Journal of Network and Information Security
11 (2) 2023, 01-08
http://www.publishingindia.com/jnis

Effective Heart Disease Prediction using Machine Learning Algorithms

Joban K. Joseph[1], Jahana Jabbar[2], Manual Soman[3*], Nasweeba K. N.[4] and Athira Manikuttan[5]
[1]Student, Department of Computer Science, SCMS School of Engineering & Technology, Karukutty, Ernakulam District, Kerala, India.
Email: jobankjoseph@gmail.com
[2]Student, Department of Computer Science, SCMS School of Engineering & Technology, Karukutty, Ernakulam District, Kerala, India.
Email: jahanajabbar@gmail.com
[3]Student, Department of Computer Science, SCMS School of Engineering & Technology, Karukutty, Ernakulam District, Kerala, India.
Email: manualsoman2244@gmail.com
[4]Student, Department of Computer Science, SCMS School of Engineering & Technology, Karukutty, Ernakulam District, Kerala, India.
Email: nasweeba19@gmail.com
[5]Assistant Professer, Department of Computer Science, SCMS School of Engineering & Technology, Karukutty, Ernakulam District, Kerala, India.
Email: athira@scmsgroup.org
[*]Corresponding Author

Abstract: Cardiovascular diseases are a leading cause of death globally, resulting in 17.9 million deaths each year, according to a new report by the World Health Organization. However, with the advancement of technology, machine learning approaches have shown promise in the health industry, providing an opportunity to diagnose and treat heart disease at an earlier stage. In this research project, we aim to build a machine learning model to predict the likelihood of heart disease based on relevant factors. We use a Kaggle heart disease dataset, which includes a comprehensive list of factors related to heart disease, and employ various machine learning algorithms such as Naive Bayes, Support Vector Machine, Random Forest, K-NN, and Decision Tree. Our results indicate that Random Forest provides better prediction accuracy in less time than other machine learning approaches, making it an effective decision support system for medical professionals. This project has the potential to improve the diagnosis and treatment of heart disease and ultimately save lives.

Keywords: K-NN, Naïve Bayes, Random forest, Support vector machine.

I. Introduction

The primary focus of human civilization is on healthcare, with the World Health Organization stating that every individual has a fundamental right to good health. Despite this, heart-related illnesses account for over 31% of all global fatalities, particularly in poorer nations where access to diagnostic facilities and qualified physicians is limited. As a result, medical aid software utilizing computer technology and machine learning techniques is being developed to support early diagnosis and treatment of cardiac diseases, with the aim of reducing the risk of death through early detection.

Machine learning is a growing subfield of AI research that involves developing algorithms to handle a wide range of tasks such as prediction, classification, and decision-making. In the field of medicine, ML techniques are applied to comprehend patterns in large and complex healthcare datasets to derive predictions and insights. ML algorithms are capable of handling big data and making predictions based on real-time input and historical data. Such a machine learning framework for cardiac sickness prediction could encourage cardiologists to act more quickly and enable more patients to receive drugs in a shorter period of time, potentially saving more lives.

Numerous authors [1-3] have made significant efforts to predict heart disease using machine learning algorithms, but this paper is an additional effort to benchmark the Kaggle heart disease prediction dataset while comparing five widely used ML techniques to determine which is the most accurate. This study contributes to the wider body of literature on the subject and seeks to improve the accuracy of existing prediction models.

This paper's structure is as follows: In Section I, we underscore the critical importance of early cardiac disease diagnosis and treatment, emphasizing its role in reducing fatalities. Section II, 'Related Works,' reviews prior research on machine learning methods' classification accuracy. Section III elaborates on our methodology, detailing data collection and analysis techniques. Section IV delves into the machine learning algorithms employed in cardiac disease prediction. Section V introduces the role of Jupyter Notebook in data analysis. Section VI describes the use of machine learning techniques to predict heart disease using a subset of features from the Kaggle dataset. The best-performing algorithm is chosen based on accuracy, and various evaluation criteria like accuracy, confusion matrix, precision, recall, and f1-score are employed to assess the system's performance in predicting cardiac diseases. In Section VII, 'Results,' we thoroughly examine the outcomes of our research. We specifically highlight the Random Forest model, which achieved an impressive 80% accuracy, underscoring its superiority among the tested algorithms for cardiac disease diagnosis. Finally, Section VIII, 'Conclusion and Future Scope,' summarizes our findings and offers insights into potential future research and application prospects, providing a comprehensive overview of our study.

