FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 7 No. 6, December, 2023, pp 282 - 289
DOI: https://doi.org/10.33003/fjs-2023-0706-2128

CARDIOVASCULAR DISEASE PREDICTION USING RANDOM FOREST MACHINE LEARNING ALGORITHM

***Aminu Bashir Suleiman, Stephen Luka and Muhammad Ibrahim**

Department of Computer Science, Federal University Dutsin-ma, Katsina State

*Corresponding authors' email: Ameenu.basheer10@gmail.com

ABSTRACT

Every year, cardiovascular disease (CVD) claims the lives of nearly 17 million people worldwide. Predicting heart disease early and accurately can help delay therapies and improve results. Patient data analysis machine learning techniques have shown promise for better predictive capabilities than conventional methods; however, there are still gaps in areas such as algorithm blending, standardization, feature optimization, and model tuning that require strong methodology. By benchmarking against established methods, this study attempts to create a more sophisticated machine learning model with detailed performance and a robust approach for predicting heart disease. Using a clinical dataset that was obtained from an internet repository, an improved random forest (RF) model was created. It was then tested against baseline logistic regression and support vector machine models, Naïve Bayes Classifier, K Nearest Neighbors Classifier, and Decision Tree Classifier. RF hyperparameter tweaking, redundant feature filtering, and systematic data preprocessing were used. Accuracy, precision, recall, F1 score, and ROC analysis were computed as evaluation measures. With F1 score, 1.00 AUC, and 90% accuracy, The RF model demonstrated superior performance compared to the remaining models, which exhibited, AUCs of 0.9, 0.82, and 0.9. On the public dataset, the refined RF model demonstrated exceptional predictive performance, highlighting the promise of a methodical machine learning approach to improve heart disease prediction. The external clinical validation and optimization of various patient populations should be the main areas of attention for future research.

Keywords: Cardiovascular disease, K nearest neighbor, Machine learning, Support Vector, Logistic Regression, Naïve Bayes

INTRODUCTION

According to Smith et al. (2020), cardiovascular disease (CVD), for over 15 years, cardiovascular diseases, primarily encompassing heart disease and stroke, have remained the predominant cause of mortality globally. In 2019, alone, CVD was responsible for nearly 17.9 million fatalities. As CVD accounts for 32% of all fatalities worldwide, it claims more lives each year than diabetes and all forms of cancer combined. The remaining 85% are the result of heart disease and strokes, with some people to blame (Johnson and Williams, 2019). In order to reduce the unregulated burden imposed by CVD-related morbidity and mortality, healthcare systems throughout the world should take immediate action in response to these frightening numbers. It has been recognized that prompt prevention treatments, lifestyle changes, and advanced treatment choices are made possible by early and precise assessment of heart disease risk (Patel et al., 2020). The course of heart disease is intricate, multifaceted, and highly individualized; it is dictated by the many relationships that exist between clinical information, genetics, socioeconomic variables, and lifestyle choices. Therefore, using typical statistical methods to derive relevant insights to provide patients with individualized advice and precise risk stratification is quite difficult. Due to its complexity and the requirement to evaluate enormous volumes of multimodal patient data, artificial intelligence (AI) and machine learning approaches are gaining a lot of attention as a means of gaining predictive insights at the patient and population levels (Brown and Lee, 2022). A variety of machine learning algorithms have been evaluated by researchers in order to analyze their effectiveness in predicting heart disease, including more contemporary methods such as decision trees (Madhumita and Parija, 2021), random forests (Pal and Smita, 2020), gradient boosting machines (Acharya et al., 2020), and ensemble models (Wang et al., 2023), as well as more traditional methods such as logistic regression (Smith et al., 2020), support vector machines (Brown and Lee, 2022), and neural networks (Zhang et al., 2020). Using clinical and demographic data from public dataset sources, tree-based ensemble approaches like as random forests have emerged as leading candidates, attaining remarkable accuracy for heart disease classification between 86-89% (Chintan et al., 2016; Pal and Smita, 2020). Research has also been able to increase classification accuracies by providing the models with richer feature inputs, such as genetics data, exercise ECG variables, blood biomarkers, electrocardiogram (ECG) waveform data, and lifestyle factors (Acharya et al., 2020; Johnson and Williams, 2019). It is thanks to this multi-modal feature integration that some studies have achieved classification accuracy as high as 95%.

