International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 1, ISSN No. 2455-2143, Pages 36-46
Published Online May 2020 in IJEAST (http://www.ijeast.com)

PERFORMANCE ANALYSIS OF SOME SELECTED MACHINE LEARNING ALGORITHMS ON HEART DISEASE PREDICTION USING THE NOBLE UCI DATASETS

Lamido Yahaya
Department of Computer Science
Gombe State University,
Gombe, Gombe State, Nigeria

Nathaniel David Oye
Department of Computer Science
Modibbo Adama University of Technology
Yola, Adamawa State, Nigeria

Abubakar Adamu
Department of Mathematics
Gombe State University,
Gombe, Gombe State, Nigeria

Abstract: **Heart disease is one of the major causes of morbidity and mortality in the world. The diagnosis and treatment are very complex, especially in the low income countries, due to the rare availability of efficient diagnostic tools and shortage of physicians which affect proper prediction and treatment of patients. Lack of awareness, inadequate preventive measures, lack of experienced medical professionals are among the factors that contribute to high risk of heart disease occurrences. Although, large proportion of heart diseases could be prevented but they continue to rise mainly because preventive measures taken are inadequate. Nowadays, several clinical decision support systems on heart disease prediction have been developed using the most popular machine learning algorithms and tools. This paper analyses the performances of these algorithms on heart disease prediction using the noble UCI datasets. They include Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT-J48), Random Forest (RF), K-Nearest Neighbor (KNN) and Neural Network (NN). From our investigation, these algorithms were mostly used in which RF appeared the best in the prediction of heart diseases using the mentioned datasets. From the 34 researches investigated, RF was used 10 times and appeared the best 4 times, followed by SVM whose frequency of usage was 18 times with 6 best performances. From the most popular algorithms, KNN was employed 10 times but appeared the best only once. Others such as LR and MLP were used 7 and 5 times respectively but none recorded a single best performance in the prediction of heart diseases, while FCM and Vote were not popular and were rarely considered.**

Keywords-- **Machine Learning, Algorithms, Heart Disease, Classification, Prediction**

I. INTRODUCTION

The heart is a kind of muscular organ which pumps blood into the body and is the central part of the body's cardiovascular system [1]. It forms a complex system in collaboration with arteries, veins and capillaries, which control blood flow throughout the body. Any defect or damage that causes obstruction in the normal blood circulation or flow from the heart would result in severe complications of heart diseases. These are commonly called cardiovascular diseases (CVDs) and according to the World Health Organization (WHO) reports, are among the deadliest diseases in the world. Cardiovascular disease is caused by disorders of the heart and blood vessels, and includes coronary heart disease (heart attacks), cerebrovascular disease (stroke), raised blood pressure (hypertension), peripheral artery disease, rheumatic heart disease, congenital heart disease and heart failure [2]. CVDs include diseases of the heart, vascular diseases of the brain, and diseases of blood vessels [3]. CVDs are generally preventable through various measures, such as changes in life style but they are still on the increase every day. According to [4], CVDs are responsible for the deaths of 17 million people each year, or approximately one-third of global deaths annually. Another WHO reported in 2007 stated that of an estimated 58 million deaths globally from all causes in 2005, CVDs accounted for 30%, and between 2006 and 2015, deaths due to non-communicable diseases (half of which would be due to CVDs) were expected to increase by 17% [5]. The latest WHO report in 2017 stated that 17.9 million people die each year from CVDs, an estimated 31% of all deaths worldwide [6]. The various reports by the World Health Organization have indicated that deaths due to CVD cases have been on the increase, which are mainly attributed to inadequate preventive measures, poor medical facilities and shortage of physicians, especially in the low income countries. There are certain risk factors that increase a person's chances of having a cardiovascular disease. Some of these factors as enumerated by [7] include the following:

1. High Blood Pressure;
2. Abnormal Blood Lipids;
3. Use of Tobacco;
4. Obesity;
5. Physical Inactivity;
6. Diabetes;
7. Age;
8. Gender; and
9. Family Generation

Other risk factors include abnormal HDL (good) cholesterol, and high fat diet. With these factors and more, physicians generally make diagnoses by evaluating a patient's current test results and previous diagnoses made on other patients with the same results. Cardiovascular diseases are of many types, some of which as listed by [8] include the following:

1. Arrhythmia;
2. Coronary Artery Disease;
3. Dilated Cardiomyopathy;
4. Myocardial Infarction; and
5. Cardiac Arrest

The increasing rate of cardiovascular diseases has become a global concern. Therefore, the healthcare industry needs to shape and intensify the way these diseases are handled in order to minimize the impact in the society. According to [9], huge data is available in the healthcare industry. Specifically, the CVD data is also available, which needs to be efficiently analyzed for effective decision making, from which efficient predictive model could be developed. Based on data, statistics, clinical records and hospital management, it is claimed that in every three years medical data doubles up and making health industry a multi-billion dollar domain [10]. Machine learning techniques play a vital role in the analysis of such data for better control and prevention of heart diseases.

