WWJMRD 2022; 8(05): 27-32
www.wwjmrd.com
International Journal
Peer Reviewed Journal
Refereed Journal
Indexed Journal
Impact Factor SJIF 2017: 5.182 2018: 5.51, (ISI) 2020-2021: 1.361
E-ISSN: 2454-6615

Alpha Alimamy Kamara
Faculty of Information Communication and Technology, University of Creative Technology, Freetown, Sierra Leone.

Abdul Joseph Fofanah
Department of Mathematics and Computer Science, Milton Margai Technical University, Freetown, Western Area Rural, Sierra Leone.

Correspondence:
Alpha Alimamy Kamara
Faculty of Information Communication and Technology, University of Creative Technology, Freetown, Sierra Leone.

Heart Disease Prediction Support System using Machine Learning Approaches

Alpha Alimamy Kamara, Abdul Joseph Fofanah

Abstract
Heart disease (cardiovascular disease) is a common source of death in the world and a major health threat. According to WHO research, a cardiovascular disease caused 17.9 million deaths worldwide in 2017. Unfortunately, the mortality and morbidity of cardiovascular disease (heart disease) are increasing year by year, especially in developing countries. According to reports, almost 80% of heat-related deaths occur in middle-income and low-income countries. In addition, in low-income countries, the age of these deaths is younger than in high-income countries. The poor economic transition in developing countries has led to environmental changes and unhealthy lifestyles; in addition, the aging of the population may increase the risk factors of cardiovascular disease and the incidence of cardiovascular disease (heart disease). The patients and the whole society were hit hard by heart disease. Therefore, strategies for improving the diagnosis and treatment of heart disease are needed in the future.
Machine learning may now solve this problem. This study used four different machine learning algorithms to develop and implement a predictive model for heart (cardiovascular) disease detection. The findings of this study show that all developed models, including random forests, decision trees, neural networks, and XGBoost, have high classification accuracy and are similar in predicting heart disease cases. However, the comparison based on the true positive rate shows that the random forest model performs slightly better in predicting heart disease, with a classification accuracy rate of 94.96 %.

Keywords: Cardiovascular Disease, Decision Tree, Machine Learning, Random Forest, Neural Network, XGBoost, Heart Disease.

1. Introduction
Heart disease refers to a variety of factors that affect the normal operation of the bloodstream, which involves the heart and also blood vessels. There are various forms of heart diseases, such as cardiovascular disease, which impact the heart and as well as blood vessels, causing blood to not be drained and distributed properly in the body. Since cholesterol and fat are stored within the walls of the blood vessels that provide blood to the heart in coronary heart disease, the heart does not receive the volume of blood it needs.
In the May 2017 WHO report, an estimated 17.9 million people die yearly from cardiovascular diseases (CVDs) worldwide, more so from heart and stroke. This constitutes 31 % of the year's global deaths [1]. The prevalent symptoms and health behaviors that lead at an early stage to the development of cardiovascular diseases must be established. However, different studies are performed to assess cardiovascular diseases. Blood pressure, blood sugar, cholesterol, etc. are some of these measures. A category of diseases covering the blood vessels (arteries and veins) and the heart are cardiovascular diseases. There is a high and troubling rate of cardiovascular disease in many countries around the world at present. Among men and women over sixty-five years of age, cardiovascular disorders have been a principal source of death worldwide. Cardiovascular disease is presently seen as a "second epidemic" in many nations, replacing infectious diseases as the primary cause of death
Several classification methods such as Decision Trees, Random Forest, and Support Vector Machine (SVM) etc are used to predict the presence of cardiovascular diseases. They are also used to identify the relevant factors that mostly contribute to the causes of cardiovascular

(heart) diseases, different classification methods are used, such as Decision Trees and Neural Networks. They are also used to classify the variables that contribute most to the cause of cardiovascular (heart) diseases

2. Related Work

[2] Developed an architecture with the competence to predict possible diabetes patients. Their key research objective was to develop a web app based on an advanced machine learning algorithm's higher prediction accuracy. They used a standard dataset that is proficient in predicting the complications of diabetes using a diagnostics manner. In their research, Artificial Neural Network showed a high improvement with an accuracy of 82.35% prediction rate which drives the design of a User-friendly Web Application for Diabetes Predictions.