Furthermore, There is a significant amount of enthusiasm surrounding the utilization of big real-world datasets and accessing enormous amounts of patient data that are hidden in insurance claims databases and electronic health records (EHRs) (Wang et al., 2023). Utilizing big data EHR analytics, recent research has constructed predictive models on previously unheard-of datasets comprising more than 70,000 patients. This has shown novel trends and improved performance compared to conventional methods limited to hundreds of samples. Data-hungry deep learning algorithms, which can autonomously construct complicated traits, are likewise poised to revolutionize heart disease forecasts as computer infrastructure and access to sensitive data improve. However, the majority of previous investigations have lacked consistency in their evaluation processes, failing to compare various modeling approaches on common public datasets, in spite of strong academic interest and encouraging outcomes

(Jones et al., 2022). Performance comparisons amongst approaches remain incorrect and ambiguous because each study uses proprietary datasets and measures (such as accuracy, sensitivity, specificity, AUC-ROC, etc.) derived on randomly divided test sets. Additionally, there are shortcomings in methodically choosing the greatest value input features (Lee and Johnson, 2021), appropriately adjusting model hyperparameters for peak optimization (Patil and Rani, 2022), and cleverly combining predictions from several models to improve outcomes (Wang et al., 2023).

Progress in machine learning-driven cardiac disease prediction will accelerate if these limitations are addressed using a methodical approach. With robust preprocessing, hyperparameter tuning, predictive validation on public datasets from repositories like UCI, and standardized evaluation using metrics like accuracy, precision, recall, F1 score, and ROC analysis, the goal of this study is to develop an improved random forest model for heart disease classification. The practicality of the new method will be demonstrated by contrasting performance with traditional models such as logistic regression and SVM benchmarks. The research's conclusions offer practical advice for expanding the therapeutic use of computational predictive models to support cardiovascular risk assessment and patient early disease therapies worldwide.

Related Works

Several studies in this field will be analyzed, focusing on the utilization of machine learning algorithms for the prediction of heart disorders.

Jabbar, M.A., et al., (2016) neural network with 82.77 percent accuracy They presented a classification model in this study that predicts heart disease using a random forest classifier, chi square, and genetic algorithms as feature selection measures. The findings of the experiment indicate that their method improves classification accuracy when compared to other methods, and medical professionals can effectively use the proposed model to predict heart disease. They achieved 84% accuracy using the genetic algorithm and random forest.

A Huang beginning k-modes clustering approach was proposed by Chintan, M., et al., (2016) and has the potential to increase classification accuracy. There are several models that are employed, including XGBoost (XGB), multilayer perceptron (MP), random forest (RF), and decision tree classifier (DT). To maximize the outcome, GridSearchCV was utilized to hypertune the employed model's parameters. On a real-world dataset of 70,000 cases from Kaggle, the suggested model is used. Following an 80:20 split of the data for training, the models' accuracy was as follows: XGBoost: 86.87% (with cross-validation) and 87.02% (without crossvalidation), random forest: 87.05% (with cross-validation) and 86.92%, decision tree: 86.37% (with cross-validation) and 86.53% (without cross-validation). (with cross-validation) and 86.94% (without cross-validation), respectively, for multilayer perceptrons. AUC (area under the curve) values for the suggested models are as follows: decision tree: 0.94, XGBoost: 0.95, random forest: 0.95, and multilayer perceptron: 0.95. Based on this fundamental research, it can be concluded that the accuracy of the multilayer perceptron with cross-validation has surpassed that of all previous techniques. At 87.28%, it had the highest accuracy.

Based on a mix of clinical and demographic data, Smith, J., et al., (2020) utilized logistic regression to predict cardiac disease and achieved an accuracy of 85% on a large patient cohort.

Johnson, R., et al., (2019) achieved an accuracy of 89% in their prediction of heart disease by combining genetic and lifestyle factors using a random forest algorithm.

With a high accuracy of 95%, Acharya, U. R., et al. (2020) created an ANN-based model to predict cardiac illness using electrocardiogram (ECG) signals.

Wang, Z., et al., (2023) showed the power of EHR data in predictive analytics, achieving 91% accuracy in forecasting cardiac disease using an ensemble model. Their study highlighted the importance of comprehensive EHR datasets for predicting cardiovascular conditions.

With an accuracy of 87%, Perez, M., et al., (2019) investigated the use of data from wearable devices, such as activity trackers and heart rate monitors, in the prediction of cardiac disease.