II. Related Works

Many research studies have been conducted to assess the classification accuracy of different machine learning methods. The importance of ML techniques in various fields has been demonstrated through multiple applications explored by Yu-Xuan Wang and colleagues [4]. In their study, they proposed a novel approach to create a functional framework utilizing various machine learning techniques. The data miner examined all the information gathered from the structure to produce accurate results. Thus, Wang *et al.* work highlight the potential of machine learning in developing effective frameworks for real-world applications. The different testing revealed that the suggested strategy produced excellent outcomes. A prior piece on analytics and data mining applications was proposed by Zhiqiang Ge *et al*. in 2017. These processes were employed in the business world for a variety of reasons. They have examined 10 supervised learning algorithms and 8 unsupervised learning algorithms here [5]. In their study, Wang *et al.* showcased the application of semi-supervised learning algorithms. Interestingly, in industry, a majority of applications (between 90-95%) rely on a combination of supervised and unsupervised machine learning techniques. This highlights the critical role that machine learning plays in designing unique and effective applications across fields such as industry and medical services. Overall, the research underscores the importance of utilizing a range of machine learning approaches to tackle real-world problems.

III. Methodology of System

The initial step in the processing of our system involved collecting the dataset. We obtained the

dataset from Kaggle, a well-known platform for data science competitions and projects. The dataset we used has been widely used and verified by numerous researchers in the field.

A. Data Collection

The first building a heart disease prediction system starts with collecting and selecting appropriate datasets for training and testing. A split of 73% was chosen for the training dataset and 37% for the testing dataset. The Scikit-learn library provides functions for data splitting, including the train_test_split() function, which we used to split the dataset randomly. This function ensures that the distribution of the data is maintained in both the training and testing sets. This step is crucial in ensuring that the model is trained on a diverse range of data and can generalize well to unseen data during testing.

B. Attribute Selection

Selecting the most relevant attributes or features for a machine learning model is crucial for accurate predictions, especially for heart disease prediction. Attributes such as age, gender, blood pressure, cholesterol level, family history, smoking, and exercise habits are typically used. Choosing the right set of attributes involves domain knowledge, data analysis, and experimentation to identify the features that have the most significant impact on the target variable while minimizing irrelevant or redundant features. Overfitting and underfitting can occur if there are too many or too few features, respectively. Preprocessing techniques such as scaling, normalization, imputation, and feature engineering can also improve model performance by making the data more meaningful and reducing noise and redundancy.