A. Machine Learning Techniques

Machine learning techniques play an important role in analyzing the accumulated medical data for effective diagnosis. The increasing morbidity and mortality due heart diseases or generally CVDs worldwide has attracted the attention of researchers to conduct many studies in their effort to minimize the rates [11]. Machine learning techniques have been widely used in the implementation of clinical decision support systems for heart disease prediction. Algorithms such as Naïve Bayes (NB), Neural Network (NN), Decision Tree (DT-J48), K-Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM) are among the most popular techniques used in heart disease data analysis and prediction. Other algorithms include Logistic Regression (LR), Multilayer Perceptron (MLP), Fuzzy C Means (FCM), and Vote. These algorithms can be used to enhance the data storage for practical and legal purposes [12].

In fact, there exists a wide gap between the accuracy of traditional heart disease prediction done by medical professionals and that of the modern approaches such as the machine learning techniques. Reference [13] stated that according to a survey conducted by WHO, the medical professional is able to correctly predict only 67% of heart disease. At the same time some of the developed models can predict an accuracy up to 90% or more. Therefore, correct heart disease prediction is a serious concern that created a vast scope for research in the healthcare sector. This is where the machine learning techniques play their role to improve the prediction accuracy to the highest percentage, so that the number of morbidity and mortality due to heart diseases resulting from inefficient diagnosis would be minimized. Machine learning and data mining applications may benefit the healthcare industry immensely but this depends on how clean the data is [14]. One of the most standard data that is mostly used by machine learning researchers is the UCI heart disease data, which is publicly available online.

B. The UCI Heart Disease Datasets

The UCI is an acronym for the University of California, Irvine in the United States. The UCI machine learning repository is a publicly available database that contains a vast amount of several datasets for machine learning researchers, in which the heart disease datasets are inclusive. There are four (4) heart disease datasets contributed from different sources, as enumerated by [15]:

1. Hungarian Institute of Cardiology. Budapest, data contributed by Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland, data contributed by William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland, data contributed by Matthias Pfisterer, M.D.

4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation, data contributed by Robert Detrano, M.D., Ph.D.

Each of the four datasets has the same instance format, while they have 76 raw attributes, but only 14 of them are actually used by researchers. The number of instances vary: Cleveland data has 303, VA Long Beach has 200, and Hungarian has 273 while that of Switzerland has 102 instances. Of the four datasets, the Cleveland data has been used mostly. Our recent investigation showed that more than 60% of the 53 research articles, employed the UCI data for heart disease prediction.

II. LITERATURE SURVEY BASED ON THE UCI HEART DISEASE DATA

This section presents a survey on those research works conducted based on the UCI heart disease data. There are 34 different articles extracted from our previous work, which is a comprehensive review on heart disease prediction [11]. The extracted research works were mostly conducted in 2018 and 2019, while very few came from 2015, 2016 and 2017. The fundamental objective of this paper is to analyze the performances of the selected algorithms solely based on the UCI data. The extracted papers are as follows:

Reference [16] presented a heart disease prediction framework using some supervised machine learning algorithms in R programming language. The algorithms used include Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Naïve Bayes (NB). The Cleveland datasets from the popular University of California, Irvine (UCI) machine learning repository consisting 303 instances and 76 features were used. The data was preprocessed due to missing values and the sample became 302 instances and only 14 heart disease features in size. The data was split into 70% and 30% for models training and testing respectively. It was a comparative analysis of the selected techniques in which the experimental results showed that the NB classifier performed the heart disease prediction better than the SVM and KNN, with an accuracy of 86.6%.

Reference [17] investigated a method termed ensemble classification, which is used for improving the accuracy of weak algorithms by combining multiple classifiers. The Cleveland heart disease datasets from the UCI machine learning repository were used for experiments. The sample contains 303 instances and 76 features in which 8 of the features are categorical while 6 are numeric. The sample was split into training and testing datasets for the classifiers, which was cleaned and preprocessed for missing values and invalid data removal. A comparative analysis of various classification algorithms was performed to determine the weak ones among them. To improve the performance of the weak algorithms, ensemble algorithms such as bagging, boosting, voting, and stack were employed. Then SVM, NB, RF, Bayes Net, C4.5, MLP, and PART classifiers were used for the classification of the datasets. C4.5, MLP, and PART were found to be weaker hence, the ensemble strategy was used to improve their performance. 10-fold cross validation was employed to evaluate the performance of the classification models. Experimental results of the individual classifiers showed that the accuracy rates of NB, RF, BN, C4.5, MLP, and PART were found in the range of 75.58% to 83.17%. The NB classifier appeared the best with 83.17% accuracy, while C4.5, MLP, and PART showed comparatively poor performance with accuracy less than 80%. The final results showed that ensemble techniques, such as bagging and boosting are effective in improving the prediction accuracy of weak classifiers, and exhibited satisfactory performance in identifying risk of heart diseases. A maximum increase of 7% accuracy for wear classifiers was achieved with the help of ensemble classification.

Reference [18] proposed a diagnostic system for predicting heart disease using Multi-Layer Perceptron Neural network (MLP) with back propagation as the training algorithm. The performance of the developed system was evaluated based on sensitivity, specificity, precision and accuracy. The Cleveland data of the UCI machine learning repository containing 303 instances and 76 features was employed for model training and testing. Data preprocessing was performed to remove 6 instances which contain missing values. Of the 76 features, only 14 were used as the most relevant to heart disease. Based on the experiments performed, the MLP-NN proposed model gave high accuracy of 93.39% for 5 neurons in hidden layer with running time of 3.86 seconds in the heart disease prediction.