[3] Proposed a machine learning ensemble technique that merged several machine learning methods to construct a more reliable and stable model for predicting the menace of heart (cardiovascular) disease. Their Ensemble model had a 90% accuracy rate, which was higher than the accuracy of every individual classifier.

Their model can be used to help physicians evaluate patient cases in order to support their diagnoses and minimize human.

[4] Introduced Naive Bayes and Decision Tree classification algorithms that were implemented to the dataset to estimate the probability of a patient having heart disease. They compared the two algorithms on the same dataset in order to decide which was the most accurate. In terms of their performance, the Decision tree algorithm correctly predicted heart disease patients 91% of the time, while the Nave Bayes classifier correctly predicted 87% heart disease, patients. Their project concluded that the Decision Tree Classification algorithm is the safest and best for working with medical datasets.

[5] Developed a hybrid algorithm-based prediction of heart (cardiovascular) disease, which they followed up with appropriate patient advice. Among the various individual algorithms used in their model, the hybrid models were the most successful and produced the best accuracy, specificity, and sensitivity. The Body Mass Index (BMI), family history, waist circumference, and physical activities in everyday life were all increased in their hybrid model. With 82.11% accuracy and 91.47% sensitivity, this Hybrid model even outperforms the competition. This study points to a new research direction in which these ideas can be applied to smart devices and data science to revolutionize heart disease diagnosis and care. This initiative has the possibility to promote awareness about heart disease's complexities.

Data classification Algorithms

Predictive modeling is an approach to building a model capable of making predictions. This model includes a machine learning algorithm that enables data-driven models to learn specific information from observed data in a training data set to make those predictions. In this project, the proposed model has a particular target used to predict output for new use cases, which determines whether heart disease is present or not. Hence supervised learning classification algorithms are used to train and test the model. The following Machine Learning algorithms (Random Forest, Decision Tree, XGBoost, and Neural Networks), are used to build the proposed model

Random Forest

Random forests are a category of tree-based learning algorithms that aims to increase the predictive accuracy of tree-based learning algorithms. They were originally designed to improve decision trees, but they can also be used to improve regression trees. The basic knowledge is to create a collection of decision trees in which the best attribute in each node is selected from a limited set of randomly chosen candidate attributes. The size of the forest (the number of trees generated) is usually at least 100, but it can be much larger.

Random Forest denotes an ensemble of very simple tree predictors, with each tree producing a different outcome. The tree outcome is a prediction of the dependent value given the predictors in the regression model. In the classification model, on another angle, the tree outcome is a membership of a class that categorizes a series of independent values based on the predictors that are present in the dependent value [6].

Decision Tree

Another example of a supervised machine learning algorithm is the Decision Tree Classification Algorithm. It can deal with both categorical and numerical information. It gives you a categorical answer based on conditions like Yes/No, True/False, 1, or 0. Medical databases are often handled by using the Decision Tree Classification algorithm. For medical datasets, this algorithm's accuracy is usually much higher than the other algorithms.

The immediate aim of decision trees is to find the descriptive characteristics that provide the most information about the targeted feature and then divide the dataset along with the parameters of these attributes so that the target attribute values for the sub-datasets are as clean as possible. This explains why the descriptive feature that leaves the goal feature the barest is considered the most informative. The method of locating the most informative function continues until a stopping criterion is met, at which point we arrive at leaf nodes.

XGBoost

The eXtreme Gradient Boosting is represented as XGBoost. However, the term XGBoost corresponds to the engineering objective of pushing boosted tree algorithms' computational capabilities to their limits. A few of the motives XGboost is so famous for are because of this.

XGBoost is a straightforward application. This implementation is extremely fast as compared to other gradient boosting methods.

