International Journal of Mathematical Analysis and Modelling

(Formerly Journal of the Nigerian Society for Mathematical Biology)

Volume 6, Issue 2 (Oct.-Nov.), 2023

ISSN (Print): 2682 - 5694
ISSN (Online): 2682 - 5708

International Journal of Mathematical Analysis and Modelling
Volume 6, Issue 2, October-November 2023, pages 333 – 344

A machine learning model and application for heart disease prediction using prevalent risk factors in Nigeria

F.A. Ekle*†, N.M. Agu†, F.S. Bakpo†, C.N. Udanor†, and A.H. Eneh†

Abstract

Over the past two decades, heart disease remains the principal source of death globally. According to the World Health Organization, about 17.9 million deaths occur every year internationally as a result of heart disease. The vast number of deaths is common amongst low and middle-income countries and Nigeria is at the top on the list of the most affected sub-Saharan African Nations. With the ravaging impact of heart disease in developing countries, there is need to have a reliable, accurate and efficient approach to make an early diagnosis of the disease to achieve prompt therapy or management. The purpose of this research is to develop a machine learning model and application for heart disease prediction using prevalent heart disease risk factors in Nigeria. Quantitative, qualitative, and experimental research methods were used in implementing this research. The experiment for the training and testing of the models were carried out using Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB) and Voting Ensembles Classifier (VEC) with NB, DT, and RF as base learners. The DT, NB, RF and VEC machine learning algorithms gave an accuracy and error rate values of (95.83%, 0.042), (97.22%, 0.028), (97.92%, 0.021) and (96.53%, 0.035) respectively. This research developed a heart disease prediction model and application using RF and Django python web framework respectively. The heart disease dataset used in training the model was gotten from Federal Medical Centre Keffi, Nassarawa State, Nigeria. The developed system acts as a decision support tool for cardiologists.

Keywords and phrases: machine learning; supervised learning machine learning algorithm; heart disease prediction; prevalent risk factors

1 Introduction

According to World Health Organization (WHO), heart disease also referred to as cardiovascular disease is currently the number one killer disease in the world and also three-quarters of the world's deaths from cardiovascular disease occur in low and middle-income nations [17]. This is because it most of the time shows no known symptoms, making it difficult to detect on time, which is why it is called a silent killer. Nigeria is top among the most affected Sub-Saharan African countries and there are projections that the figures will keep rising due to urbanization, poor eating habits, and stressful lifestyles [18]. The currently existing traditional ways of diagnosing cardiovascular disease in Nigeria sometimes cause vague, inaccurate prognoses, and lack prompt prescription for the disease or its risk

*Corresponding Author; E-mail: francis.ekle.pg88249@unn.edu.ng
†Department of Computer Science, University of Nigeria, Nsukka, Nigeria

factors. This is due to human errors, lack of modern-day cardiovascular disease detection equipment, work pressure, and stress on the few available cardiologists.

Several risk factors are associated with cardiovascular disease and it has become imperative to provide an accurate and innovative approach to make an early diagnosis of the risk or presence to achieve prompt treatment or management of the disease. Machine Learning (ML) is a commonly used technique for processing enormous data in the healthcare sector. Researchers in the field of Artificial Intelligence (AI) apply several ML techniques to analyze huge and complex medical data,helping healthcare professionals to predict several diseases [2]. With the large amount of locally available primary cardiovascular disease datasets generated from heart disease risk factors daily in cardiovascular care clinics in teaching hospitals and medical centers across the country, an accurate, efficient, affordable, simple to use, and readily available ML decision support tool for the early detection of cardiovascular disease presence and its risk was developed. This decision support tool is used by cardiologists to reduce the escalating prevalence of morbidities, disabilities, and mortalities due to cardiovascular disease in Nigeria to a manageable level.