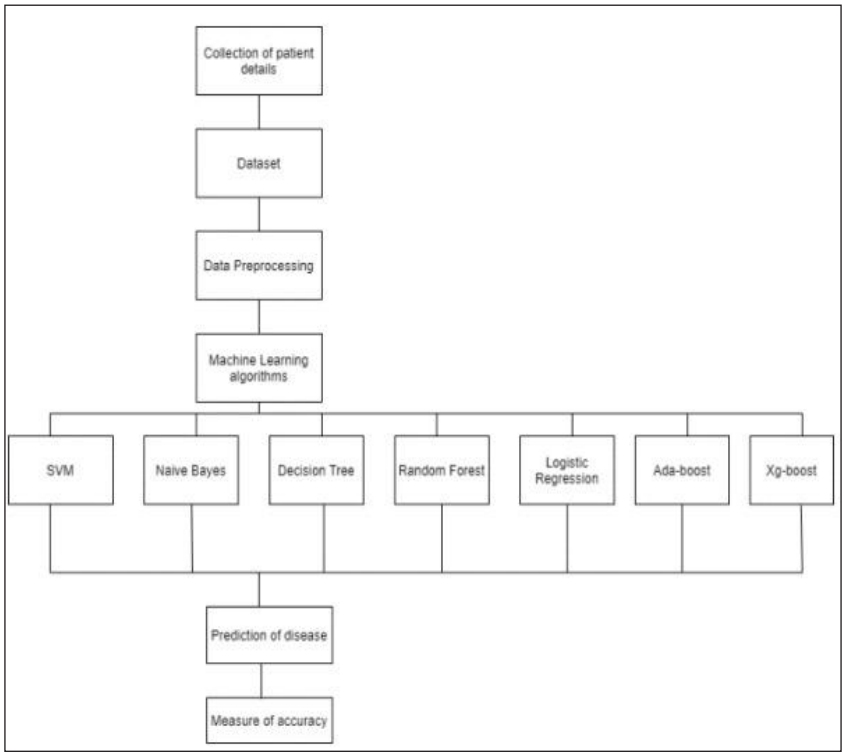

Fig. 1: Architecture of Prediction System

TABLE I: ATTRIBUTES OF THE DATASET

No	Features	Description	Scale
1	Age	Age in years	29 - 77
2	GD	Gender	Female (0), Male (1)
3	CP	Chest pain type	Typical angina (1), Atypical angina (2), Non-angina pain (3), Asymptomatic (4)
4	trestbps	Resting blood pressure on admission to the hospital (mm/Hg)	94 - 200
5	chol	Serum cholesterol (mg/dl)	126 - 564
6	Fbs	Fasting blood sugar is greater than 120 mg/dl	No (0), Yes (1)
7	Restecg	Resting electrocardiographic results	Normal (0), Having ST-T wave abnormality (1), Showing probable or definite left ventricular hypertrophy by Estes' criteria (2)
8	Thalach	Maximum heart rate achieved (ppm)	71 - 202
9	Exang	Exercise induced angina	No (0), Yes (1)
10	Oldpeak	ST depression induced by exercise relative to rest	0 - 6,2
11	slope	The slope of the peak exercise ST segment	Up sloping (0), Flat (1), Down sloping (2)
12	ca	Number of major vessels colored by fluoroscopy	0-3
13	Thal	The heart status	Normal (3), Fixed defect (6), Reversible defect (7)
14	num	Diagnosis of heart disease	Healthy (0), Patient has heart disease (1)

C. Preprocessing of Data

Preprocessing is an essential step in achieving high-quality results from machine learning algorithms [6]. For instance, the Random Forest algorithm cannot handle null values, which means that we must handle missing values in the original raw data. This can be done by either deleting the rows or filling the missing values with appropriate methods such as mean, median or mode.

Another preprocessing step is converting categorical values into numerical values. One of the common methods to do this is through dummy coding, where each category is converted into a binary variable that takes on a value of 0 or 1. This allows the algorithm to handle categorical variables as numeric variables.

D. Data Balancing

Data balancing is a critical step for obtaining accurate results from machine learning algorithms. By balancing the data, we ensure that both target classes are equally represented, which can be seen in the balanced distribution of the classes on a graph. In the context of heart disease prediction, the target classes are represented by "0" for patients with heart diseases and "1" for patients without heart diseases.

IV. MACHINE LEARNING ALGORITHMS

- *Logistic Regression [7]:* A sigmoid function is used in the binary classification technique of logistic regression to forecast the likelihood that an observation will belong to a particular class. The continuous output is converted to a binary value using the sigmoid function after fitting a linear regression model to the input features.
- *Naive Bayes:* Using Bayes' theorem [8], Naive Bayes is a probabilistic classification algorithm [9] that estimates the likelihood that an observation will fall into a particular class. It determines the conditional probability of each feature given the class under the assumption that the input features are independent of one another.
- *Support Vector Machine (SVM) [10]:* Support Vector Machine (SVM) is a binary classification method that identifies the best hyperplane for classifying the input characteristics. In order to achieve better separation, the algorithm translates the input features to a high-dimensional space and maximises the distance between the hyperplane and the closest data points from each class.
- *K-Nearest Neighbours (KNN) [11]:* This non-parametric classification technique predicts the class of an observation based on the classes of its k-nearest neighbours in the feature space. The algorithm determines the separation between the training data and the input features before choosing the k-nearest neighbours to produce the prediction.
- *Decision Tree [12]:* A binary classification algorithm, Decision Tree iteratively divides the input features into various branches based on the features' values. By reducing the impurity of

the resulting subsets, the algorithm chooses the best split and predicts the class of an observation based on the path it travels through the tree.