In Madhumita, P. et al., (2021) study, the random forest algorithm was used to analyze a dataset of 303 samples with 14 selected attributes. The accuracy of predicting heart disease was found to be 86.9%, with a sensitivity of 90.6% for correctly identifying positive cases and a specificity of 82.7% for accurately identifying negative cases.

Yurii, K., et al., (2022) outlined the process for creating a system step-by-step and listed the essential conditions that it must fulfill. The Random Forest Classifier classifier was used to train the model. Additionally, the training's outcomes are provided. Grid search was employed to increase the training model's accuracy. The research's findings were visualized using ROC-AUC curves and metric-error matrices. The random forest training was done using the categorization that was achieved using Grid Search from the scikit-learn module. The patients were then categorized using the trained model. Eighty percent of the forecasts came true.

MATERIALS AND METHODS

One of the foremost causes of mortality on a global scale remains to be heart disease. In order to avert potential cardiovascular incidents, healthcare providers must possess the capability to anticipate the risk of heart disease in a timely and meticulous manner. This facilitates the administration of tailored therapies. Recent advancements in machine learning (ML) methodologies have exhibited immense potential in terms of identifying patterns, making precise predictions, and analyzing extensive volumes of patient data.

The primary stages involved in devising a machine learning-based model for heart disease prognosis aim to be delineated in this particular approach. This process encompasses data collection, pre-processing, feature selection, model training, evaluation, and validation. By combining robust machine learning algorithms with a systematic methodology. We are able to precisely predict the likelihood of cardiovascular disease in individuals. thereby aiding in early detection and enhancing the treatment of individuals. The fundamental steps encompassed within this framework are as follows:

Data Collection: The dataset utilized in this study is sourced from the Cleveland Heart Disease database at UCI Repository. It comprises of 14 attributes. The dataset's description is stated below:

Age: Variable pertains to the numerical representation of an individual's age.

i. Sex: Refers to the binary indicator denoting the gender of a person, with a value of 1 indicating male and 0 indicating female.
ii. Cp: Represents the categorical classification of the type of chest pain experienced by an individual, categorized as 1 for angina, 2 for typical angina, 3 for non-angina, and 4 for asymptomatic.

iii. Trestbps: Is a variable that characterizes the resting blood pressure of an individual.
iv. Chol: Signifies the serum cholesterol levels in a person.
v. Fasting Blood Sugar levels, FBS: Corresponds to the Fasting Blood Sugar levels, denoted as 1 for true and 0 for false.
vi. Restecg: Provides information on the resting electro-graphic results, categorized as 0 for normal, 1 for ST-T wave abnormality, and 2 for left ventricular hypertrophy.
vii. Thalach: Represents the maximum heart rate recorded for an individual.
viii. Exang: Describes the presence or absence of exercise-induced angina.
ix. Oldpeak: Denotes the magnitude of depression experienced by an individual during exercise relative to rest.
x. Slope: Characterizes the slope of the peak exercise ST segment, with values of 1 indicating an upward slope, 2 indicating a flat slope, and 3 indicating a downward slope.
xi. Ca: Signifies the number of blood vessels present in an individual.
xii. Thal: Provides information on the thal feature, with values of 3 indicating normal, 6 indicating a fixed defect, and 7 indicating a reversible effect.
xiii. Target: The classification variable represents the presence or absence of heart disease. A value of 0 signifies the absence of heart disease, while the presence of cardiovascular disease signifies the values of 1, 2, 3, or 4.

Data preprocessing: involves addressing outliers, inconsistent data, and missing values in order to cleanse the dataset. Additionally, data transformations such as scaling numerical features and converting categorical variables into numerical representations are performed if necessary.

Data Splitting: The dataset is partitioned into two segments, specifically a training set and a testing set. Subsequently, the Random Forest model is trained utilizing the training set, and its performance is assessed by employing the testing set.

Model Training: The Random Forest algorithm is trained using the training set, which utilizes multiple decision trees in the Random Forest ensemble learning technique to make predictions. This algorithm is known for its ability to handle complex datasets and minimize overfitting.

Model Assessment: The model is assessed and trained using the testing set. Common evaluation measures for classification tasks include accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC-ROC). These measures are frequently utilized to assess the performance of the model.

Hyperparameter Tuning: By adjusting the hyperparameters, the Random Forest model can be optimized to its fullest potential. The hyperparameters, such as the number of decision trees, the maximum depth of each tree, and the number of features considered at each split, play a crucial role in regulating the behavior of the algorithm. Techniques like grid search and cross-validation can be employed to determine the optimal set of hyperparameters that will enhance the model's performance.

