Evaluating the Efficacy of Machine Learning Algorithms in Heart Disease Prediction

Vaishnavi Devineni[1], Vaishnavi Ratnam Movva[1], Gopichand Medisetty[1], Srilatha Tokala[1], Murali Krishna Enduri[1], Satish Anamalamudi[1]

[1] Algorithms and Complexity Theory Lab, Department of Computer Science and Engineering, SRM University-AP, Neerukonda, Andhra Pradesh, India.

Contributing authors: vaishnavi_devineni@srmap.edu.in; vaishnaviratnam_m@srmap.edu.in; gopichand_m@srmap.edu.in; srilatha_tokala@srmap.edu.in; muralikrishna.e@srmap.edu.in; satish.a@srmap.edu.in;

Abstract

Heart disease analysis, prediction, and early detection are vital challenges within the healthcare domain. The availability of expensive therapies and medical interventions emphasizes the significance of anticipating heart diseases before they reach critical stages. By performing a thorough examination into the prediction of heart illness using various machine learning algorithms, such as XGBoost, Naive Bayes, SVM, Logistic Regression, Random Forest, and LSTM, etc., this work intends to contribute to global healthcare. The comparison of various machine learning algorithm's predictive ability, as measured by their accuracy, recall, precision, and F1-score is the study's primary objective. We conduct this to find the best and most reliable models for predicting heart disease and predict that XGBoost is performing for all the measures. The results of this study may enable healthcare professionals and decision-makers to choose early intervention strategies that will ultimately enhance patient outcomes.

Keywords: heart disease analysis, prediction, machine learning, health care professionals, predictive performance

1 Introduction

Heart disease is the general term for any condition that makes it difficult for the heart to pump blood. Heart illness comes in various forms, but the abilities that are most prevalent are coronary artery disease and heart failure [1]. There will be 17.9 million fatalities worldwide, predicts the World Health Organization [2]. India's top cause of death is thought to be heart disease. As a result, the diagnosis of cardiac disease necessitates the employment of highly skilled and knowledgeable medical personnel [3]. Making diagnostic decisions has benefitted doctors much from the advancement of computer science and machine learning theory. With the advent of artificial intelligence and machine learning, the area of medical diagnostics has undergone a paradigm shift. By utilizing vast amounts of medical data, machine learning algorithms have demonstrated the capacity to uncover intricate patterns and relationships within complex datasets, enabling better-informed clinical decision-making. Machine learning is a technique for manipulating and extracting implicit knowledge from data that was either previously unknown or known [4]. Machine learning is a very diverse field with a wide range of applications that are constantly growing in scope. These days cardiovascular diseases a term that describes a variety of disorders that could harm your heart have grown quite widespread.

Early Diagnosis and Proactive Intervention means the developed heart disease prediction model can be integrated into healthcare systems to aid in early diagnosis. By accurately predicting heart disease risk, healthcare facilities can allocate resources more efficiently. High-risk patients can be prioritized for screenings and interventions, ensuring that critical medical resources are utilized effectively. In remote or underserved areas, the heart disease prediction model can be integrated into telemedicine platforms. This facilitates remote monitoring and enables healthcare professionals to reach patients who may have limited access to specialized medical services. The machine learning methods employed in this study encompass a range of algorithms, including XGBoost, Naive Bayes, Support Vector Machine, Logistic Regression, Random Forest, KNN classifier, Decision tree, as well as various neural network models like ANN, RNN, and LSTM. The selection of the optimal algorithm hinges on meticulous performance evaluation metrics such as Accuracy, Precision, F1 score, and Recall for each algorithm. These metrics provide crucial insights into the machine's accuracy in disease prediction based on the provided parameters. These models and algorithms offer numerous essential advantages, including the ability to assess risks effectively and more.

To examine the provided data, it is crucial to use machine learning to find discrete patterns that are concealed. This prediction must be made utilizing Python libraries along with machine learning algorithms. Machine learning techniques are utilised after data analysis to help in the early detection and prognosis of cardiac disease. For this investigation, a data collection with the following variables was used (marked $0 or 1$): RestingECG, ExerciseAngina, RestingBP, Cholesterol, FastingBS, MaxHR, Sex, Oldpeak, ST_Slope, Age, ChestPainType, and HeartDisease.