Scalability throughout all scenarios is still the most relevant feature of XGBoost's efficiency. The system is 10 (ten) times faster than current standard solutions on either a single computer, and it extends to billions of instances in decentralized or memory-constrained settings. Its scalability is dependent on several major aspects and algorithmic improvements. A new tree learning algorithm for sparse data has been developed, as well as a technically supported weighted quantile draft procedure for handling instance weights in estimated tree learning and a theoretically validated weighted quantile draft procedure for estimate tree learning. Parallel asynchronous learning accelerates learning [7]

Neural Networks

A neural network is a sequential model composed of a number of nodes linked by directional links, as the name implies. The ties between nodes define the correlation between linked nodes, and then each node is a representative processing unit. All nodes are adaptive, which means that their outputs are dependent on node-specific parameters that can be changed [8].

Neural networks, particularly descriptive and predictive data mining, may be used for many purposes. In the original field of machine learning, they were programmed to mimic the neurophysiology of the human brain in a very integrated system by integrating computational simple elements (neurons)

3. Implementation Of the Proposed Algorithm

A. Dataset description

There are eleven characteristics and one target variable in the cardiovascular dataset used in this project. These features are divided into six (6) nominal variables and 5 numeric variables. Below is a detailed description of all the features in the dataset.

Table 1: Description of the dataset attributes and their value.

Attribute	Description
Age	This represents the ages of patients in years <100. It is a numerical feature.
Sex	This is the gender of the patients. It is a nominal feature. 1 represents Male and 0 represents Female.
Chest Pain Type	This represents the different types of chest pains experienced by patients. Chest pains are categorized into the following: typical = 1, typical angina = 2, non-anginal pain = 3, asymptomatic = 4
Resting BPS	it is the level of blood pressure of patients at the resting mode in mm/HG
Cholesterol	it is the Serum cholesterol of patients in mg/dl
Fasting blood sugar	It is the Blood sugar levels of patients on fasting greater than 120 mg/dl represents as 1 in case of true and 0 as false
Resting ECG	It is the result of an electrocardiogram while at rest are represented in 3 distinct values: Normal =0, Abnormality in ST-T wave = 1, Left ventricular hypertrophy = 2
Max heart rate	The represents the highest heart rate achieved
Exercise angina	It is angina induced by exercise. 0 shows NO and 1 depicts Yes
Exercise-induced ST	It is the depression in comparison with the state of rest
ST slope	It is the ST segment measured in terms of the slope during peak exercise. It is a numeric feature. It is divided into the following: Normal = 1, Upsloping = 2, Flat = 3, Downsloping=4
Target	It is the target variable or labels the model will predict. If the predicted value is 1, it means the patient is suffering from heart risk and if the predicted value is 0 means, it means the patient is normal.

Distribution of Heart disease among patients in the dataset

Distribution of Heart disease among patients in the dataset. The distribution of heart disease among patients is presented below. According to the result obtained, 628 (53%) patients have heart disease and 561 (47%) of the patients do not have heart disease.