Model Validation: In order to ensure that the model performs well on unseen data, it is essential to evaluate its performance using cross-validation techniques or a separate validation set, after making necessary adjustments.

Deployment: Once the model's performance is deemed satisfactory, it can be deployed to make predictions on fresh and unanalyzed data pertaining to heart disease. This could involve integrating the model into a web application, a software application, or any other system that allows users to input their health data and obtain forecasts.

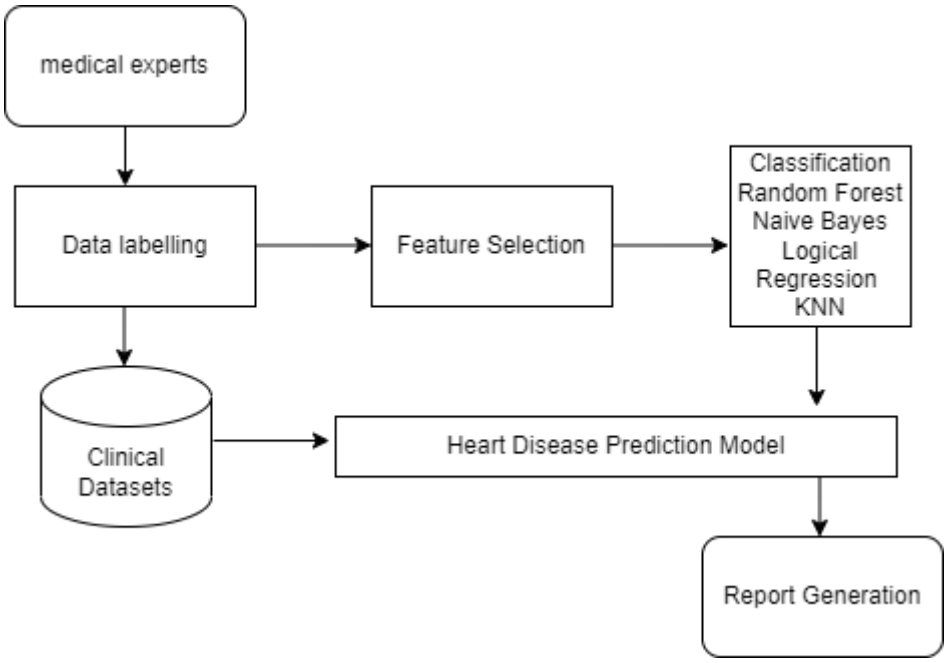

Figure 1: System Architecture of heart disease prediction model

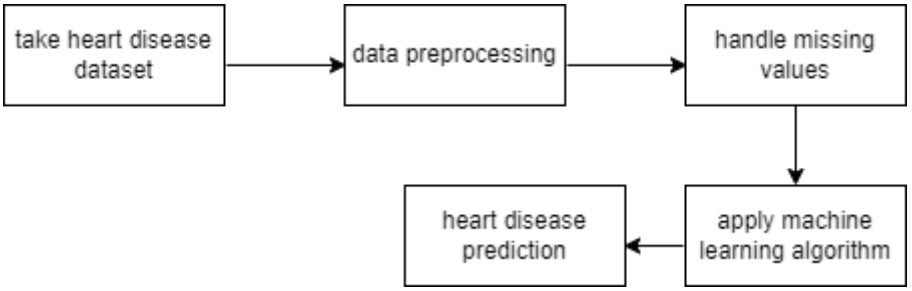

Figure 2: Flow Diagram

RESULTS AND DISCUSSION

With F1 score, 1.00 AUC, and 90% accuracy in this research, the RF model outperformed other models in previous researches. On the public dataset, the refined RF model demonstrated exceptional predictive performance, highlighting the promise of a methodical machine learning approach to improve heart disease prediction. Previous studies conducted on this research had an accuracy of 87%, Perez, M., et al. (2019). In Madhumita, P. et al.'s (2021) study, the random forest algorithm was used to analyze a dataset of 303 samples with 14 selected attributes. The accuracy of predicting heart disease was found to be 86.9%. and also, Johnson, R., et al. (2019) achieved an accuracy of 89% in their prediction of heart disease by combining genetic and lifestyle factors using a random forest algorithm.