- *Random Forest [13]:* An ensemble classification technique known as Random Forest mixes various decision trees to increase accuracy and decrease overfitting. Individual decision trees are trained using subsets of features and data points that are randomly chosen by the algorithm, and the predictions of the trees are then combined to get the final classification.
- *XGBoost [14]:* This gradient boosting approach turns several weak learners into a powerful learner by combining them. Decision trees are used by the method as weak learners, and fresh trees are iteratively trained to fix the mistakes of the old ones. The weighted average of all the trees' forecasts make up the final projection.

V. About Jupyter Notebook

Jupyter Notebook has become a popular tool for data scientists and researchers due to its user-friendly interface and versatile functionalities. With its ability to combine rich text elements and code in a single document, Jupyter Notebooks provide a perfect platform for sharing and communicating data analysis results. Users can include images, equations, links, and interactive visualizations to enrich their analysis description, making it more accessible and understandable for others. In addition, the Notebook allows for the execution of code and real-time data analysis, making it an ideal tool for exploring data and testing hypotheses. It supports various programming languages, including Python, R, and Julia, and can be used for a wide range of applications, from data cleaning and preprocessing to machine learning and deep learning. With Jupyter Notebook, users can collaborate, share and reproduce their research findings in a more efficient and transparent way. Overall, Jupyter Notebook is a powerful and flexible tool that provides a unique and interactive environment for data analysis and visualization.

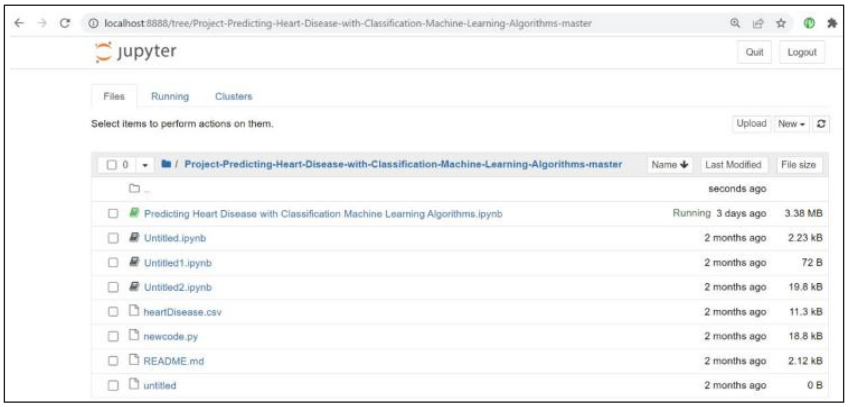

Fig. 2: Jupyter Notebook

VI. Performance Analysis

Several machine learning techniques, including SVM, Naive Bayes, Decision Trees, Random Forests, Logistic Regression, and XGBoost, are used to predict heart disease. Out of the 76 features in the Heart Disease Kaggle dataset, only 14 are taken into account for the prediction. Gender, kind of chest discomfort, fasting blood pressure, serum cholesterol, exang, etc. are examples of patient characteristics. The algorithm that offers the highest accuracy is selected after each one's accuracy has been compared. Accuracy, confusion matrix, precision, recall, and f1-score are some of the evaluation criteria. The ratio of (TP+TN) to (TP+FP+FN+TN) represents accuracy [15]. Confusion matrix offers a matrix output that shows the system performance as a whole. Accurate cardiac disease prediction

depends on these evaluation metrics. The research demonstrates that the maximum accuracy, or 80%, is provided by the random forest classifier. In order to modify their lifestyle and lower complications, high-risk patients must receive an early diagnosis of heart disease. Heart disease can be detected and treated by the medical community and patients with the application of the right technology, such as machine learning.