2 Literature review

In today's digital and Artificial Intelligence-driven world, several clinical decision support systems on heart disease prediction have been developed to simplify and ensure efficient diagnosis. Das et al. [2] developed a model using machine algorithms such as Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF) and k-Nearest Neighbors (kNN) for prediction of heart diseases. The highest accuracy level was achieved by the kNN classifier, with an accuracy score of 86.84%. Shah et al. [16] developed a ML model for heart disease prediction using 4 ML techniques, kNN ML method achieved the highest accuracy score. Arumugam et al. [1] in a study that developed a fine-tuned ML model for forecasting the likelihood of heart disease in diabetic individuals used Naïve Bayes (NB), SVM, and DT ML techniques. Olaniyi et al. [10] developed and implemented an intelligent heart disease prediction system that is modeled on feed forward multilayer perceptron Artificial Neural Network (ANN) and SVM. Their findings showed that SVM gives a better result with an accuracy score of 87.5% using the Cleveland Heart Disease University of California Institute Repository Dataset (CHDUCIRD). DT produced the highest accuracy score of 90.5%.

Desai et al. [3] developed a system for monitoring the health status of heart disease patients using SVM, kNN, ANN, Logistic Regression (LOR), and Gradient Boosting Trees (GBT) ML algorithms and cloud computing technology. The LOR ML model gave the best accuracy score of 85.96% and was used in developing the system using the CHDUCIRD dataset. Gupta et al. [9] developed a heart disease prediction models using LOR, SVM, NB, DT, kNN and RF supervised learning ML algorithms and the CHDUCIRD dataset. LOR produced the best model with an accuracy score of 92.30%. Qureshi and Warke [14] developed heart disease prediction models using RF, ANN, NB, and DT supervised learning ML algorithms. The RF model produced the best model with an accuracy score of 82.25% using the CHDUCIRD dataset.

Sarah et al. [15] carried out a comparative study on the performance of supervised learning ML algorithms for heart disease prediction using LOR, SVM, DT, kNN, NB, and RF supervised learning ML algorithms. LOR and SVM algorithms produced the best models with an accuracy score of 85.25% respectively. Dutta et al. [5] carried out a comparative study on the performance of kNN, DT, XG Boost (XGB), and RF ML algorithms in the prediction of heart disease. kNN produced the best accuracy score of 85.71% using the CHDUCIRD dataset.

Based on the reviewed literature above, none of the previous works used Voting Ensembles Classifier (with NB, DT and RF as the base classifiers) ML algorithm in predicting the presence of heart disease. Secondly, health data is demographic in nature [8], but from the extant literature, most of the existing projects carried out in the prediction of heart disease using ML algorithms and risk factors causing heart disease used CHDUCIRD, a secondary dataset from University of California Institute dataset database existing online that were gotten from other climes. Therefore, this research is based basically on locally available primary heart disease dataset and its risk factors gotten from Nigeria.

3 Methodology

Quantitative, qualitative, and experimental research methodologies were adopted in implementing this research. The experiment for the training and testing of the models were carried out using Decision tree (DT), Random Forest (RF), Naïve Bayes (NB) and Voting Ensembles Classifier (VEC) with NB, DT, and RF as base classifier algorithms as shown in the flowchart in figure 1. The performances of the algorithms were measured by the accuracy and error rate gotten from Confusion Matrix.