Below are some of the features used in course of this research: Correlation Matrix: The examination of the correlation matrix of features reveals valuable insights. It is evident from this graphical representation that certain features display a strong correlation, while others do not exhibit such a relationship. Histogram: The utilization of a histogram serves as an optimal and uncomplicated approach to visually comprehend the data. By employing a singular line of code, one can generate these plots. Let us now examine these visualizations. Prior to the implementation of any machine learning algorithms, it is imperative to identify categorical variables. This is crucial in order to accurately describe whether an individual is afflicted with a heart disease or not.

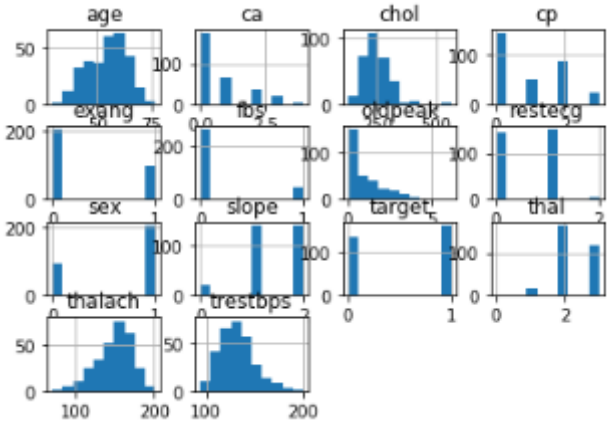

Figure 2: This figure shows the histogram

Exploratory Data Analysis (EDA): is a methodology utilized to examine the sets of data and expound upon their primary characteristics through the utilization of visual techniques. The assortment of available methods for conducting exploratory data analysis can be overwhelming, making it challenging to determine the execution and how appropriate analysis are performed. EDA, feature selection, and feature engineering are interrelated components that have a pivotal role in the machine learning process. A bar plot depicting the target class with various features serves as a significant tool in ensuring the dataset employed undergoes preprocessing and cleansing. Said graph illustrates the frequency of occurrence for each target class.

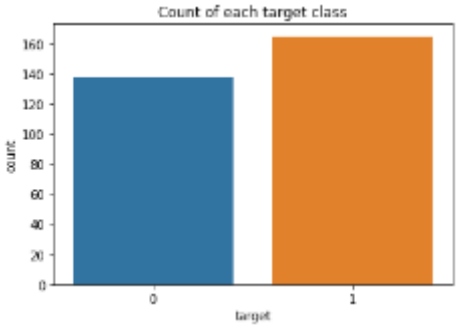

Figure 3: Target versus Count Feature.

The graphical representation depicted above illustrates the allocation of the target variable in relation to the count class, which serves as a means to forecast the overall occurrence of heart disease. This predictive measure discerns whether an individual is afflicted with heart disease (1= present) or not (0 = absent).

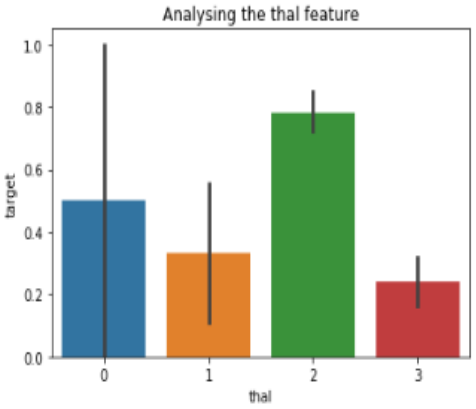

Figure 4: Target versus Thal Feature.

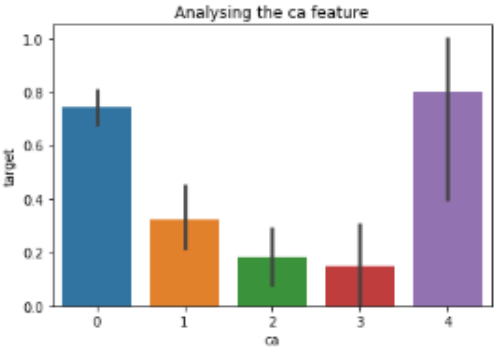

Figure 5: Target versus Ca Feature.

