluated based on accuracy, precision, recall, and F1 score, aiming to identify the best-performing model for real-world healthcare applications. Through this research, we aim to contribute to the ongoing efforts to integrate machine learning into clinical settings for improved patient outcomes and reduced cardiovascular mortality.

II. METHODOLOGY

This section outlines the methodology to predict heart disease using machine learning algorithms. The process includes dataset collection, feature selection, model development, and evaluation of six distinct machine learning algorithms: Random Forest, Decision Tree, Support Vector Machine (SVM), AdaBoost, XGBoost, and Logistic Regression. The detailed steps are as follows:

A. Dataset Collection

The dataset used for this study is the publicly available UCI Heart Disease dataset, which contains 76 attributes related to heart disease. However, for prediction and optimizing performance, only 14 key attributes were selected based on their relevance to heart disease prediction shows in Table 1.

Table 1: Attributes of the Dataset

Sl No	Type	Description	Attribute
1.	Numerical	Age of the patient (29 to 77)	Age
2.	Nominal	Patient Gender (male-0, female-1)	Sex
3.	Nominal	Type of Chest Pain	Cp
4.	Numerical	Resting blood pressure (measured in millimeters of mercury on admission to the hospital, with readings ranging from 94 to 200)	Trestbps
5.	Numerical	Serum Cholesterol (measured in milligrams per deciliter, with levels ranging from 126 to 564)	Chol
6.	Nominal	Fasting blood sugar level more than 120mg/dl (true-1 false-0)	Fbs
7.	Nominal	The electrocardiographic result at rest (0 to 1)	Resting
8.	Numerical	Maximum heart rate (71 to 202) attained	Thali
9.	Nominal	Agina (yes-1, no-0) was included in the exercise.	Exang
10.	Numerical	Exercise-induced ST depression (0 to.2) compared to rest	Oldpeak
11.	Nominal	The peak exercise ST segment slop (0 to 1)	Slope
12.	Numerical	Count of major vessels (0-3)	Ca
13.	Nominal	3 – normal	Thal
14.	Nominal	1 or 0	Targets

These 14 attributes represent numerical and categorical data and serve as input variables, with the **Target** variable being the output that predicts whether a patient is at risk of heart disease (binary classification: 0 for no heart disease, 1 for heart disease).

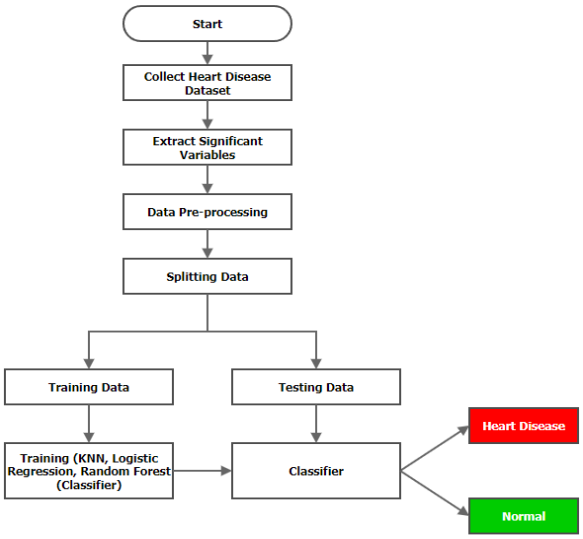

Figure 1: System architecture of our proposed System

Figure 1 illustrates the step-by-step process for heart disease prediction using machine learning algorithms. The flowchart starts with collecting the heart disease dataset, which is subsequently processed through several stages. The significant variables are extracted from the dataset, reducing the dimensionality and focusing only on the key attributes relevant to heart disease prediction. Next, the data undergoes a pre-processing phase where operations like handling missing values, normalization, and encoding categorical variables are performed to prepare the data for model training. After pre-processing, the dataset is split into two parts: training data and testing data. The training data is used to build and train various machine learning classifiers, including K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest. The trained models are then applied to the testing data, which evaluates the classifiers' performance. The classifier makes predictions based on the test data, determining whether a patient is likely to have heart disease (represented in red) or is considered normal (represented in green) based on the significant attributes provided by the dataset.