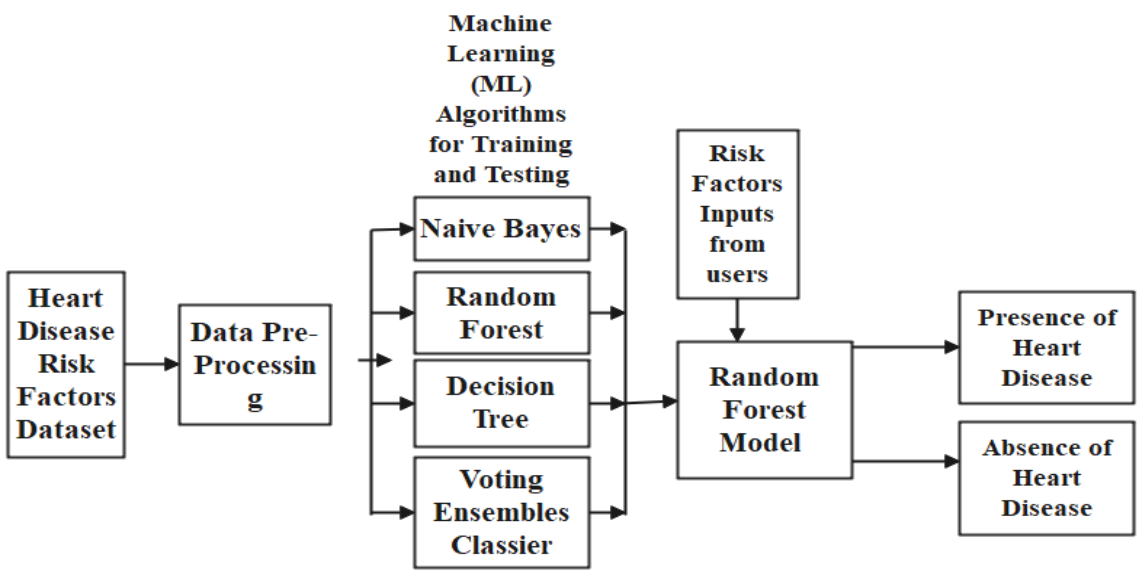

Figure 1: Flowchart showing the flow of activities in implementing the machine learning model for heart disease prediction

Decision Tree (DT): According to Shah et al. [16], DT is a frequently used data mining ML method that has the ability to solve classification and regression problems, but is more commonly used for resolving classification tasks. This algorithm categorizes a population into branch-like segments that construct a reversed tree with a root node, internal nodes, and leaf nodes. The ML method is non-parametric and can competently deal with large, complex, and complicated data without imposing a complex parametric structure. DT learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a DT. Learned trees can also be shown as sets of if-then rules to improve human legibility. The DT model makes analysis based on three basic nodes, namely:

- Root node: principal node, based on this node all-other nodes functions.
- Interior node: handles various features or attributes of the dataset.
- Leaf node: represent the outcome of each test, i.e., the class of the dependent variable in classification problems.

The DT algorithm divides the data into two or more analogous sets based on the most significant indicators. The entropy of each feature is computed, and then the dataset is divided, with predictors having the minimum entropy. The formula for calculating the entropy of an attribute is shown in Eq. (1).

$$Entropy(s) = \sum_{i=1}^{c} -p_i \log_2 p_i \qquad (1)$$

where in Eq. (1), c is the total number of classes and the probability of samples belonging to a class at a given node can be denoted as p_i. A simple basic structure of a DT showing the three basic nodes is shown in figure 2.

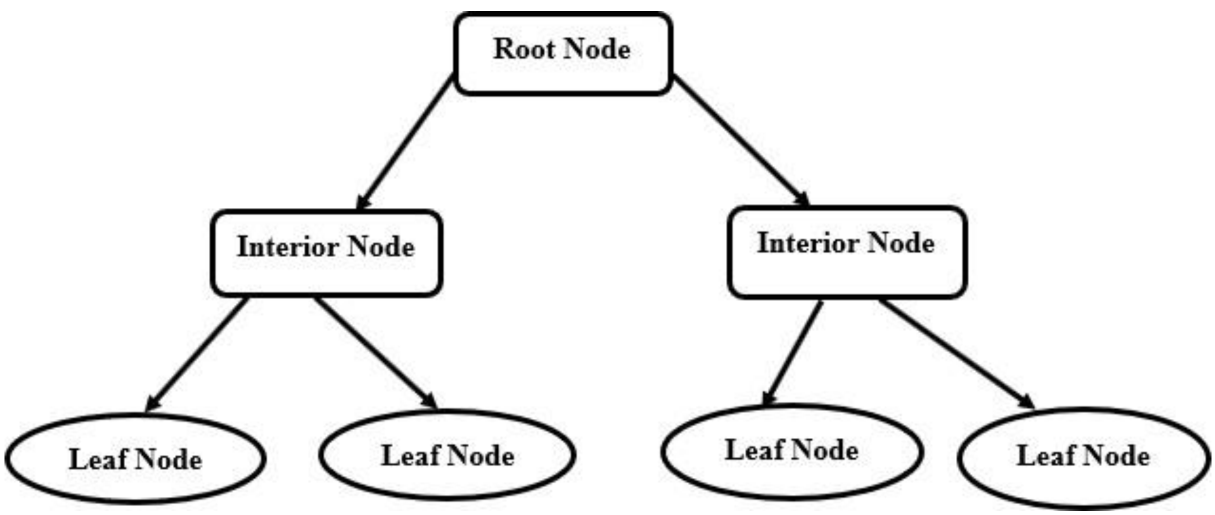

Figure 2: Structure of a Decision Tree showing the root node, interior nodes represented by a rectangle, and leaf nodes represented by an oval shape.