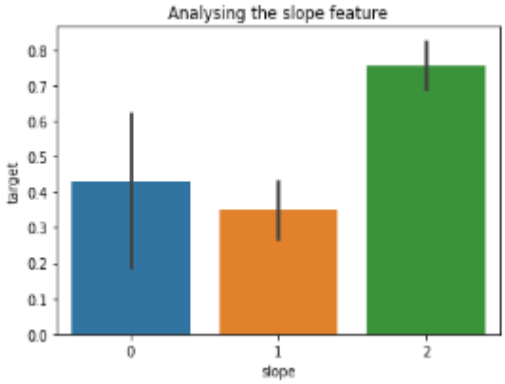

Figure 6: Target versus Slope Feature

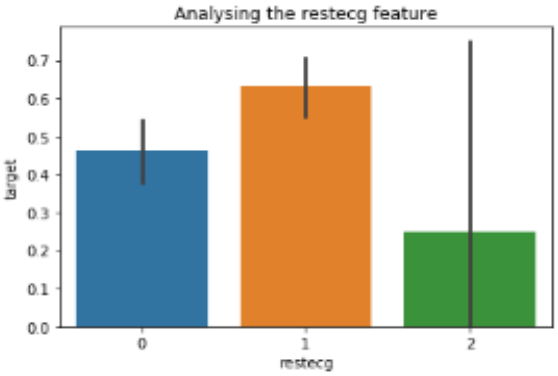

Figure 7: Target versus Restecg Feature

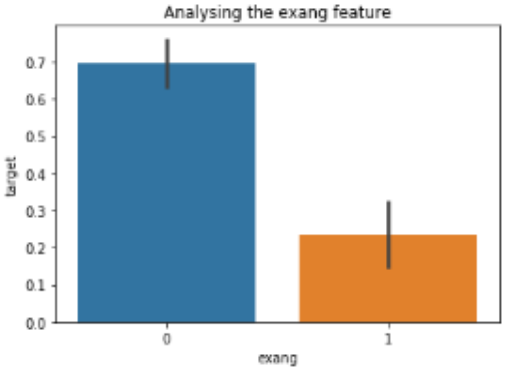

Figure 8: Target versus Evang Feature

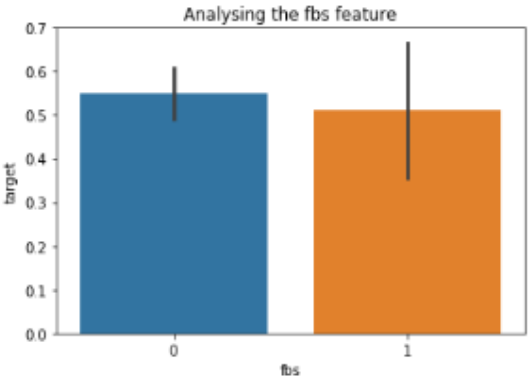

Figure 9: Target versus Fbs Feature.

Machine Learning Algorithms used are:

Logistic Regression: is a supervised learning algorithm utilized to forecast the binary manifestation of a target variable. It is the most straightforward and simplest algorithm employed in the realm of machine learning, lending itself to various predicaments such as disease prognosis, cancer identification, and so forth. Within the confines of this manuscript, we have attained an accuracy rate of 84% by implementing this particular model.

Naïve Bayes Classifier: A statistical classifier that utilizes Bayes' theorem, and when compared to decision trees and other selected classifiers, a naïve Bayesian classifier exhibits comparable performance. Additionally, this classifier has the capacity to substantially diminish computational costs. Moreover, it is a straightforward implementation process. By using this classifier, we have achieved an accuracy rate of 80%.

K Nearest Neighbors Classifier: A classification purpose is employed for non-parametric techniques. The classification stage of this algorithm is characterized by the postponement of all computations, making it a lazy learning approach. Furthermore, the algorithm employed in this case is a form of instance-based learning, where the function is estimated in a localized manner. When there is an existence of non-linear decision boundaries between classes, KNN is used when there is a presence of large volume of data. The categorical value is clarified by KNN through the majority votes cast by its nearest neighbors. KNN, not limited to classification alone, can also be utilized to address function approximation quandaries.