Reference [19] proposed a Deep Neural Network (DNN) method to develop and design an automated system for heart attack prediction. They employed the Cleveland dataset from the UCI machine learning repository, which commonly contains 303 records and 76 attributes. Data preprocessing was carried out to remove anomalies, such instances with missing values and data redundancy. Their proposed DNN model contains 5 stages. Data selection stage is the first that collects the data, then data preprocessing stage which removes missing values and the like. The third stage is the classification stage where the DNN algorithm was used for the heart disease prediction, then validation stage and finally the result stage. Based on the performed experiments, their proposed DNN model performed better with an improved classification accuracy in heart disease prediction.

Reference [20] proposed a new approach for features selection using collaboration between well-known features selection techniques by accumulating features rank for all selected features selection techniques for heart disease prediction. They used the Cleveland datasets from the UCI machine learning repository, which contains 303 instances with 76 features, for algorithm training and testing. Information Gain, Gain ratio, Relief F, Symmetrical Uncertainty, and One-R feature selection techniques were applied on the datasets to select important features for the heart disease prediction. Of the 76 features, only 14 were used in the feature selection process. Applying these techniques, 4 features were removed on the basis of eliminating features with accumulated rank less than threshold (< 1). The eliminated features were tresbps, chol, fbs and restecg. The datasets were split into 2/3 and 1/3 for model training and testing respectively. WEKA tool was used to perform the classification accuracy, recall, precision and F-measure for the dataset. The classification techniques applied include NB, KNN, SVM, ANN, and DT-J48. All algorithms produced better prediction rate according to the experimental results obtained. But KNN classifier recorded the best enhancement rate in accuracy (+3.7%), precision (+8.1%), recall (+0.3.5%), F-measure (+0.5.7%), then followed by ANN, J48, SVM and NB.

Reference [21] proposed a heart disease prediction framework based on RF algorithm in machine learning using Python. They used the Cleveland heart disease datasets obtained from the UCI machine learning repository for the algorithm training and testing. This sample originally contains 303 instances with 76 features but after preprocessing and manual attribute selection of features, only 9 features were used. 75% of the sample was used for algorithm training while 25% was used for testing. A graphical user interface (GUI) was developed using Visual Studio Code for visualization of the experiments. The RF classifier was employed for the classification, where an accuracy of 97.56% was achieved. The heart disease diagnosis was divided into four (4) stages based on artery blockage, where an artery blockage greater than 50% indicates the presence of heart diseases.

Reference [22] proposed a web-based application for predicting heart diseases using machine learning techniques. The algorithms used for classification were SVM, LR, and NB. Heart disease datasets from the UCI machine learning repository were proposed for algorithm training and testing, divided into 75% and 25% respectively. The proposed system would have user interface, through which heart disease patients enter their information and a database implemented using MySQL, which stores patients' medical history. Data preprocessing was carried out to remove inconsistencies and missing values. From the three classifiers selected, SVM performed better with an accuracy up to 64.4%, and was selected for the main application. The remaining two algorithms, which are NB and LR had classification accuracies of 60% and 61.45% respectively. Based on their proposed experiments, the system would give prediction if a patient had a heart disease risk greater than 60%.

Reference [23] proposed a heart disease prediction framework called "Hybridization" that combined several machine learning algorithms into a single model. The Cleveland datasets from the online machine learning repository of the UCI consisting of 303 instances and 14 features were used in the model training and testing processes. Data preprocessing was carried out to reduce the attributes from 14 to 12. The range of classification algorithms applied included the NB, SVM, KNN, NN, J48, RF, and GA, taking into account their accuracies, sensitivities and specificities in the heart disease prediction. They were applied on the same dataset and features one after the other. The results of the experiments showed that NB and SVM performed better in the heart disease prediction with the same accuracy of 89.2%.

Reference [24] proposed a Map-Reduce based framework for heart disease prediction in patients. It was to extract the needed information from the records of heart disease patients. The UCI datasets of the Cleveland heart disease data containing 303 instances and 76 features were used. The data was preprocessed to remove missing values after which whole sample size became 297 instances with only 14 attributes. The sample was divided into 250 and 47 instances for model training and testing processes respectively. They compared the performance of meta-heuristic approach and that of trained persistent fuzzy neural network on the datasets. Experimental results showed that their proposed Map-Reduce based algorithm performed better in the heart disease prediction with a classification accuracy of 98.12%.

Reference [25] performed a comparative study on heart disease classification and prediction using machine learning techniques. The algorithms used include NB, DT, RF, SVM, and LR in the Rapid-Miner. The common Cleveland heart disease datasets from the UCI machine learning repository consisting of 303 instances and 14 attributes were used. During learning and of the model, 10-fold cross validation technique was used. From the results of the experiments, DT algorithm appeared the highest in the heart disease prediction accuracy followed by SVM at 93.19% and 92.30% respectively.

Reference [26] performed a comparative analysis on some of the popular machine learning algorithms used for heart disease prediction. WEKA 3.6 version was used to study four classifiers including RIPPER, DT, ANN, and SVM. The usual UCI datasets for Cleveland containing 303 instances and 14 attributes were used for the model training and testing. Data preprocessing operation was performed which subsequently reduced the sample size to 296 instances. The essence was to remove records with missing values. The performances of the selected algorithms were compared with other classifiers which include the KNN, NB and MLP. The experimental results showed that the selected algorithms performed better, with SVM having the performance of 90.00% accuracy.