Random Forest (RF): RF is an ensemble ML algorithm. This algorithm works by combining many DT into a forest [11]. In the RF algorithm, each separate tree outputs a class expectation, and the class with the most votes becomes the model's forecast. The greater the number of trees in a RF classifier, the better the chance of increased accuracy. It can be used for both classification and regression tasks, although it excels more at classification problems and can overcome the shortcomings of DT and other ML methods, such as overfitting and missing values.

Naive Bayes (NB): According to Diwakar et al. [4], NB is a classification and regression supervised learning ML technique based on Bayes' theorem with the assumption of strong (Naive) independence amongst predictors. In simple terms, an NB classifier believes that the existence of one attribute in a class is unrelated to the presence of any other attribute. For example, if a fruit is green, spherical, and roughly 4 inches in diameter, it is considered an orange. Even if these attributes depend on each other or upon the existence of the other attributes, a NB classification technique would consider all of these characteristics to independently contribute to the probability that this fruit is an orange. NB is a simple, easy-to-implement, and efficient algorithm that handles non-linear and complex data. Mathematically, the Bayes theorem is written as shown in Eq. 2.

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)} \tag{2}$$

where in Eq. (2), $P(A/B)$ is the posterior probability, $P(B)$ is the predictor prior probability, $P(A)$ is the class prior probability, $P(B/A)$ is the possible probability of predictor.

Voting Ensembles Classifier (VEC): Similarly, to the RF ML algorithm, the VEC classifier estimates several base models. It uses voting to combine the separate independent predictions or classifications from the various classifiers to arrive at the final results. However, the main difference lies in the base models. The VEC model does not mandate or necessitate the base models to be homogenous. Put differently, we can train different base learner ML models to arrive at the final result [6]. Both hard and soft voting mechanisms are implemented in the VEC algorithm. We forecast the final class label using the hard voting approach as the class label that has been forecasted the most frequently by the classification trained models. Soft voting, on the other hand, predicts class labels by averaging class probabilities. Mathematically, hard voting is represented by Eq. (3).

$$\hat{y} = \operatorname{mod} e\{A_1(x), A_2(x), A_3(x)\} \tag{3}$$

In Eq. (3), $\hat{y}$ is the target variable class label. A_1, A_2 and A_3 are they base classifiers used in the VEC classifier respectively. x is the data points or records. Figure 3 shows the framework of the VEC algorithm using NB, RF, and DT as the base models or classifiers.