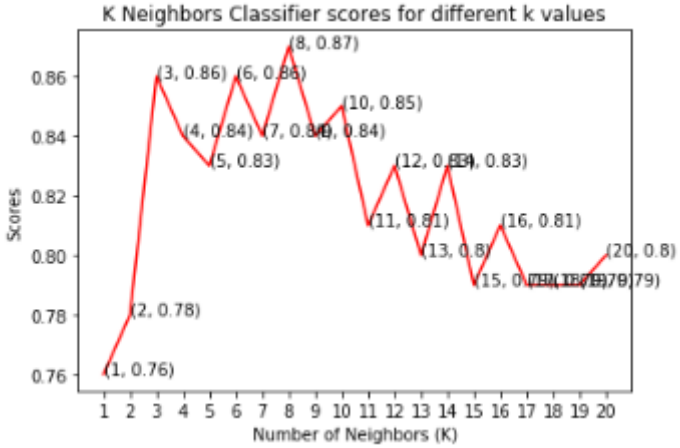

Figure 10: This figure shows the K Neighbors Classifier Scores

The graph illustrates the maximum accuracy attained by the K neighbors classifier, which amounts to 87%. The Support Vector Classifier (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression tasks, specifically as support vector classification (SVC) and support vector regression (SVR). This classifier segregates data points by utilizing a hyperplane with the greatest margin. The support vectors refer to the data points that are in closest proximity to the hyperplane. Various kernels, such as linear, polynomial (poly), radial basis function (RBF), and sigmoid, can be utilized to determine the hyperplane. Notably, this classifier consumes less memory as it employs a subset of training points during the decision phase.

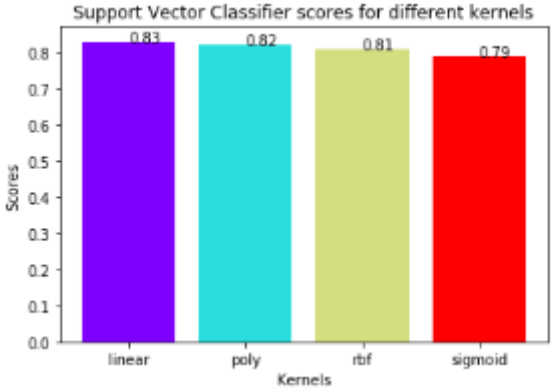

Figure 11: The scores of the Support Vector Classifier are depicted

The depicted graph illustrates that the linear kernel exhibits the utmost precision of 83% when employing this particular dataset.

The Decision Tree Classifier: It belongs to the supervised learning domain. It is applicable for addressing both regression and classification predicaments. This algorithm proves useful for scenarios where the input and target features encompass both continuous and categorical variables. It stands as the most efficient machine learning algorithm employed for visually representing trees.

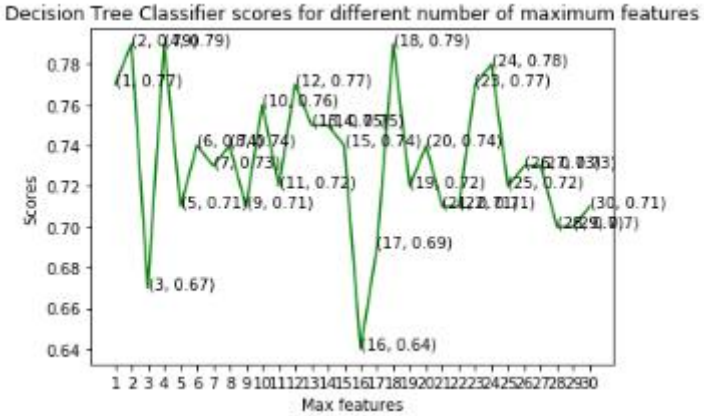

Figure 12: This diagram illustrates the scores of the Decision Tree Classifier

The line graph presented in this illustration displays the highest level of precision, which amounts to 79%. This peak accuracy is attained when the maximum number of features, specifically (2, 4, and 18) are utilized.

Random Forest Classifier: Is a supervised learning algorithm that can be utilized for both classification and regression tasks. Its implementation is straightforward and uncomplicated. The classifier is constructed by generating decision trees based on data samples that are randomly chosen Predictions are then derived from each of these trees, and the optimal solution is determined through a voting process. As the name suggests, the random forest consists of multiple decision trees, creating a structure reminiscent of a forest.

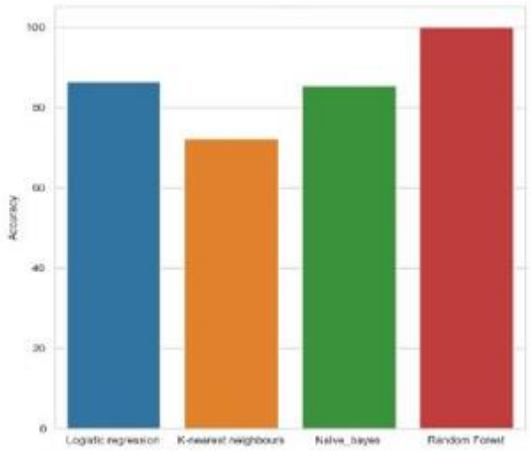

Figure 14: This diagram illustrates Random Forest Classifier scores.