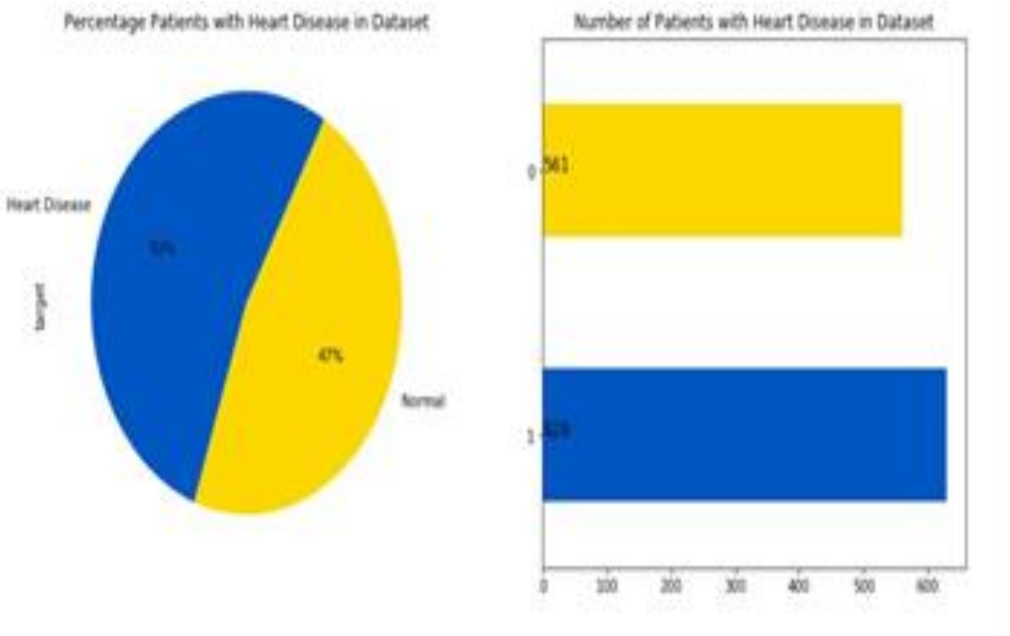

Fig. 1: Distribution of Heart disease among patients in the dataset.

Gender and Age distribution of heart disease

From the result shown below, it is concluded that 76% of patients who have heart disease are males and 24% are females. And the mean (average) age of patients in the dataset is approximately 55 years.

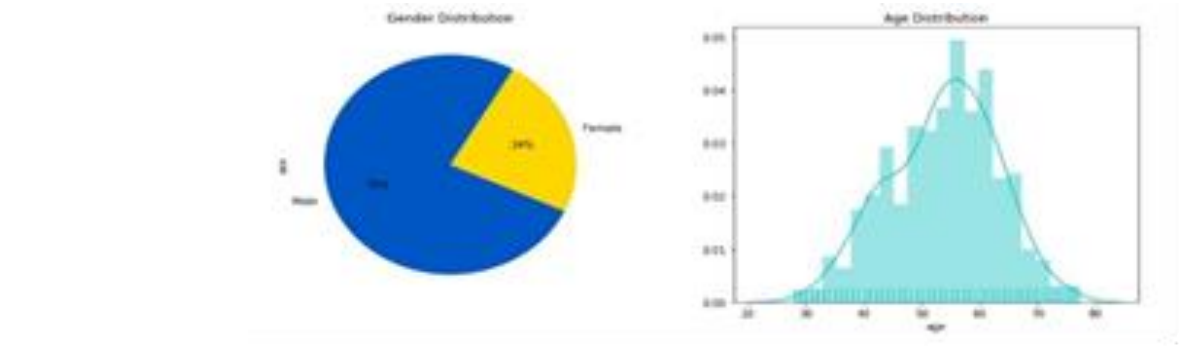

Fig. 2: Gender and Age distribution of heart disease.

B. **Performance measurement**

To evaluate the validity of the predictive model, various measurements can be calculated such as sensitivity, specificity, accuracy, and precision, by using the confusion matrix (Table 2).

A confusion matrix is an error matrix, also described as a confusion matrix. It is a table layout that lets you see how well a classification model performs [9]. A two-by-two confusion matrix could be generated using a binary classifier and a collection of instances. A binary classifier with expected positive (p) and negative (n) values, for example, has four possible outcomes:

- It is a true positive if the real instance is positive and it is expected to be positive (TP).
- It is a false positive if the real instance is negative but the expected result is positive (FP).
- A false negative occurs when the real event is positive but the expected result is negative (FN).
- It is a real negative if the actual event is negative and the expected result is negative (TN).

Table 2: Confusion Matrix.