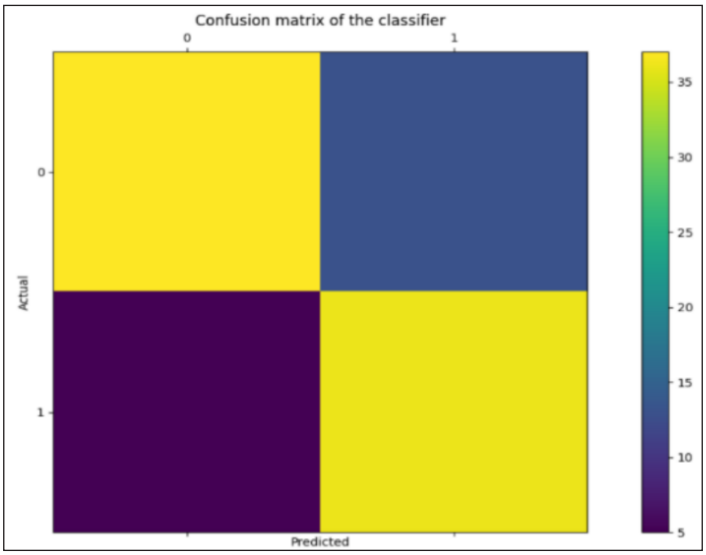

Fig. 3: Confusion Matrix

Where

TP: True Positive

FP: False Positive

FN: False Negative

TN: True Negative

Correlation Matrix: The correlation matrix, which depicts the interdependence of several attributes, is used in machine learning to choose features.

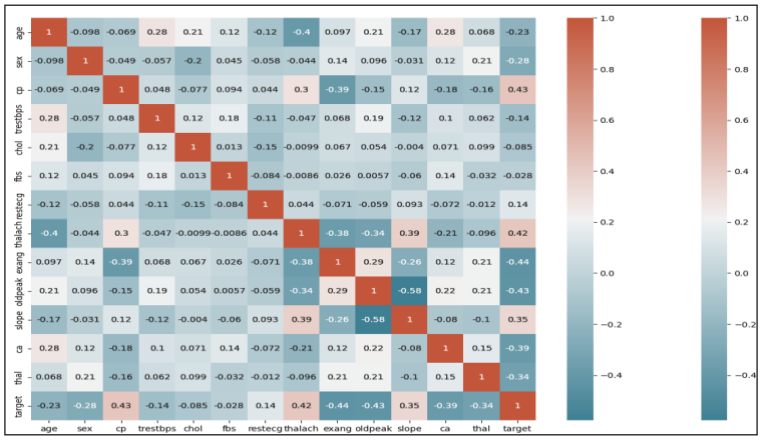

Fig. 4: Correlation Matrix

Pairplots: Pairplots are also a great way to immediately see the correlations between all variables. With so many features, it can be difficult to see each one. So instead a pairplot is made with only continuous features.

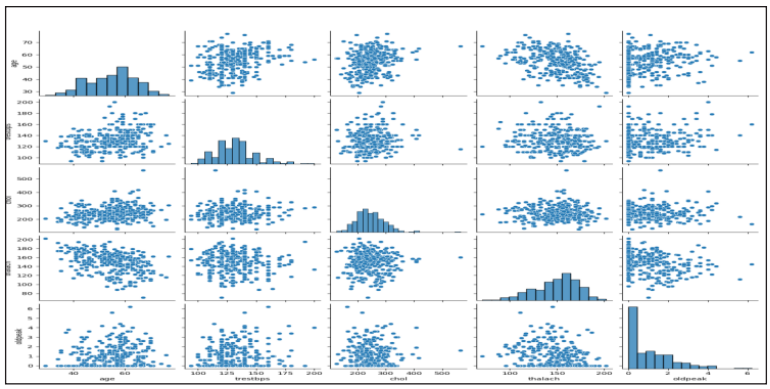

Fig. 5: Pairplot

Precision - It is the proportion of correctly positive results to all positive results that the algorithm correctly predicted.

Recall - This is the proportion of correctly positive results to all positively anticipated results by the system.

F1-Score - The range of the F1- score, which evaluates test accuracy is 0 to 1, and it is the harmonic mean of Precision and Recall.

VII. Result

After training and testing different machine learning algorithms, it was found that Random Forest had a higher accuracy compared to the other models. The accuracy was calculated using the confusion matrix for each algorithm, which provided the count of TP, TN, FP, and FN. By using the accuracy equation, the accuracy value was determined to be 80%. Based on this, it can be concluded that Random Forest is the best model among the ones tested. A comparison of the accuracy of different models is presented below.

The highest accuracy is given by the Random Forest algorithm.