Reference [27] proposed a heart disease prediction based on machine learning techniques using NB and DT algorithms in Python. The datasets used for training and testing of the model were obtained from the Kaggle website, which contains 13 heart disease features. Another dataset from the UCI machine learning repository was used for the simulation. The proposed model was implemented on the Scipy environment. Form their experiments, results showed that DT algorithm performed better than the NB in the prediction of heart diseases.

Reference [28] performed a comparative study on some of the most popular classification models used in data mining. They include K-Nearest neighbor (KNN), Support Vector Machine (SVM) and Artificial Neural Network (ANN) using MATLAB multilayered feed forward back propagation. Cleveland heart disease data containing 303 instance with 76 features from the UCI machine learning repository was used. They performed data preprocessing to remove records with many missing values, where the data size became 270 instances with only 13 attributes. Half of the data was used for models training and the other half for testing. Their experimental results showed that SVM outperformed both the KNN and ANN based on the classification accuracy at 85% while KNN at 82% performed better than ANN at 73% approximately.

Reference [29] conducted a study to identify the most significant features in heart disease prediction. In their system framework, seven classification algorithms in the Rapid-Miner studio were used, which include the KNN, DT, NB, LR, SVM, NN, and Vote. The Cleveland data containing 303 instances with 76 features obtained from the UCI machine learning repository was used. They performed a cross validation on the data using 10 folds cross validation approach. One subset was used for training and the remaining for testing. From the results of their experiment, Vote classifier appeared the best in the heart disease prediction with an accuracy of 87.4%.

Reference [8] also carried out a comparative investigation on heart disease prediction using support vector machine, decision tree, and k-nearest neighbor algorithms. They used the VA Long Beach dataset obtained from the UCI machine learning repository, which comprises of 270 instances and 12 attributes for the algorithm training and testing purposes. The model was evaluated based on accuracy, sensitivity, and specificity using confusion matrix. Their experimental results showed that Support Vector Machine (SVM) performed better than KNN and DT in classifying the heart disease patients, with an accuracy of 92%, sensitivity of 100%, and specificity of 83%.

Reference [30] developed a machine learning based hybrid intelligent system framework for heart disease patients' diagnosis using seven of the popular classification algorithms using Python. They include KNN, ANN, DT, SVM, NB, LR and MLP. Cleveland dataset containing 303 instances with 76 features was used for model training and testing. They applied a 10 folds cross validation approach on the data. Feature selection algorithms which include Relief, Minimal-Redundancy-Maximal-Relevance (mRMR) and Least Absolute Shrinkage and Selection Operator (LASSO) were used to select the best heart disease correlated features. The data was preprocessed to remove the instances with large missing values. This reduced the data size to 297 instances with only 14 features. Applying the feature selection algorithms reduced the features to 6 only as heart disease related. They tested each of the classifiers with any of the feature selection algorithms in order to get the best performing model. Their experimental results showed that SVM with LASSO feature selection algorithm appeared the best performing combination, as compared with other feature selection algorithms and classifiers.

Reference [31] proposed a hybrid model combining Genetic and Naïve Bayes algorithms in python for heart disease prediction. They used the popular UCI dataset of Cleveland comprising 303 instances and 14 heart disease features. Based on the comparative analysis performed with other data mining techniques such as Weighted Fuzzy Rules, Logistic Regression, as used in previous studies, their proposed model, GA_Fuzzy_Naive was found to be more accurate at 97.14%.

Reference [1] proposed a tentative design of a cloud-based heart disease prediction system using machine learning techniques. Two of the UCI datasets: Cleveland heart disease data consisting of 303 instances with 14 features and VA Long Beach data consisting of 270 instances with also 14 features were merged together making a bigger dataset. Five machine learning algorithms, including MLP, LR, NB, RF, and SVM in the Java-based open access platform (WEKA) were applied in the classification and prediction processes. Of the five algorithms, SVM appeared the best classifier with a classification accuracy of 97.53%.

Reference [32] proposed a framework that combined the popular Naïve Bayesian classifier and Particle Swarm Optimization (PSO) feature selection algorithm for efficient heart disease prediction. The UCI dataset of VA Long Beach consisting of 270 instances and 14 features was used for the model training and testing processes. Of the 14 features, only 7 were selected for the heart disease prediction. From the experimental results, the Naïve Bayes predictive model performance was 79.12% accurate but escalated to 87.91% when integrated with the PSO selection algorithm. It was concluded that the NB+PSO model improved the heart disease classification accuracy, which is 8.79% better than the original NB performance.

Reference [33] analyzed different data mining techniques and procedures for testing their precision and execution on preparing medical information index. The UCI dataset of Cleveland consisting 303 instances and 76 features were used for the classifiers training and testing processes. Data preprocessing was carried out and 10 folds cross validation process was used for data validation. Of the 76 heart disease features, only 14 were used for the prediction using NB, j48, RF, AdaBoost, Bagging, MLP, and Simple Logistic Regression classifiers. Based on the experimentation, NB, Simple Logistic Regression and RF performed better in the heart disease classification and prediction.