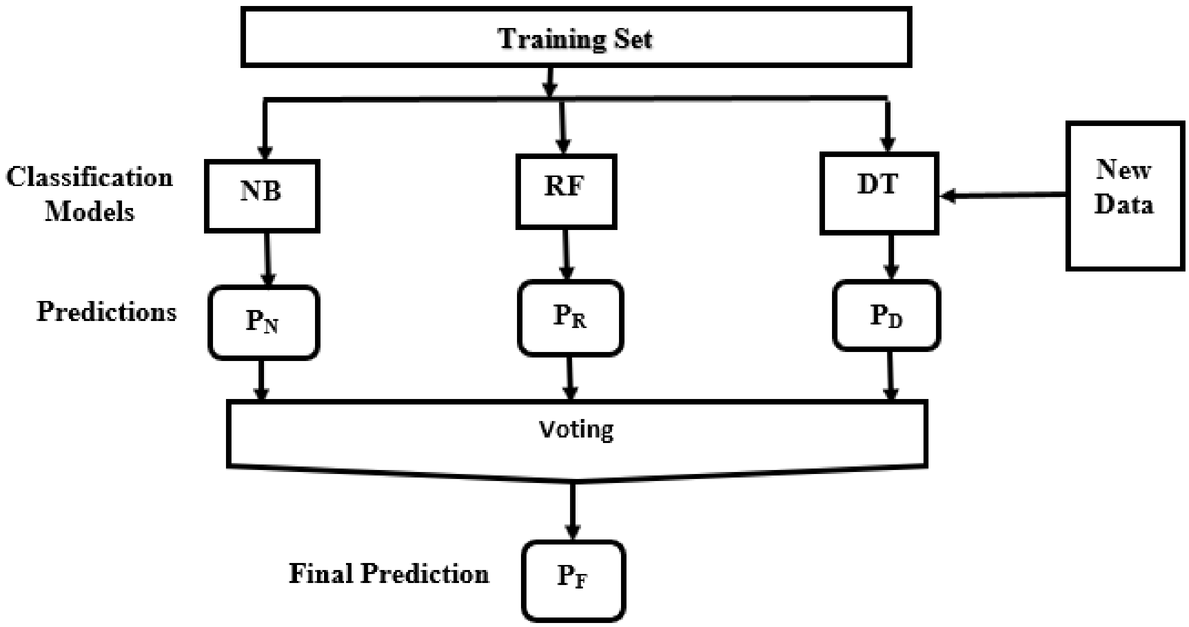

Figure 3: Figure showing the framework of the VEC algorithm using NB, RF and DT asbase models.

In figure 3, P_N is the prediction value for NB classifier, P_R is the prediction value for RF classifier, P_D is the prediction value for DT classifier and P_F is the final prediction for the VEC model. Training Set is the dataset used for training the model, while New Data is the dataset used for testing the model.

4 Data preprocessing

The heart disease dataset used for this thesis was collected from heart disease clinic in Federal Medical Centre Keffi. The collected dataset is made up of 720 instances and 13 variables. 12 are independent variables while 1 is a dependent variable. The heart disease risk factors inputs data considered for this research are; sex, age, hypertension status, weight, height, physical activity level, smoking habit, alcohol consumption habit, diabetes status, drug abuse status, genetic history and dyslipidaemia status as shown in details in table 1. The format of the heart disease dataset gotten from Federal Medical Centre Keffi is shown in table 1 below.

Table 1: Attributes and details of dataset of heart disease gotten from Federal Medical Centre Keffi.

Sr.no.	Attribute	Details
1	Sex	1= male; 0=female
2	Age	Patients age in years
3	Diabetic	1-Yes; 0-No
4	Weight	Patients weight in kilogram(kg)
5	Height	Patients height in metres(m)
6	Hypertensive	1-Yes; 0-No
7	Dyslipidaemia	Level of good or bad lipids or cholesterol. 1-Yes; 0-No
8	Genetic history	2=Parent; 1=Other than parents; 0=None

9	Abuse drugs	1-Yes; 0-No
10	Smoke	3=Daily; 2=Occasionally; 1=Used to Not now; 0=Never
11	Alcohol	3=Significantly; 2=Occasionally; 1=Used to Not now; 0=Never
12	Physical activity level	0=Active; 1=Moderately Active; 2=Sedentary
13	Heart disease prediction	0=Absence of heart disease; 1=Presence of heart disease

4.1 Applying Pearson Correlation

In this section, the correlations between attributes of the FMC Keffi dataset and heart disease prediction are analyzed and applicability to the classification analysis is examined. Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. The correlation between two variables is a measure of how well the variables are related. The most common measure of correlation in statistics is the Pearson Correlation which shows the linear relationship between two variables. Pearson correlation coefficient analysis produces a result between -1 and 1. A result of -1 means that there is a perfect negative correlation between the two values, while a result of 1 means that there is a perfect positive correlation between the two variables. Results between 0.5 and 1.0 indicate high correlation. For the purpose of this research, independent variables of the heart disease dataset (risk factors) with Pearson correlation coefficient values from, 0.3 (30%) and above with the dependent variable (Heart disease prediction) shown in the Seaborn Heatmap visualizing FMC Keffi data correlations (Figure 4) is used for training the model.