2 Related Works

Many scientists and academics are working to apply machine learning techniques in various healthcare sectors [5]. The healthcare sector is undergoing a significant transition and development thanks to machine learning. Many researchers have created prediction models using their own methodologies, and a few of them have also tried fusing different strategies to create hybrid models in an effort to increase accuracy.

Using different algorithms Kavitha *et al.* created a hybrid model that relies on random forest probabilities. rain data is enhanced by incorporating the probabilities from a random forest model and then fed into the decision tree algorithm [6]. Likewise, test data is augmented with probabilities obtained from the decision tree model. Subsequently, predictions are made using these processed data points. Remarkably this model showed an accuracy of 88%.

In their survey, Chala Beyene and Pooja Kamat highlighted the variability in skills and experience among medical practitioners, leading to potential errors in decision-making and critical situations for patients [7]. To address this issue, disease occurrence prediction has become essential. They employed various algorithms and however, a limitation of their method is the absence of performance prediction for each algorithm in implementing the proposed system. On the positive side, their approach offers several advantages. Through the use of various feature selection techniques and algorithms, it enhances the existing decision-making process. The proposed approach facilitates automatic diagnosis, leading to faster results, which, in turn, enhances Quality of Services (QoS) and reduces costs associated with saving a patient's life through timely intervention in heart disease prediction. This survey was done in the year 2018 and according to their research, the model presented by them has shown an accuracy of 94.56% in predicting early cardiovascular diseases.

Arun *et al.* propose a Cloud Computing (CC) based cardiovascular illness prediction system using Naive Bayes (NB) and AES algorithms [8]. The method ensures secure execution and analysis of diverse symptoms and interventions. The regulated dissemination of discovered information to the patient's family via a social networking system is a downside, though. Nevertheless, the advantage lies in AES encryption for securing sensitive but unclassified data. The CAM system uses key private proxy re-encryption to move the computational load to the cloud for mobile health monitoring that protects privacy.

Several computational methods, such as Genetic Algorithms (GA) and Recurrent Fuzzy Neural Networks (RFNN), are suggested for investigating heart ailments by Kaan *et al.* [9]. This method's requirement for the participation of medical professionals during the evaluation of many factors that could affect the decision-making process is one of its drawbacks. On the other hand, the method uses 297 worth of patient data instances altogether, of which 45 are used for training and 252 are used for testing. It remarkably obtains an accuracy of 97.78% on the testing set. Using information from a heart disease testing dataset, the study is able to generate a number of parameters, including accuracy, Root Mean Square Error (RMSE), risk of misclassification error, sensitivity, specificity, F-score, and precision.

A survey done by Ramya *et al.* put forward interesting research by comparing various prediction-based models done by fellow researchers and summarized the existing research concerning heart disease prediction [10].

3 Methodology

In this project, we will assess, forecast, and identify cardiac disease using a variety of machine learning methods.

3.1 Preprocessing the datasets

By identifying the properties that are most important to the analysis, EDA serves as a first step for feature selection or engineering and tries to give a comprehensive grasp of the dataset's structure and features. Analysts can understand the distribution of variables, spot probable outliers, and spot trends or clusters in the data by using histograms, scatter plots, box plots, and other graphical tools created by EDA. A correlation map as shown in Fig.1 shows the relationship between each characteristic property.

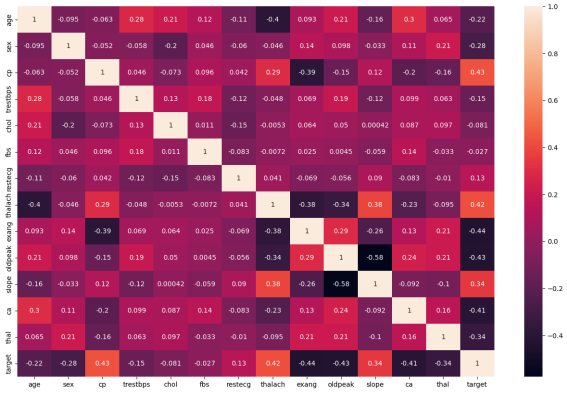

Fig. 1 Correlation matrix of heart disease dataset.

3.2 Implementing various machine learning models

The next phase of this research will involve using several machine learning algorithms to forecast cardiac disease. Collecting pertinent data and resolving missing values, outliers, and feature encoding are the initial steps in preprocessing a dataset before applying a machine learning algorithm to it. Create training, optional validation, and test sets from the data after that. Depending on the type of problem and the amount of the dataset, choose the most relevant features and machine learning technique. Train the model and tweak its hyperparameters to get the best results. Utilise validation data to gauge the model's performance before putting it to the test with fictitious data. If the model performs as expected, use it to create predictions based on fresh

data. Continuously monitor and maintain the model to ensure accuracy and ethical considerations. This process is represented in the flowchart as shown in Fig.2.