B. Data Preprocessing

Before model development, the dataset underwent the following preprocessing steps to ensure data quality:

i. Handling Missing Data

Any missing or null values were handled by removing the rows or columns with substantial missing data or imputing missing values using appropriate techniques (e.g., mean imputation for numerical data or mode imputation for categorical data).

ii. Normalization/Scaling

For attributes like age, cholesterol levels, and resting blood pressure, the data was normalized to ensure all variables were on a similar scale, as some machine learning algorithms, particularly SVM and Logistic Regression, are sensitive to the scale of the input data. Min-Max Scaling was applied to rescale the values between 0 and 1.

iii. Encoding Categorical Variables

Categorical variables such as Gender, Chest Pain Type, Fasting Blood Sugar, and Thalassemia were encoded using one-hot encoding for multiclass attributes and label encoding for binary attributes, converting them into numerical format for compatibility with machine learning algorithms.

C. Feature Selection

To reduce dimensionality and improve model performance, feature selection techniques were applied. A correlation matrix was generated to examine the relationship between the attributes and the target variable. Highly correlated features were retained, while minimal correlation to the target variable was removed to reduce noise.

D. Model Development

This study employed six machine learning algorithms, each trained on the preprocessed dataset. The following is a brief description of each algorithm and its implementation:

i. Support Vector Machine (SVM)

SVM is a robust supervised learning algorithm used for classification. SVM aims to find the optimal hyperplane that maximally separates the two classes (presence or absence of heart disease). A linear kernel was initially applied, followed by experiments with non-linear kernels, such as the radial basis function (RBF), to handle any non-linearity in the data [9].

ii. Decision Tree Classifier

The Decision Tree is a non-parametric model that splits the dataset based on specific features to create a tree-like structure. The CART algorithm (Classification and Regression Tree) was used, with entropy and Gini index applied as criteria to select the optimal split at each node [10].

iii. Random Forest Classifier

Random Forest is an ensemble learning method that constructs multiple decision trees using random subsets of the dataset. The final prediction is based on the majority vote of all the individual trees. It improves the model's accuracy and generalization by reducing overfitting [11].

iv. AdaBoost

AdaBoost, or Adaptive Boosting, is a boosting technique that combines multiple weak classifiers to form a robust classifier. Each iteration of the algorithm focuses on previously misclassified instances, thus improving overall classification accuracy [12].

v. XGBoost

XGBoost is an advanced boosting algorithm that builds decision trees sequentially. It reduces errors in previous trees by adjusting the weights of incorrectly predicted instances. XGBoost also incorporates regularization to prevent overfitting and handles missing data efficiently [13].

vi. Logistic Regression

Logistic Regression is a statistical model used for binary classification tasks. It uses the logistic function to estimate the probability of a binary outcome (heart disease or no heart disease). It is a simple yet effective method that serves as a baseline model in many classification tasks [14].

E. Model Evaluation

Each model was trained and tested using a 70/30 train-test split, where 70% of the data was used for training and 30% for testing. To evaluate the performance of each algorithm, the following metrics were computed:

- **Accuracy**: The proportion of correct predictions out of the total predictions.
- **Precision**: The ratio of true positives to the sum of true positives and false positives.
- **Recall**: The ratio of true positives to the sum of true positives and false negatives.
- **F1-Score**: The harmonic mean of precision and recall, providing a balanced evaluation metric.

III. RESULT ANALYSIS & DISCUSSION

The experimentation showed that XGBoost is the most reliable model for predicting heart disease, with an accuracy of 82%, followed closely by SVM at 81%. XGBoost's ability to handle various data types and superior regularization techniques make it

well-suited for this task. On the other hand, SVM's strength lies in its effectiveness with binary classification problems, but its performance can vary with different kernels and hyperparameters.