Reference [34] used three of the most popular data mining techniques: RF, NB and DT to develop a prediction system in order to analyze and predict the possibility of heart diseases. Their fundamental objective was to identify the best classification algorithm suitable for providing maximum accuracy when classification of normal and abnormal person was carried out. The UCI dataset of VA Long beach consisting of 270 instances and 13 heart disease features were used for models' training and testing processes. The dataset was split into 80% and 20% for models training and testing respectively. Their experimental results showed that RF classifier performed better than NB and DT in the heart disease prediction.

Reference [35] presented a framework based on neural network (NN) to develop an effective heart disease prediction system (EHDPS) for predicting the risk level of heart disease. The Multi-Layer Perceptron (MLP) neural network (NN) with back propagation was used as training algorithm. The UCI dataset of the Cleveland consisting of 303 instances and 15 heart disease features was used for the model training and testing processes. The data was divided into 40% and 60% for the training and testing respectively. Data preprocessing operation was carried out to remove noisy data and missing values. Their experimental results showed that the proposed model was able to predict heart diseases with 100% accuracy.

Reference [36] proposed a hybrid approach for heart disease prediction using machine learning algorithms optimized by Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) techniques. Fast Correlation based Feature selection (FCBF) method was used to remove redundant features from the datasets, for improved performance of the classifiers. The Cleveland heart disease datasets of the UCI machine learning repository were employed for algorithm training and testing. The sample size is commonly known with 303 instances and 76 features, form which only 7 features were left for heart disease prediction when the feature selection techniques were applied. The classification experiments were carried out in WEKA environment using five algorithms, which include RF, KNN, SVM, NB, and MLP using 10-fold cross validation method. The effectiveness of the classifiers was evaluated based on time to build the model, correctly classified instances, incorrectly classifies instances, and accuracy. The evaluation was performed in the first instance without optimization, then with FCBF optimization and finally with the FCBF, PSO, and ACO optimizations. Confusion matrices were created to represent the evaluation scenario. Accuracies of the classifiers were measured based on precision, recall, TP rate and FP values. Comparative analysis of the result obtained showed that the proposed optimized model with FCBF, PSO and ACO techniques achieved an accuracy of 99.65% in predicting patients with heart diseases, with the KNN classifier and 99.60% with RF classifier.

Reference [37] proposed a "vertical system integration of a sensor node and toolkit of machine learning algorithm for predicting heart diseases in patients". The system could be used for both heart disease monitoring and diagnosis. The pulse rate sensor (AMPED) and

Bluetooth were integrated to monitor the Heart Rate Variability (HRV) and send the data to mobile application. The graph of the heart rate and heart disease prediction could be seen through the mobile application. The prediction is done based on the increase or decrease of the HRV value. They used the common Cleveland heart disease datasets from the UCI machine learning repository for algorithm training and testing. The sample size is commonly known with 303 instances and 76 features. Half of the sample was used for training and building of the classification model. The machine learning algorithms used were DT, NB, SVM, KNN, LR, and RF. Experimentation with these algorithms yielded a result with an accuracy of 89% in predicting the heart disease using the RF classifier. For the monitoring purposes, two experiments were carried out. In the first instance, 20 healthy individuals were used, from which 5 were asked to hold the model and play sorts, no alarm was raised, because they were all healthy. In the second instance, also 20 but unhealthy individuals were used and alarm was raised, an SMS was sent to the nearest hospital. Their results showed that in both cases an accuracy of 100% was achieved.

Reference [38] presented a heart disease prediction framework using Naïve Bayesian classifier and K-means algorithms. Their datasets were obtained from the UCI machine learning repository, but not clearly specified. They used 13 heart disease features such as age, gender, high blood pressure, cholesterol, etc. to estimate the possibility of heart disease for a person. The system allows patients to enter their information, which would be classified as either normal or heart disease stages 1, 2, or 3. Using the NB and improved K-means algorithms, the risk rate of heart disease was detected and accuracy level also provided obtained according to the number heart disease features entered.

Reference [39] presented a heart disease prediction framework that shows how synthetic data would be used to address privacy concerns and overcome constraints inherent in small medical research datasets. They examined the used of surrogate datasets comprising synthetic observations for modelling the system. The data was generated based on the characteristics of original observations and compared the prediction accuracy results obtained from LR, DT and RF. The UCI dataset of the Cleveland heart disease data containing 303 instances with 76 features, which was preprocessed to become 279 instances and 14 features was used as the original data. The experiment was divided into three stages. In the first stage, the base line models and their results were established, which fundamental objective was to validate and compare the accuracy and stability of the results of the proposed models to those in the previous studies. In the second stage, the same original data (Cleveland: 279 instances and 14 features) was used to generate 50,000 records, and it was used to train and test the previous LR, DT, and RF models. In the last stage, 60,000 records were generated from the same original dataset, and was used to train and test the ANN model of the perceptron forward and backward propagation algorithm type. From their experiments using the traditional models (LR, DT, and RF) with the surrogate data, they achieved an improved prediction stability within 2% variance at around 81% using 10-fold cross validation. While using the ANN with the surrogate data, they improved the heart disease prediction accuracy by nearly 16% to 96.7% while maintaining stability at 1%.

Reference [40] presented a medical diagnosis system framework for predicting the risk of cardiovascular disease using genetic algorithm (GA) and multilayered back propagation neural network (NN) models. The UCI unspecified heart disease dataset was used in training and testing of their models. Data preprocessing was carried out to remove instances with large missing values for improved prediction performance. Min-Max Normalization was adopted to replace the missing values by the most probable value in the format of (####). The hybrid system used backward propagation algorithm for learning and training the neural network. The classification accuracy obtained using this framework was up to 90.17%. The result was displayed in a graphical format either using pie chart or bar chart.