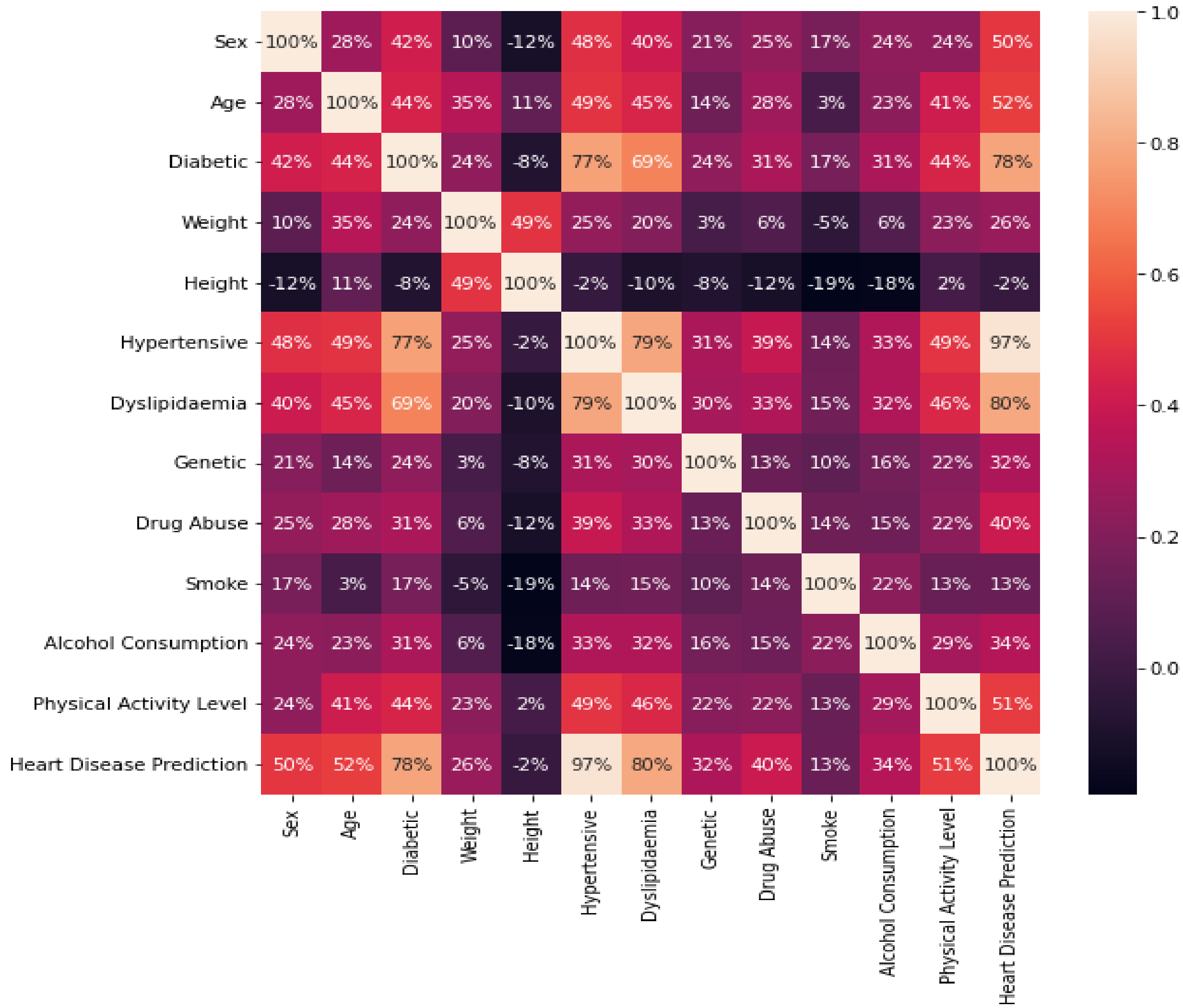

Figure 4: Seaborn Heatmap visualizing FMC Keffi data correlations

5 Model evaluation

For evaluating the model, the dataset is divided into 2 parts. 80 percent was used in training the model, while the remaining 20 percent was used in testing the model.

5.1 Evaluation metrics

The evaluation metrics used for this research such as accuracy and error rate were gotten from the confusion matrix table shown in table 2. In the field of machine learning and precisely the problem of classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of machine learning algorithm. The template for any binary confusion matrix is made up of the followings; true positives, false negatives, false positives, and true negatives along with the positive and negative classifications as shown in table 2 [7][12-13].

Table 2: Binary Confusion Matrix.