		Actual values	
		Positive	Negative
Predicted values	Positive	TP	FP
	Negative	FN	TN

$$Sensitivity = \frac{TP}{TP+FP}$$

$$Specificity = \frac{TN}{TN+FP}$$

$$Accuracy = \frac{(TN+FP)}{(TN+FP+FN+TP)}$$

$$Precision = \frac{TP}{TP+FP}$$

4. Experimental Results

Different experiments are conducted to evaluate the performance and the validation of the developed model. Results are estimated using confusion matrix measurements and to compare the accuracy using different algorithms. The first experiment used the train-test split validation technique where the dataset is divided into two parts, and made several tests with different percentages, the best splitting achieved is 90% of the data for training and 10% for testing. Fig. 2 shows the results obtained by applying XGBoost, Random Forest, Neural Network, and Decision Tree algorithms. Based on the experimental results shown in Fig. 2, it is clear that the classification accuracy of the Random Forest algorithm is the highest, followed by XGBoost compared to other algorithms.

Table 3: Comparison of the performance of ML algorithms.

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC
Random Forest	0.9496	0.9118	1	0.8947	0.9538	0.9474
Decision Tree	0.8824	0.9444	0.8226	0.9274	0.8793	0.885
Multi-Layer Classifier	0.8487	0.94	0.7581	0.9474	0.8593	0.8527
XGBoost	0.9328	0.9355	0.9355	0.9298	0.9355	0.9327

A. Performance Comparisons of the Models

True Positive Rate (Sensitivity) and True Negative Rate were the other success performance metrics used to compare the performance of the model. Again, relative to the other models, both the Random Forest and the Decision Tree models performed extremely well here. The True Positive Rate and the True Negative Rate for Random Forest, Decision Tree, Neural Network and XGBoost models were (TP Rate, TN Rate) = (100%, 89.47%), (82.26%, 92.74%), (75.81%, 94.74%) and (93.55%, 92.98%).

The Random Forest model scored the highest TP Rate (100%) and the multi-layer perceptron model scored the lowest True Positive Rate (75.81%). In comparison to the other models, it was simpler for the Random Forest model to The Random Forest and XGBoost classifiers outperformed the other algorithms in terms of True Positive Rate, True Negative Rate, and F-Measure values. The Random Forest classifier once again obtained the highest ROC Area score.

Correctly identify positive cases compared to the other models, whereas a model constructed from Decision Tree struggled a little to accurately identify positive cases compared to the others. One interesting thing found here was that all the models relative to the positive ones were better at predicting negative events. When it comes to the ROC area, Table 3 shows that the Random and XGBoost classifiers are the most reliable when looking at the area under the curve (AUC) as a measure of separation consistency.

Model Selection

The next move was the selection of the best model based on such comparisons after analyzing the performance of the models. It is important to see things from the perspective of the clinician to do so. Since cardiovascular disease is a fatal illness, a medical practitioner may tend to keep the number of false positives low and maintain high levels of true positives. It is detrimental to tell a healthy patient that he or she is sick. The main factor for successful treatment is the early diagnosis of a disease. The classification models are therefore supposed to perform well in identifying positive cases, and the focus is more on the True Positive rate when choosing the best model. The total accuracy, then, for this research, the aim is to correctly classify patients with heart disease as much as possible. The True Positive Rate is preferably supposed to be closer to 100% as is realistically feasible. In other words, in return for improving the accuracy of positive classifications, one should be able to compromise the accuracy of negative classifications. Based on this assumption, the best predictive model for this research is the Random Forest classifier. The experimental findings have shown that the Random Forest algorithm has generally outperformed other models in the field of predicting cases of heart disease.

Performance of the Selected Model

Based on the exploration of the heart disease dataset and the goals of this research, the data mining objective of this research was established. The goal was measured against the model selected. The results show that the chosen model achieved the stated objective. The model could be used in the diagnostic process of cardiovascular disease. It will, therefore, assist cardiologists to interpret echo recordings.

The machine learning aim for this research was to give patients' echo results for each attribute; classify patients into two categories; those who are diagnosed with heart disease and those who are free from the disease. The selected model built with Random Forest classier was able to answer this question by predicting 94.96% of the cases correctly. As the classification accuracy value is high, cardiologists could rely on it for assisting heart disease diagnosis.