Reference [41] developed a heart disease prediction system which they compared with NB algorithm performance. The Cleveland data obtained from the UCI machine learning repository was employed during model training and testing scenarios. The sample size consists of 303 instances and 76 features. The proposed system receives patients information which would be classified a 0 absent “Absent” or 1 which means “Present” of heart disease. Results of the comparative analysis with the NB classifier showed that their proposed model performed better with classification accuracy of 97%.

Reference [7] performed a comparative study on various machine learning algorithms for predicting patients with heart disease through graphical representation of results. For the algorithm training and testing, the Cleveland heart disease datasets from the UCI machine learning repository were used. The sample is originally composed of 303 instances with 76 features, from which 14 were used for the classification. Classifiers in WEKA, which include NB, SVM, DT, and

KNN were used. Experimental results were represented graphically, which showed that NB classifier performed better than the other classifiers in predicting patients with heart diseases correctly.

Reference [12] presented a detailed study on some of the common classification algorithms including NB, and DT for heart disease prediction. The Cleveland dataset obtained from the UCI machine learning repository containing 303 instances and 76 features were used for model training and testing. Of the 76 heart disease, 19 were selected and used for the prediction. From the experimental results, DT outperformed NB classifier in terms of accuracy in the heart disease prediction.

Reference [42] compared the performances of J48, Logistic Model Tree (LMT), and Random Forest (RF) algorithms in WEKA for heart disease prediction task. The Cleveland datasets from the UCI, which commonly consist 303 instances and 76 features were used for the training and testing using the 10-fold cross validation method. The evaluation was based on accuracy, sensitivity, and specificity. Experimentation was carried out on Core i3 with 2.4GHz CPU and 4GB RAM. Results showed that J48 appeared with the highest classification accuracy of 56.76% followed by LMT at 55.77%, then RF.

Reference [43] proposed a framework that can proficiently find the tenets to foresee the risk level of patients in view of the given parameter about their health. The main objective was assist non-specialized doctors to make predictions about the heart disease risk level. The Cleveland datasets obtained from the UCI machine learning repository, containing 303 instances and 76 features were used for training and testing. The model was trained and tested using 10-fold cross validation. Data preprocessing was carried out to remove noisy data and missing values. They used covering rule model which C4.5 in WEKA for the classification process. Knowledge Extraction based on Evolutionary Learning (KEEL) was used for the implementation. Experiments were performed together with other well-known algorithms which include SVM, NN, and MLP, in which the proposed classifier (C4.5) performed better in the heart disease prediction with an accuracy of 86.7%.

Reference [44] designed a framework for heart disease prediction using data mining techniques. One of the UCI datasets was used to train and test the system using 10-fold cross validation method. SVM, NB, KNN, C4.5, Back Propagation classifiers were used and performances were compared. SVM classification algorithm appeared the best in terms of accuracy, sensitivity, precision, low specificity, mean absolute error and low computing times in all feature combinations. The SVM classifier accuracies at 13, 12, 11, 10, 9, 8 and 7 feature combinations were 83.70%, 84.00%, 84.00%, 84.10%, 84.40%, 84.80%, and 85.90% respectively.

Reference [2] presented a framework called “Heart Attack Prediction using Data mining Techniques” using Fuzzy C Means classifier to predict the risk of heart attack in patients. The Fuzzy C means is an unsupervised machine learning clustering algorithm that allows one piece of data to belong to two or more clusters. The datasets used for the model training and testing were obtained from the UCI machine learning repository. The sample size contains 270 instances and 76 heart disease features, in which only 13 were used for the heart attack prediction. Data preprocessing was carried out to remove missing values. The results of classification experiment performed showed that the proposed classifier (Fuzzy C Means) achieved better accuracy than most of the existing classification algorithms.

III. PERFORMANCE ANALYSIS

The UCI heart disease data, especially the Cleveland data has been used extensively for research in heart disease classification and prediction. Prominent machine learning tools used include Python, WEKA, R and MATLAB. Popular algorithms from either of the tools, which include NN, NB, DT-J48, KNN, RF, and SVM were frequently applied for the heart disease predictions. Other algorithms include MLP, LR, FCM and Vote. Very few researchers considered using hybrid approaches, such as Genetic Algorithm-Neural Network (GA_NN), Fuzzy Neural-Genetic Algorithm (FNGA) and Genetic Algorithm-Naïve Bayes (GA_NB). From our empirical investigation on comparative studies conducted, it was observed that the NB algorithm had the highest frequency of usage, which is up to 20 times, but appeared the best in the prediction of heart diseases using the UCI data only 6 times, followed by J-48 and SVM with frequencies of usage 18 times each, with best performances as 4 and 8 respectively. Of the 6 most popular algorithms selected, RF performed better with 10 frequencies of usage and 4 best performances, while KNN remain the last with 10 frequencies of usage and only 1 best performance. Algorithms like LR and MLP were used 5 and 7 times respectively with no single best performance. From the cumulative 34 researches investigated, FCM algorithm was used 2 times with 1 best performance, while Vote was used only once. The two algorithms are very unpopular and were rarely used on the data. The three hybrid methods were used only once each by some researchers to predict heart diseases using the UCI data. Their performances were not far better from some of the original algorithms, such as the RF and SVM. Based on our study, hybrid approaches are also not popular in predicting heart diseases, especially on the UCI data. The summary of this discussion is as shown in table 1 below:

Table 1: Performance Comparison of Algorithms

S/N	Algorithm	Frequency of Usage	Best Performance	Frequency of Failure
1	FCM	2	1	1
2	J-48	18	4	14
3	KNN	10	1	9
4	LR	7	0	7
5	MLP	5	0	5
6	NB	20	6	14
7	NN	12	4	8
8	RF	10	4	6
9	SVM	18	8	10
10	Vote	1	1	0

IV. CONCLUSION & FUTURE WORK

Efficient heart disease prediction is essential due to the fact that the morbidity and mortality rates remain very high despite of the vast researches conducted every year. Classification or prediction accuracy depends largely on the type of data used. A classifier can perform differently on different datasets. The UCI heart disease data is a standard data used by machine learning researchers to predict heart diseases. Based on the investigation conducted with this data, it was discovered that NB algorithm has the highest frequency of usage followed by J-48 and SVM. But RF with only 10 frequencies of usage seems to be best in the prediction accuracy, where it appeared the best 4 times, followed by the SVM with 8 best performances out of 18. Algorithms such as the KNN, LR and MLP performed poorly in the prediction. Other algorithms such as FCM and Vote are very unpopular and were rarely used. Therefore, it’s recommended that with the UCI data, machine learning models developed using RF and SVM algorithms are more accurate and should be considered the baseline for heart disease prediction henceforth. This might help to improve the prediction accuracy thereby controlling the escalating rate of mortality due to heart disease or CVDs in general. Next moment, the performances of these algorithms would be investigated on the heart disease datasets used other than the UCI data.

V. ACKNOWLEDGEMENTS

The authors acknowledge the support of the Tertiary Education Trust Fund (TETFUND), Nigeria for the sponsorship of both local and international scholarships.

VI. REFERENCES

[1] S. Nashif, M. Raiban, M. Islam and M. H. Imam, "Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system," World Journal of Engineering and Technology, vol. 6, pp. 854-873, 2018.

[2] G. R. Banu and J. H. B. Jamala, "Heart attack prediction using data mining technique," International Journal of Modern Trends in Engineering and Research, vol. 2, no. 5, pp. 428-432, 2015.

[3] WHO, "Global Atlas on cardiovascular disease prevention and control," WHO Library Cataloguing, Geneva, 2011.

[4] WHO, "Integrated Management of Cardiovascular Risk," WHO Library Cataloguing, Geneva, 2002.

[5] WHO, "Prevention of cardiovascular disease: guidelines for assessment and management of cardiovascular risk," WHO Press, Geneva, Switzerland, 2007.

[6] WHO, "Global action plan for the prevention and control of noncommunicable diseases," WHO Library Cataloguing, Geneva, 2017.

[7] S. K. Sen, " Prediction and diagnosis of heart disease using machine learning algorithms," International Journal of Engineering and Computer Science, vol. 6, no. 6, pp. 21623-21631, 2017.

[8] K. Hariharan, W. S. Vigneshwar, N. Sivaramakrishnan and V. Subramaniyaswamy, "A comparative study on heart disease analysis using classification techniques," International Journal of Pure and Applied Mathematics, vol. 119, no. 12, pp. 13357-13366, 2018.

[9] A. Solanki and M. P. Barot, "Study of heart disease diagnosis by comparing various classification algorithms," International Journal of Engineering and Advanced Technology, vol. 8, no. 2S2, pp. 40-42, 2019.

[10] A. Kashyap, "Artificial intelligence and medical diagnosis," Scholars Journal of Applied Medical Sciences, pp. 4982-4985, 2018.

[11] L. Yahaya, N. D. Oye and E. J. Garba, "A comprehensive review on heart disease prediction using data mining and machine learning techniques," American Journal of Artificial Intelligence, vol. 4, no. 1, pp. 20-29, 2020.

[12] S. Nikhar and A. M. Karandikar, "Prediction of heart disease using machine learning algorithms," International Journal of Advanced Engineering, Management and Science, vol. 2, no. 6, pp. 617-621, 2016.

[13] V. Kirubha and S. M. Priya, "Survey on data mining algorithms in disease prediction," International of Journal of Computer Trends and technology, vol. 38, no. 3, pp. 24-128, (2016.

[14] S. A. Lashari, R. Ibrahim, N. Senan and N. S. A. M. Taujuddin, "Applications of data mining techniques for medical data classification: a review," MATEC Web of Conferences, 2018.

[15] D. Dua and C. Graff, "UCI Machine Learning Repository," Irvine, CA, 2019.

[16] S. Anitha and N. Sridevi, " Heart disease prediction using data mining techniques," Journal of Analysis and Computation, vol. 8, no. 2, pp. 48-55, 2019.

[17] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informatics in Medicine Unlocked, 2019.

[18] K. Subhadra and B. Vikas, "Neural network based intelligent system for predicting heart disease," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 5, pp. 484-487, 2019.

[19] M. Ashraf, M. A. Rizvi and H. Sharma, "Improved heart disease prediction using deep neural network," Asian Journal of Computer Science and Technology, vol. 8, no. 2, pp. 49-54, 2019.