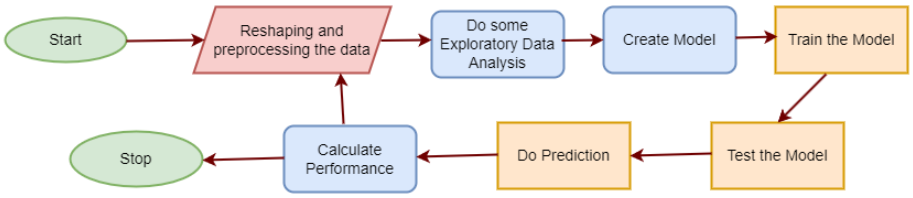

Fig. 2 Flowchart for various machine learning models.

3.2.1 XGBoost

The advanced gradient boosting approach known as XGBoost successively assembles a group of weak learners (often decision trees) [11]. XGBoost employs decision trees as its weak learners due to their ability to capture complex relationships in data. Sequentially adding trees allows for the correction of earlier trees' mistakes. XGBoost uses gradient descent optimization to minimize the objective function with regard to the model's predictions, XGBoost determines the gradient of the loss function and updates the model in the opposite direction of the gradient. XG Boost allows for k-fold cross-validation to estimate model performance and select hyperparameters effectively.

3.2.2 Random Forest

The supervised learning method, which includes Random Forest, contains predictions from several decision trees built during training and is the recommended algorithm to improve accuracy and avoid overfitting. A form of cluster classification model called random forest is applicable to classification, regression, and feature selection issues [12]. Despite its strengths, Random Forest might not perform well on datasets with strong linear relationships or highly correlated features. It might also struggle with capturing complex interaction patterns in data.

3.2.3 Naive Bayes

Notably, the Naive Bayes Classifier has showcased its effectiveness as a classifier, consistently yielding favorable results across various tasks and domains [13]. Naive Bayes can handle both categorical and continuous features, making it versatile in dealing with different types of medical data commonly found in disease prediction studies. The probabilistic nature of Naive Bayes makes it interpretable. It can help physicians and researchers understand the model's decision-making process by revealing the likelihood that a particular attribute would affect the disease prognosis.

3.2.4 Support Vector Machine

In order to perform classification tasks such as predicting heart disease this machine learning method, is frequently utilized. SVM's uniqueness also lies in its use of kernel functions for implicit feature mapping and its strong generalization capabilities [14]. By appropriately tuning its parameters and preprocessing the data, SVM can be a valuable tool for accurate heart disease prediction in a research setting. The approach can implicitly translate the input features to a higher-dimensional space thanks to kernel functions. In the initial feature space, this transformation aids SVM in locating non-linear decision boundaries.

3.2.5 Long Short-Term Memory

Long short-term memory is employed to address the operation, storage explosion, and potential gradient disappearance of the long input sequence during training, in contrast to the normal cyclic neural network [15]. LSTMs make use of a memory cell that can maintain data across lengthy sequences. Data must be normalized and reshaped to satisfy the model's input requirements in order to be appropriately prepared for LSTMs.

3.2.6 Logistic Regression

The most popular use of the supervised machine learning method logistic regression is binary classification, where the goal is to calculate the probability that a certain instance belongs to a particular class. By utilizing the sigmoid (logistic) function in logistic regression, the linear combination of input characteristics is transformed into a probability score between 0 and 1. Unlike linear regression, which predicts continuous values, logistic regression predicts the probability of a binary result (such as 0 or 1). To make class predictions, the anticipated probability might be thresholded. A binary, or something that can take two values and be mixed with true or false, yes or no, etc., is the end outcome in most asset disposal models [16].

4 Results

The assessment metrics listed below may be used to compare different machine learning algorithms for predicting cardiac disease.

Confusion matrix: Following data processing, the confusion matrix of the majority of classification models is taken into consideration. An analysis of a model's performance on a set of test data is summarized by a confusion matrix. The confusion matrix functions by contrasting the expected value produced by machine learning with the actual value [17]. Let us consider u is True Positives, v is True Negatives, w is False Positives, and x is False Negatives.