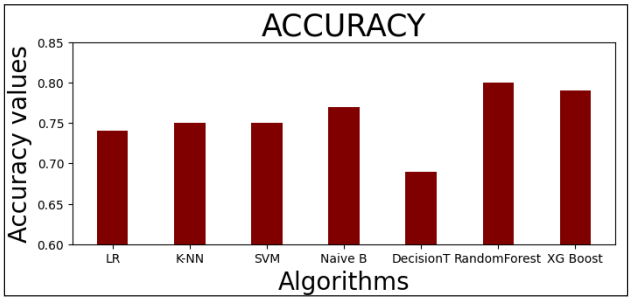

Fig. 6: Accuracy Comparison of Algorithms

VIII. Conclusion and Future Scope

The detection of heart disease is crucial since it is a leading cause of death in India and worldwide. Employing advanced technology, such as machine learning, in early detection of heart disease can have a significant impact on society. Early diagnosis can assist high-risk patients in deciding whether to adjust their lifestyles, which can minimize issues and represent a significant breakthrough in the medical field. As the number of individuals developing cardiac diseases increases every year, early detection and treatment are critical. The appropriate use of technology can benefit both patients and the medical community in this area. This study uses seven machine learning algorithms, including SVM, Decision Tree, Random Forest, Naive Bayes, Logistic Regression, KNN, and XG Boosting, to evaluate performance. The dataset comprises 76 features that contribute to heart disease in individuals, and 14 important characteristics are chosen to assess the system. When all features are considered, the author receives less efficiency from the system. To increase efficiency, attribute selection is used, and n characteristics must be chosen to evaluate the model that provides greater accuracy. Eliminating some dataset features that have virtually equal correlations improves efficiency. If all attributes in the dataset are considered, efficiency decreases significantly. The study compares the accuracy of each of the seven machine learning techniques to create a prediction model. Evaluation metrics such as the confusion matrix, accuracy, precision, recall, and f1-score are used to effectively predict the disease. After comparing all seven methods, the random forest classifier provides the highest accuracy of 80%.

References

[1] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in *2017 IEEE Symposium on Computers and Communications (ISCC),* Heraklion, 2017, pp. 204-207, doi: 10.1109/ISCC.2017.8024530.

[2] S. Dhar, K. Roy, T. Dey, P. Datta, and A. Biswas, "A hybrid machine learning approach for prediction of heart diseases," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, 2018, pp. 1-6, doi: 10.1109/CCAA.2018.8777531.

[3] C. Raju, E. Philipsy, S. Chacko, L. Padma Suresh, and S. Deepa Rajan, "A survey on predicting heart disease using data mining techniques," in *2018 Conference on Emerging Devices and Smart Systems (ICEDSS)*, Tiruchengode, India, 2018, pp. 253-255, doi: 10.1109/ICEDSS.2018.8544333.

[4] Y.-X. Wang, Q.-H. Sun, T.-Y. Chien, and P.-C. Huang, "Using data mining and machine learning techniques for system design space exploration and automatized optimization," in *Proceedings of the 2017 IEEE International Conference on Applied System Innovation*, vol. 15, pp. 1079-1082, 2017.

[5] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data mining and analytics in the process industry: The role of machine learning," *IEEE Access*, vol. 5, pp. 20590-20616, 2017.

[6] Y. Zhang, R. Fogoros, J. Thompson, B. H. Kenknight, M. J. Pederson, A. Patangay, and S. T. Mazar, U.S. Patent No. 8,014,863. Washington, DC: U.S. Patent and Trademark Office, 2011.

[7] https://en.wikipedia.org/wiki/Logistic_regression

[8] https://en.wikipedia.org/wiki/Bayes27_theorem

[9] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[10] https://en.wikipedia.org/wiki/Support_vector_machine

[11] https://en.wikipedia.org/wiki/Knearest_neighbors_algorithm

[12] https://en.wikipedia.org/wiki/Decision_tree_learning

[13] https://towardsdatascience.com/understanding-random-forest58381e0602d2

[14] https://en.wikipedia.org/wiki/XGBoost

[15] https://blog.paperspace.com/deep-learning-metrics-precision-recall-accuracy/

[16] https://www.who.int/hrh/links/en/

[17] https://en.wikipedia.org/wiki/Machine_learning