	Predicted Condition	
Total Population= P+N	Positive (PP)	Negative (PN)
Positive (P)	True Positive (TP)	False Negative (FN)
Negative (N)	False Positive (FP)	True Negative (TN)

From table 2 above, condition Positive (P) is the number of real positive cases in the data, condition Negative (N) is the number of real negative cases in the data, True Positive (TP) a test result that correctly indicates the presence of a condition or characteristic, True Negative (TN) a test result that correctly indicates the absence of a condition or characteristic, False Positive (FP) a test result which wrongly indicates that a particular condition or attribute is present, False Negative (FN) a test result which wrongly indicates that a particular condition or attribute is absent.
The performance metrics used in this thesis are;

Accuracy: Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0. It can also be calculated by 1 – ERR. The best accuracy in percentage is 100%. Mathematically it is shown in Eq. (4).

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \qquad 4$$

Error Rate: Error rate (ERR) is calculated as the number of all incorrect predictions divided by the total number of the dataset. The best error rate is 0.0, whereas the worst is 1.0. Mathematically it is shown in Eq. 5.

$$ERR = \frac{FP + FN}{TP + TN + FN + FP} \qquad 5$$

6 Results

This thesis implemented the ML based model using Python programming language. It trained they ML algorithms used in this thesis with heart disease risk factors and their corresponding output datasets gotten from Federal Medical Centre, Keffi. The dataset was divided into two parts, 80 percent for training and 20 percent for testing. The accuracy score gotten from the testing of the various ML algorithms used in this thesis is shown in the Bar graph in figure 5. RF ML algorithm gave the most accurate result with an accuracy score of 97.92 %. RF was used in developing the heart disease prediction application because of its high accuracy result as shown in table 3. While, the output, is the presence or absence of heart disease. The application development was carried out using Django, a python web development framework.

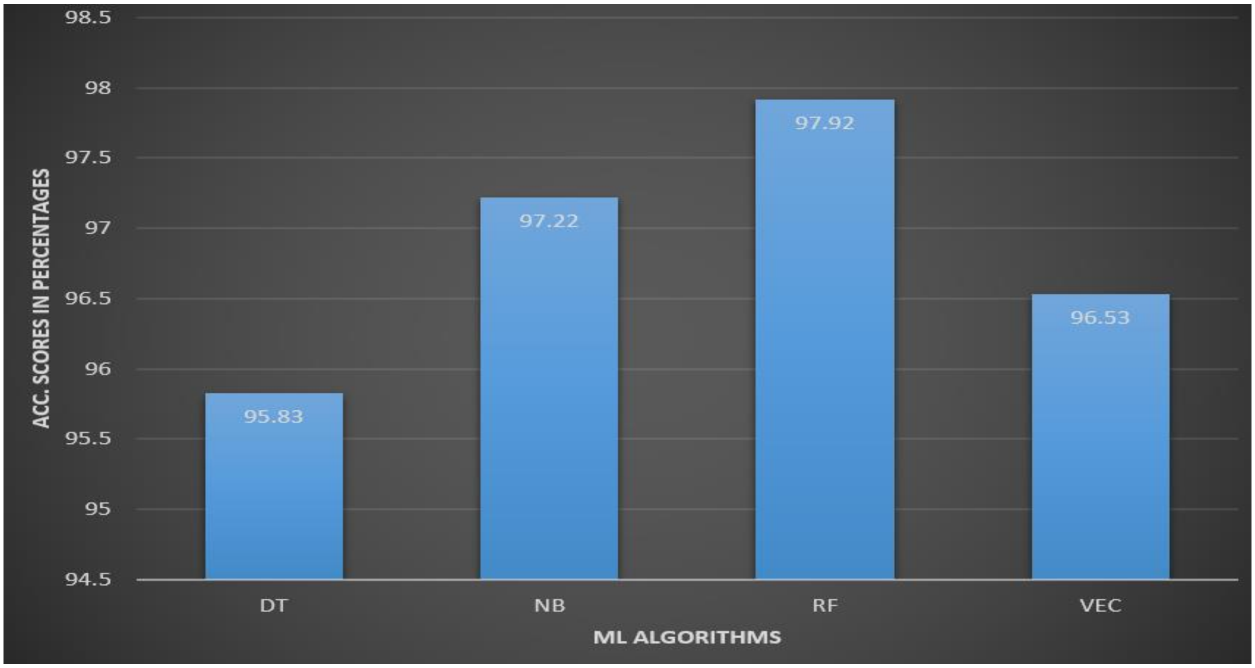

Figure 5: A Bar Chart showing the Comparison of ML Algorithms in terms of Accuracy Score in Percentage after Testing of the Models.