Accuracy: Accuracy is one factor to consider when rating categorization models. It assesses the overall accuracy of the system. It establishes the proportion of

predictions that were accurate.

$$Accuracy = \frac{u + v}{(u + v + w + x)}$$

Precision: The precision formula determines the proportion of correctly anticipated outcomes among all possible outcomes. It demonstrates the system's propensity for correctly forecasting cardiac problems.

$$Precision = \frac{u}{(u + w)}$$

Recall: Recall, sometimes referred to as sensitivity, is the ratio of the model's accurate predictions to its total accuracy. It demonstrates how well the technology detects heart ailments.

$$Recall = \frac{u}{(u + x)}$$

F1 Score: A common performance metric is the F1-score, which is often used in machine learning and binary classification tasks.

$$F1Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

In this paper for every model, the performance analysis is done by calculating four performance factors. For each model considered in our study, we meticulously calculated these four performance metrics, which are widely recognized in the machine learning community for assessing classification and prediction tasks. These metrics provide a holistic view of a model's effectiveness by considering different aspects of its performance.

Fig.3 visually depicts the outcomes of our performance analysis, showcasing the relative performances of the different machine learning models we examined. The graphical representation enables a clear comparison between the models across the different performance metrics, aiding in the identification of trends and patterns. Upon conducting an in-depth analysis of the results, we found that among the various models under consideration, XGBoost emerged as the standout performer across all four metrics: Accuracy, Precision, Recall, and F1 score. XGBoost's consistently strong performance in these metrics suggests its suitability for heart disease prediction. This research is being done to identify the most accurate models for foretelling heart disease. The results of this study may enable healthcare professionals and decision-makers to choose early intervention strategies that will ultimately enhance patient outcomes. Table.1 demonstrates the overall performance analysis of ten algorithms for the heart disease dataset.

5 Conclusion and Future Work

In conclusion, this research endeavor delved into an extensive performance analysis of ten distinct models for heart disease prediction. Through meticulous evaluation, it

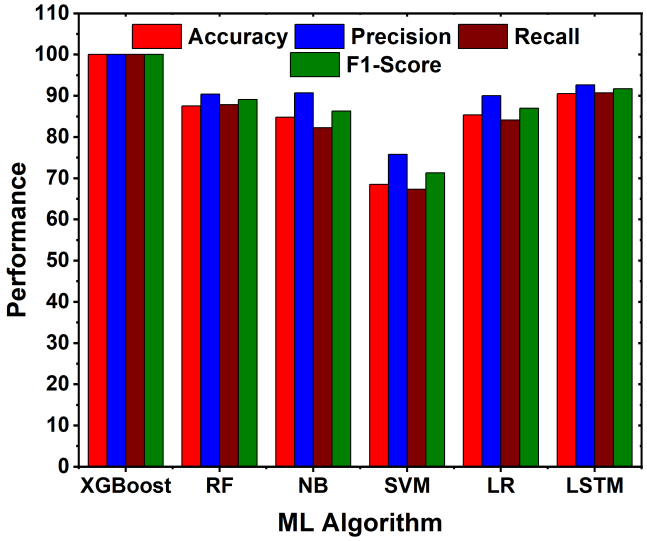

Fig. 3 Bar Plots showcasing various performance measures namely using machine learning algorithms.

Table 1 Performance of 10 distinct machine learning models showing its capability in heart prediction.

ALGORITHM	ACCURACY	PRECISION	RECALL	F1-SCORE
XGBoost	99.98	99.98	99.98	99.98
Naive Bayes	84.78	90.72	82.24	86.27
Random Forest	87.5	90.38	87.85	89.09
Logistic Regression	85.32	90.0	84.11	86.95
LSTM	90.48	92.65	90.67	91.72
SVM	68.47	75.78	67.28	71.28
K-Nearest Neighbors	86.95	91.91	85.04	88.34
Decision Tree	79.82	87.11	76.54	81.70
ANN	88.32	92.52	86.72	89.17
RNN	89.36	94.14	87.84	90.50

became evident that the XGBoost model exhibited unparalleled excellence, emerging as the frontrunner in efficiency, effectiveness, and accuracy. The empirical evidence gathered substantiates its prowess in medical prognostication, thereby warranting its precedence as a preferred choice. This study not only underscores the significance of methodical model selection but also underscores the potential of XGBoost to significantly impact clinical decision-making and patient care in the realm of cardiovascular health.