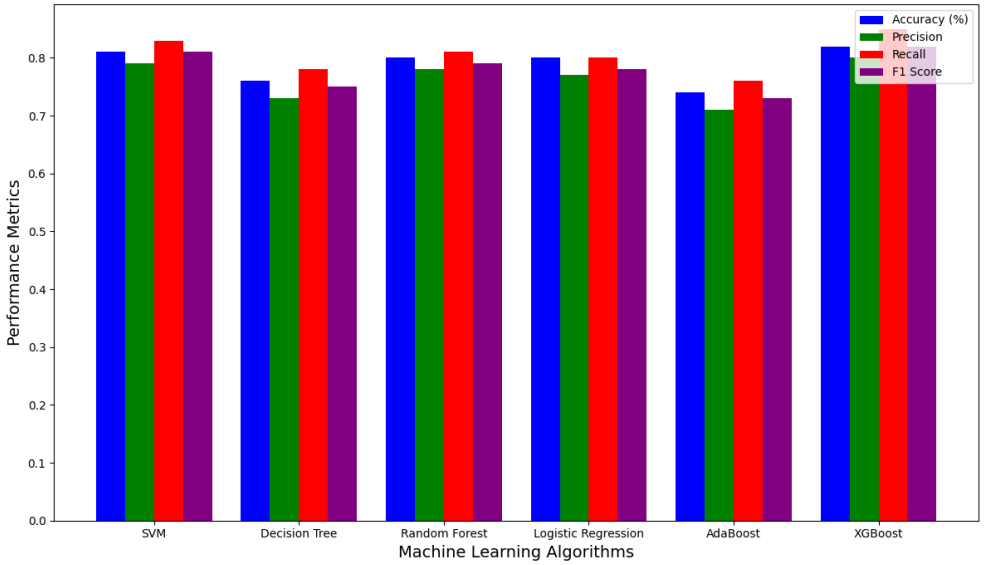

Figure 2: Performance metrics comparison for different machine learning algorithms in heart disease prediction

Figure 2 illustrates the comparison of performance metrics—Accuracy, Precision, Recall, and F1 Score—for six machine learning algorithms used in heart disease prediction: SVM, Decision Tree, Random Forest, Logistic Regression, AdaBoost, and XGBoost. Among these, XGBoost demonstrates the highest performance across all metrics, with the best accuracy (82%), recall (0.85), and F1 Score (0.82), making it the most effective model for heart disease prediction. SVM follows closely with an accuracy of 81% and recall of 0.83, showing strong classification capabilities. Both Random Forest and Logistic Regression perform similarly, each achieving an accuracy of 80%, though Random Forest shows a slightly better balance across recall and F1 Score. Decision Tree, while having decent accuracy (76%), exhibits lower recall and F1 scores, indicating more misclassifications. AdaBoost, with the lowest accuracy (74%) and precision (0.71), performs the weakest, likely due to its sensitivity to noise in the dataset. Overall, XGBoost outperforms the other algorithms across all metrics, with the highest accuracy (82%), recall (0.85), and F1 score (0.82). This indicates that XGBoost is the most effective algorithm for predicting heart disease based on this dataset.

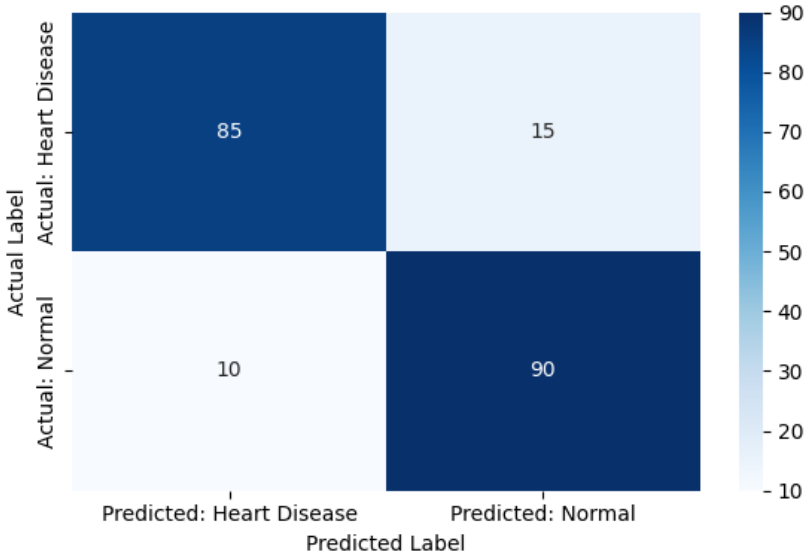

Figure 3: Confusion matrix for the XGBoost model used in heart disease prediction.