Table 3: Comparison of the performance of all the 4 machine learning algorithms used in this research.

Algorithms	Accuracy (%)	Error Rate	Training Time (s)
Decision Tree (DT)	95.83	0.042	0.02
Naïve Bayes (NB)	97.22	0.028	0.01
Random Forest (RF)	97.92	0.021	0.11
Voting Ensembles Classifier (VEC) with DT, RF and NB as base classifiers	96.53	0.035	0.02

7 Conclusion

This research used the RF trained machine learning model for implementing the heart disease prediction application. The heart disease prediction application was developed using the django python web framework. The heart disease dataset used in training the model was gotten from Federal Medical Centre Keffi, Nassarawa State, Nigeria. The developed system acts as a decision support tool for cardiologists.

Further works can be carried out in the area of heart disease dataset collection by involving more heart disease clinics in different geographical locations and regions or clime.

References

[1] Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2021). Multiple disease prediction using Machine learning algorithms. Materials Today: Proceedings.

[2] Das, S., Sanyal N. M., & Upadhyay, S. K. (2020). A Comparative Study for Prediction of Heart Diseases using Machine Learning. Proceedings of Industry Interactive Innovations in Science, Engineering & Technology.

[3] Desai, F., Chowdhury, D., Kaur, R., Peeters, M., Arya, R. C., Wander, G. S., ... & Buyya, R. (2022). Health Cloud: A system for monitoring health status of heart patients using machine learning and cloud computing. Internet of Things, 17:100485.

[4] Diwakar, M., Tripathi, A., Joshi, K., Memoria, M., & Singh, P. (2021). Latest trends on heart disease prediction using machine learning and image fusion. Materials Today: Proceedings, 37:3213-3218.

[5] Dutta, P., Paul, S., Shaw, N., Sen, S., & Majumder, M. (2022). Heart Disease Prediction: A Comparative Study Based on a Machine-Learning Approach. In Artificial Intelligence and Cybersecurity (pp. 1-18). CRC Press.

[6] Emakhu, J., Shrestha, S., & Arslanturk, S. (2020). Prediction System for Heart Disease Based on Ensemble Classifiers. In The 5th NA International Conference on Industrial Engineering and Operations Management Detroit, 10-14.

[7] Fawcett, T. (2006). Introduction to receiver operator curves. Pattern Recognit. Lett, 27:861-874.

[8] Grundy, E., & Murphy, M. J. (2015). Demography and public health. In Oxford textbook of Global public health, 6th ed., Oxford: Oxford University Press.

[9] Gupta C, Saha A, Reddy N. S., Acharya U. D. (2022) Cardiac Disease Prediction using Supervised Machine Learning Techniques. In Journal of Physics: Conference Series 2022 (Vol. 2161, No. 1, p. 012013). IOP Publishing.

[10] Olaniyi, E. O., Oyedotun, O. K., & Adnan, K. (2015). Heart diseases diagnosis using neural networks arbitration. International Journal of Intelligent Systems and Applications, 7(12):72.

[11] Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. Heart Disease, 7(1):129-137.

[12] Piryonesi, S. M., & El-Diraby, T. E. (2020). Data analytics in asset management: Cost-effective prediction of the pavement condition index. Journal of Infrastructure Systems, 26(1):04019036.

[13] Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.

[14] Qureshi, M., & Warke, N. (2022). Application of Machine Learning for Heart Disease Prediction. In Proceedings of 2nd International Conference on Artificial Intelligence: Advances and Applications (pp. 267-279). Springer, Singapore.

[15] Sarah, S., Gourisaria, M. K., Khare, S., & Das, H. (2022). Heart Disease Prediction Using Core Machine Learning Techniques—A Comparative Study. In Advances in Data and Information Sciences (pp. 247-260). Springer, Singapore.

[16] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1(6):1-6.

[17] World Health Organization. (2019, June 11). WHO updates Cardiovascular Risk Charts. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[18] World Health Organization. (2021, June 11). Cardiovascular Diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)