This work highlights the outstanding effectiveness of the XGBoost model in heart disease prediction, and highlights numerous directions for further research and development. Firstly, investigating the interpretability of the XGBoost model's predictions

could provide valuable insights into the underlying features and patterns contributing to its accuracy.Exploring ensemble approaches, which incorporate the advantages of many models, like XGBoost, may also result in even more reliable forecasting skills. Additionally, the incorporation of domain-specific medical knowledge and the integration of advanced feature engineering methodologies could further refine the model's precision and generalizability. Furthermore, as data availability and quality continue to evolve, ongoing research could encompass developing dynamic and adaptable models to accommodate changing patient demographics and health trends. Lastly, rigorous validation and clinical testing are essential before the widespread clinical implementation of the XGBoost model. Future research may fully utilize machine learning for improved heart disease prediction and patient treatment by exploring these prospective approaches.

References

[1] Katarya, R., Srinivas, P.: Predicting heart disease at early stages using machine learning: A survey. In: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 302–305 (2020). IEEE

[2] Kumar, K.P., Rohini, V., Yadla, J., VNRaju, J.: A comparison of supervised learning algorithms to prediction heart disease. In: International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering, pp. 1–5 (2023). IEEE

[3] Jabbar, M., Deekshatulu, B., Chandra, P.: Prediction of heart disease using random forest and feature subset selection. Innovations in Bio-Inspired Computing and Applications, 187 (2016)

[4] Soni, J., Ansari, U., Sharma, D., Soni, S., *et al.*: Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications **17**(8), 43–48 (2011)

[5] Zhang, J., Lafta, R.L., Tao, X., Li, Y., Chen, F., Luo, Y., Zhu, X.: Coupling a fast fourier transformation with a machine learning ensemble model to support recommendations for heart disease patients in a telehealth environment. Ieee Access **5**, 10674–10685 (2017)

[6] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y.R., Suraj, R.S.: Heart disease prediction using hybrid machine learning model. In: 6th International Conference on Inventive Computation Technologies, pp. 1329–1333 (2021). IEEE

[7] Beyene, C., Kamat, P.: Survey on prediction and analysis the occurrence of heart disease using data mining techniques. International Journal of Pure and Applied Mathematics **118**(8), 165–174 (2018)

[8] Dhanashree, S.M., Mayur, P.B., Shruti, D.D.: Heart disease prediction system using naive bayes. International Journal of Enhanced Research in Science

Technology & Engineering **2**(3), 2319–7463 (2013)

[9] Uyar, K., İlhan, A.: Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. Procedia computer science **120**, 588–593 (2017)

[10] Franklin, R.G., Muthukumar, B.: Survey of heart disease prediction and identification using machine learning approaches. In: 3rd International Conference on Intelligent Sustainable Systems, pp. 553–557 (2020). IEEE

[11] Yang, J., Guan, J.: A heart disease prediction model based on feature optimization and smote-xgboost algorithm. Information **13**(10), 475 (2022)

[12] Liu, Y., Zhang, M., Fan, Z., Chen, Y.: Heart disease prediction based on random forest and lstm. In: 2nd International Conference on Information Technology and Computer Application, pp. 630–635 (2020). IEEE

[13] Karaca, Y., Cattani, C.: Computational Methods for Data Analysis. Walter de Gruyter GmbH & Co KG, ??? (2018)

[14] Learning, M.: Heart disease diagnosis and prediction using machine learning and data mining techniques: a review. Advances in Computational Sciences and Technology **10**(7), 2137–2159 (2017)

[15] Alazab, M., Khan, S., Krishnan, S.S.R., Pham, Q.-V., Reddy, M.P.K., Gadekallu, T.R.: A multidirectional lstm model for predicting the stability of a smart grid. IEEE Access **8**, 85454–85463 (2020)

[16] Ambesange, S., Vijayalaxmi, A., Sridevi, S., Yashoda, B., *et al.*: Multiple heart diseases prediction using logistic regression with ensemble and hyper parameter tuning techniques. In: Fourth World Conference on Smart Trends in Systems, Security and Sustainability, pp. 827–832 (2020). IEEE

[17] Miranda, E., Bhatti, F.M., Aryuni, M., Bernando, C.: Intelligent computational model for early heart disease prediction using logistic regression and stochastic gradient descent (a preliminary study). In: 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI), vol. 1, pp. 11–16 (2021). IEEE