Figure 3 shows that the confusion matrix indicates that the XGBoost model accurately distinguishes between patients with heart disease. The XGBoost model demonstrates strong classification performance, correctly identifying 85 instances of heart disease (true positives) and 90 normal cases (true negatives).

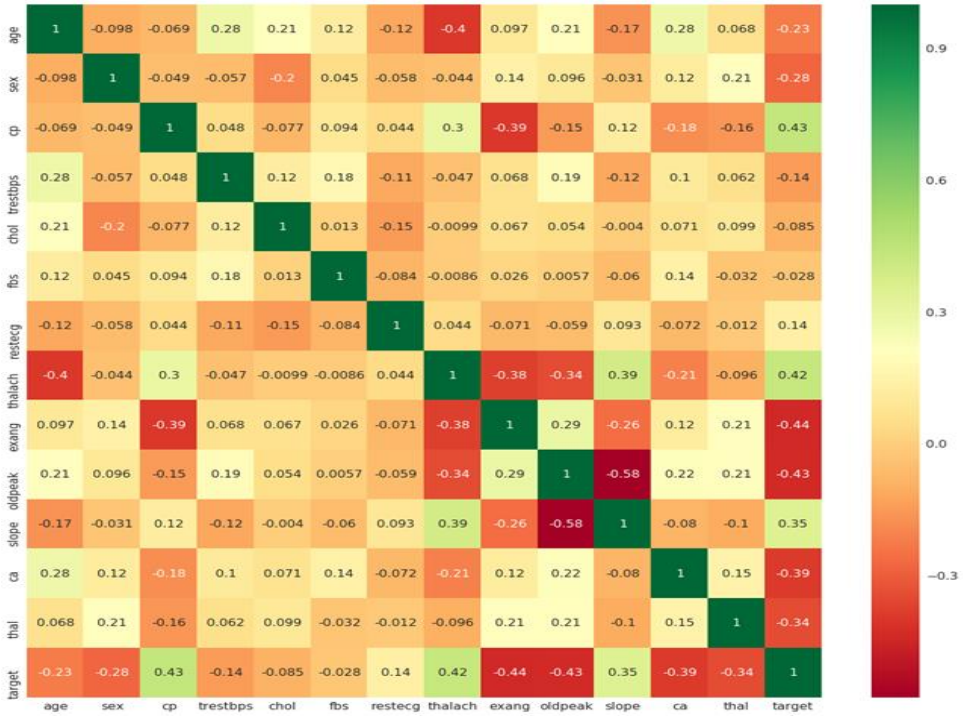

Figure 4: Correlation matrix of the attributes in the heart disease dataset.

Figure 4 shows the relationships between the attributes in the heart disease dataset. Notably, the target variable (indicating the presence of heart disease) is positively correlated with attributes such as cp (chest pain type, 0.43) and thalach (maximum heart rate achieved, 0.42), suggesting that these factors are good indicators of heart disease. On the other hand, the target variable is negatively correlated with exang (exercise-induced angina, -0.44), slope (-0.35), and ca (number of significant vessels, -0.39), implying that lower values of these attributes are associated with a higher likelihood of heart disease. Additionally, correlations among other features (such as age and cholesterol levels) are generally low, indicating limited interdependence between them.

IV. CONCLUSION AND FUTUREWORK

Our work developed a machine learning-based system for predicting heart disease using six different algorithms, including Random Forest, Decision Tree, SVM, AdaBoost, XGBoost, and Logistic Regression. Among these, XGBoost emerged as the most reliable model, achieving the highest accuracy of 82% with strong performance across all metrics, including precision, recall, and F1 score. The ability of machine learning models to predict heart disease based on key attributes such as age, cholesterol levels, and chest pain type offers a significant improvement over traditional methods, allowing for more accurate diagnosis and earlier intervention. The results from this study demonstrate that machine learning can serve as a valuable tool for healthcare professionals, aiding in identifying high-risk individuals and enhancing patient outcomes through timely treatment.