B. Classification Accuracy of the Selected Model

The ultimate goal was to build a predictive model with an accuracy (classification) of 100%. Unfortunately, the chosen model only had a classification accuracy of 94.96 percent. There were three major explanations for the 5.06 percent of cases that were misclassified.

- The structure of the dataset itself had a minor impact on classification accuracy. The dataset presented challenges such as high dimensionality and unequal groups. The True Positive Rate is higher than the True Negative Rate in all of the models. This appears to be partly because the learning algorithm used for this study developed a bias
- Despite the attempts made to find noisy data and substitute missing values, the accuracy of the model has been harmed due to missing or incomplete data and noise. It is almost impossible to capture all of the noise in a dataset, and the filled-in value might not be right in some situations.

5. Conclusion

In this study, the aim was to design a predictive model for heart disease detection using a dataset obtained from[1]. The model was built on the dataset with four different supervised machine learning algorithms i.e. Random Forest, Decision Tree, Multilayer Perceptron, and XGBoost using python. The accuracy, recall, precision, and F-measure of the model was used to test its performance. The training and test data samples were sampled using 10-fold cross-validation. All four models were effective in predicting cases of heart disease. The most effective model to predict patients with heart disease appears to be a Random Forest classifier implemented with a classification accuracy of 94.96 %.

Heart disease is a fatal disease by its nature and misdiagnosis of this disease can cause serious, even life-threatening complications such as cardiac arrest and death. The best model selected for predicting heart disease could not exceed a classification accuracy of 94.96 % and still much remains to fill the gap of 5.04% misclassified cases.

This study showed that data mining techniques can be used efficiently to model and predict heart disease cases. The outcome of this study can be used as an assistant tool by cardiologists to help them to make a more consistent diagnosis of heart disease. Furthermore, the resulting model has a high specificity rate which makes it a handy tool for junior cardiologists to screen outpatients who have a high probability of having the disease and transfer those patients to senior cardiologists for further analysis.

This research has shown that machine learning techniques could be used to predict heart disease, and the models developed as a result of this research are worthy of clinical testing. Further research should be performed using various classification algorithms such as support vector machines, and deep learning methods to enhance the classification accuracy of the models. The experiment in this study was carried out with little parameter tuning of the algorithms; however, further research should be carried out with different parameter settings to develop and extend the prediction models' capabilities. In the machine learning method, missing values, noisy data, anomalies, and outliers

[1] https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive IEE Dataport

posed a challenge. As a consequence, machine learning methods can be used to improve data quality.

References

1. R. Atallah and A. Al-Mousa, “heart disease detection using machine learning majority voting ensemble method,” in, p. 1, 2019.
2. S. K. Dey, A. Hossain, and M. M. Rahman, “Implementation of a web application to predict diabetes disease: An approach using machine learning algorithm,” in, 2018.
3. R. Atallah and A. Al-Mousa, “heart disease detection using machine learning majority voting ensemble method,” in, 2019.
4. S. K. J. and G. S., “Prediction of heart disease using machine learning algorithms.” in, 2019.
5. H. Lin, P. K. Yang, Y. C. Lin, and P. K. Fu, “On machine learning models for heart disease diagnosis,” in, 2020.
6. H. E. Hamdaoui, S. Boujraf, N. E. H. Chaoui, and M. Maaroufi, “A clinical support system for prediction of heart disease using machine learning techniques,” in, p. 3, 2020.
7. T. Chen and C. Guestrin, “Xgboost,” p. 1, 2016.
8. M. Kantardzic, p. 278, IEEE Press, 2011.
9. A Kulkarni, D. Chong, and F. A. Batarseh, “Foundations of data imbalance and solutions for a data democracy,” p. 86, 2020.
10. S. Krishnan J and S. Geetha, “Prediction of heart disease using machine learning algorithms,” in, p. 3, 2019.
11. J. H. Chen and S. M. Asch, “Machine learning and prediction in medicine —beyond the peak of inflated expectations,” vol. 376, no. 26, p. 2508, 2017.