[20] M. A. Ottom and W. Alshorman, "Heart disease prediction using accumulated rank features selection technique," Journal of Engineering and Applied Sciences, vol. 14, no. 7, pp. 2249-2257, 2019.

[21] D. Annepu and G. Gowtham, "Cardiovascular disease prediction using machine learning techniques," International Research Journal of Engineering and Technology, vol. 6, no. 4, pp. 3963-3971, 2019.

[22] A. Jagtap, P. Malewadkar, O. Baswat and H. Rambade, "Heart disease prediction using machine learning," International Journal of Research in Engineering, Science and Management, vol. 2, no. 2, pp. 352-355, 2019.

[23] M. Tarawneh and O. Embarak, "Hybrid approach for heart disease prediction using data mining techniques," Acta Scientific Nutritional Health, vol. 3, no. 7, pp. 147-151, 2019.

[24] T. Nagamani, S. Logeswari and B. Gomathy, "Heart disease prediction using data mining with mapreduce algorithm," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 3, pp. 137-140, 2019.

[25] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," International Journal of Advanced Computer Science and Applications, vol. 10, no. 6, pp. 261-268, 2019.

[26] S. N. Khan, N. M. Nawi, A. Shahzad, A. Ullah and M. F. Mushtaq, "Comparative analysis for heart disease prediction," International Journal on Informatics Visualization, vol. 1, no. 4-2, pp. 227-231, 2019.

[27] A. Sridhar and A. Kapardhi, "Predicting heart disease using machine learning algorithm," International Research Journal of Engineering and technology, vol. 6, no. 4, pp. 36-38, 2018.

[28] M. F. Rabbi, M. P. Uddin, M. A. Ali and M. F. Kibria, " Performance evaluation of data mining classification techniques for heart disease prediction," American Journal of Engineering Research, vol. 7, no. 2, pp. 278-283, 2018.

[29] M. S. Amin, Y. K. Chiam and K. D. Varathan, " Identification of significant features and data mining techniques in predicting heart disease," Telematics and Informatics, 2018.

[30] A. U. Haq, J.-P. Li, M. H. Memon, S. Nazir and R. Sun, " A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," Hindawi Mobile Information System, 2018.

[31] N. Singh and S. Jindal, "Heart disease prediction system using hybrid technique of data mining al-

gorithms," International Journal of Advanced Research, Ideas and Innovations in Technology, vol. 4, no. 2, pp. 982-987, 2018.

[32] U. N. Dulhare, "Prediction system for heart disease using naïve bayes and particle swarm optimization," Biomedical Research, vol. 29, no. 12, pp. 2646-2649, 2018.

[33] D. Kinge and S. K. Gaikwad, "Survey on data mining techniques for disease prediction," International Research Journal of Engineering and technology, vol. 5, no. 1, pp. 630-636, 2018.

[34] H. Benjamin, F. David and S. A. Belcy, "Heart disease prediction using data mining techniques," ICTACT Journal of Soft Computing, vol. 9, no. 1, pp. 1824-1830, 2018.

[35] P. Singh, S. Singh and G. S. Pandi-Jain, "Effective heart disease prediction system using data mining techniques," International Journal of Nanomedicine, 2018.

[36] Y. Khourdifi and M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," International Journal of Intelligent Engineering and Systems, vol. 12, no. 1, 2018.

[37] S. Nandhini, M. Debnath, A. Sharma and Pushkar, "Heart disease prediction using machine learning," International Journal of Recent Engineering Research and Development, vol. 3, no. 10, pp. 39-46, 2018.

[38] M. Gawali and N. Shirwalkar, " Heart disease prediction system using data mining techniques," International Journal of Pure and Applied mathematics, vol. 120, no. 6, pp. 499-506, 2018.

[39] A. Sabay, L. Harris, V. Bejugama and K. Jaceldo-Siegl, "Overcoming small data limitations in heart disease prediction by using surrogate data," 24 December 2018. [Online]. Available: https://scholar.smu.edu/datasciencereview/vol1/iss3/12..

[40] S. M. Satre, S. M. Bhagat and S. M. Thakur, "Heart disease prediction system using data mining," International Journal of Engineering Science and Computing, vol. 8, no. 2, pp. 16087-16089, 2018.

[41] S. Sharmila and M. P. I. Gandhi, " Analysis of heart disease prediction using data mining techniques," International Journal of Advanced Networking and Applications, vol. 8, no. 5, pp. 93-95, 2017.

[42] J. Patel, T. Upadhyay and S. Patel, " Heart disease prediction using machine learning and data mining techniques," IJCSC, vol. 7, no. 1, pp. 129-137, 2016.

[43] Purushottam, K. Saxena and R. Sharma, "Efficient heart disease prediction system," Procedia Computer Science, vol. 85, pp. 962-969, 2016.

[44] S. R. Voleti and K. K. Reddi, " Design of an optimal method for disease prediction using data mining techniques," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6, no. 12, pp. 328-337, 2016.

[45] M. U. Hussein, "Physics and the Cardiovascular System," 29 October 2017. [Online]. Available: https://www.researchgate.net.

[46] K. V. Nagendra and M. Ussenaiah, " A study on various data mining techniques used for heart diseases," International Journal of Recent Scientific Research, pp. 24350- 24354, 2018