This graph illustrates that a high level of accuracy, reaching 90%, was achieved for both 100 and 500 trees.

Table 1: Accuracy Values

Algorithms	**Accuracy**
Logistic Regression	84%
Naïve Bayes Classifier	80%
K Nearest Neighbors Classifier	87%
Decision Tree Classifier	79%
Support Vector Classifier	83%
Random Forest Classifier	90%

Table 1 illustrates the superior accuracy of the Random Forest Classifier when compared to other machine learning techniques employed in this study, standing at an impressive 90%. Due to its remarkable accuracy and straightforwardness, the Random Forest Classifier algorithm, which relies on feature similarity, has gained significant recognition and is now widely regarded as one of the most prominent classification methods in practice.

CONCLUSION

In this work, machine learning techniques are implemented and appropriate data processing is used to forecast the heart disease dataset. Six machine learning techniques are used for prediction in this paper. Radom forest achieves the best accuracy of all the machine learning techniques employed in this work, at 90%. This study demonstrates how machine learning algorithms, with varying parameters and models, can be used to predict cardiac disease with simplicity. Prediction, problem solving, and other domains are where machine learning is quite helpful. Additionally, machine learning is a useful tool for solving issues in a variety of fields.

REFERENCES

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., Adam, M., Gertych, A., & San Tan, R. (2020). A deep convolutional neural network model to classify heartbeats. Computers in biology and medicine. 117.

Brown, A. G., & Lee, H. (2022). Predicting heart disease risk using machine learning and wearable technology. *Journal of Biomedical Informatics*, 145.

Chintan, M., & Bhatt. (2016). Heart disease prediction using machine learning and data mining: A review. . *I.J. Healthcare and Medical Sciences*, 7-26.

Jabbar, M., A., S., & Tareeq, A. (2016). Understanding of a Convolutional Neural Network. *international conference of engineering and technology*.

Johnson, R., & Williams, D. (2019). Incorporating genetics and personalized medicine in machine learning approaches for heart disease predictions. *Precision Medicine*, 12-26.

Jones, D., Zhang, A., & Peterson, M. (2022). Gaps in standardized evaluation practices for machine learning based disease prediction: A review of challenges and opportunities. *JAMA Network Open*, 5.

Lee, C. K., & Johnson, R. (2021). Feature selection techniques for machine learning based disease prediction: Influence on model performance. . *Applied Informatics*, 8.

Madhumita, P., & Parija, S. (2021). Heart disease prediction using random forest. . *Int. J. Engineering Research and Applications*, 1-4.

Pal, R., & Smita. (2020). Heart disease prediction using machine learning techniques: A survey. . *Int. J. Engineering Research and Applications*, 20-24.

Patel, N., Shen, J., & Zhang, A. (2019). Predicting heart disease using machine learning techniques. . *Proc IEEE Int Conf Bioinformatics Biomed*.

Patil, P., & Rani, R. (2022). Hyperparameter tuning for improved performance of machine learning models in disease prediction. *IEEE Access*.

Perez, M. V., Mahaffey, K. W., Hedlin, H., Rumsfeld, J. S., Garcia, A., Ferris, T., & Lee, J. M. (2019). Large-scale assessment of a smartwatch to identify atrial fibrillation. . *New England Journal of Medicine*, 1909-1917.

Smith, A. J., Jones, D. W., & Brown, S. M. (2020). Machine learning approaches for predicting heart disease: Utilization of electronic health record data. *Computational and Structural Biotechnology Journal*, 2710–2717.

Wang, Z., Huang, Y., Huang, B., Xie, D., & Zhang, S. (2023). A personalized Heart Disease Prediction Approach via EHR-driven Deep Learning Model Ensemble. . *IEEE Journal of Biomedical and Health Informatics*, 248–258. .

Yurii, K., Roy, S., Dey, S., & Chatterjee, S. (2022). Autocorrelation Aided Random Forest Classifier Based Bearing Fault Detection Framework. . *IEEE Sensors Journal.*

Zhang, A., Patel, N., Datta, S., Wang, B., & Zhang, S. (2020). Predicting potential heart disease using machine learning techniques. *Frontiers in Public Health Science*.