Future research can focus on expanding the dataset to include more diverse and more extensive patient populations, which can improve the generalizability and robustness of the predictive models. While this study focused on traditional machine learning algorithms, incorporating deep learning techniques could yield even higher accuracy, especially with larger datasets. Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) can be explored to enhance prediction performance.

.

REFERENCES

1. Mohapatra, A., Pattnaik, S., Pattanayak, B. K., Patnaik, S., & Laha, S. R. (2022). Software quality prediction using machine learning. In *Advances in Data Science and Management: Proceedings of ICDSM 2021* (pp. 137-146). Singapore: Springer Nature Singapore.
2. Pattnaik, S., Laha, S. R., Pattanayak, B. K., Mohanty, R., Alnabhan, M., & Mohanty, M. N. (2022). Software reliability reckoning by applying neural network algorithm. *Journal of Information and Optimization Sciences*, *43*(5), 1061-1071.
3. Laha, S. R., & Nayak, D. S. K. (2024) Cybersecurity Challenges in IoT-Based Healthcare Systems: A Survey. In *Intelligent Security Solutions for Cyber-Physical Systems* (pp. 203-215). Chapman and Hall/CRC.
4. Mishra, N., Desai, N. P., Wadhwani, A., & Baluch, M. F. (2023). Visual analysis of cardiac arrest prediction using Machine learning algorithms: A health education awareness initiative. In *Handbook of Research on Instructional Technologies in Health Education and Allied Disciplines* (pp. 331-363). IGI Global.
5. Whig, P., Gupta, K., & Jiwani, N. (2022). Real-Time Detection of Cardiac Arrest Using Deep Learning. In *AI-Enabled Multiple-Criteria Decision-Making Approaches for Healthcare Management* (pp. 1-25). IGI global.
6. Moffat, L. M., & Xu, D. (2022). Accuracy of machine learning models to predict in-hospital cardiac arrest: a systematic review. *Clinical Nurse Specialist*, *36*(1), 29-44.
7. Javeed, A., Dallora, A. L., Berglund, J. S., Ali, A., Ali, L., & Anderberg, P. (2023). Machine learning for dementia prediction: a systematic review and future research directions. *Journal of medical systems*, *47*(1), 17.

8. Angraal, S., Mortazavi, B. J., Gupta, A., Khera, R., Ahmad, T., Desai, N. R., & Krumholz, H. M. (2020). Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *JACC: Heart Failure*, *8*(1), 12-21.
9. Yang, C., An, B., & Yin, S. (2018). Heart-disease diagnosis via support vector machine-based approaches. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 3153-3158). IEEE.
10. Maji, S., & Arora, S. (2019). Decision tree algorithms for prediction of heart disease. In *Information and Communication Technology for Competitive Strategies: Proceedings of Third International Conference on ICTCS 2017* (pp. 447-454). Springer Singapore.
11. Pal, M., & Parija, S. (2021). Prediction of heart diseases using random forest. In *Journal of Physics: Conference Series* (Vol. 1817, No. 1, p. 012009). IOP Publishing.
12. El Hamdaoui, H., Boujraf, S., Chaoui, N. E. H., Alami, B., & Maaroufi, M. (2021). Improving heart disease prediction using random forest and adaboost algorithms. *iJOE*, *17*(11), 61.
13. Budholiya, K., Shrivastava, S. K., & Sharma, V. (2022). An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University-Computer and Information Sciences*, *34*(7), 4514-4523.
14. Desai, S. D., Giraddi, S., Narayankar, P., Pudakalakatti, N. R., & Sulegaon, S. (2019). Back-propagation neural network versus logistic regression in heart disease classification. In *Advanced Computing and Communication Technologies: Proceedings of the 11th ICACCT 2018* (pp. 133-144). Springer Singapore.