SN Computer Science (2021) 2:350
https://doi.org/10.1007/s42979-021-00731-4

ORIGINAL RESEARCH

Machine Learning Predictive Models for Coronary Artery Disease

L. J. Muhammad[1] · Ibrahem Al-Shourbaji[2] · Ahmed Abba Haruna[3] · I. A. Mohammed[4] · Abdulkadir Ahmad[5] · Muhammed Besiru Jibrin[1]

Received: 27 May 2021 / Accepted: 31 May 2021 / Published online: 22 June 2021

Abstract

Coronary artery disease (CAD) is the commonest type of heart disease and over 80% of the deaths resulted from the diseases occurred in developing countries including Nigeria, with majority being in those victims are below 70 years of age. Though, CAD is not a well known disease in Nigeria but however in year 2014, 2.82% of the total of deaths occurred in the country were due to the disease. In this study, a machine leaning predictive models for CAD has been developed with diagnostic CAD dataset obtained in the two General Hospitals in Kano State—Nigeria. The dataset applied on machine learning algorithms which include support vector machine, K nearest neighbor, random tree, Naïve Bayes, gradient boosting and logistic regression algorithms to build the predictive models and the models were evaluated based accuracy, specificity, sensitivity and receiver operating curve (ROC) performance evaluation techniques. In terms of accuracy random forest-based machine learning model emerged to be the best model with 92.04%, for specificity Naive Bayes based machine learning model emerged to be the best model with 92.40%, while for sensitivity support vector machine based machine learning model emerged to be the best model with 87.34% and for ROC, random forest-based machine learning model emerged to be the best model with 92.20%. The decision tree generated with random forest machine learning algorithm which happened to be best model in terms accuracy and ROC can be converted into production rules and be used develop expert system for diagnosis of CAD patients in Nigeria.

Keywords CAD · Machine learning · Disease · Predictive model

Introduction

Coronary artery disease (CAD) has become a common disease that is affecting people globally, especially in developed countries. According to the world's health organization statistics, cardiovascular diseases have contributed to over 31% of the mortality rate globally [6, 17–19]. Though CAD is not

This article is part of the topical collection "Advances in Computational Approaches for Artificial Intelligence, Image Processing, IoT and Cloud Applications" guest edited by Bhanu Prakash K. N. and M. Shivakumar

✉ L. J. Muhammad
lawan.jibril@fukashere.edu.ng

Ibrahem Al-Shourbaji
alshourbaji@ut.edu.sa

Ahmed Abba Haruna
ahmadydee@gmail.com

I. A. Mohammed
ibrahimsallau@gmail.com

Abdulkadir Ahmad
aayola99@gmail.com

Muhammed Besiru Jibrin
bashmjibrin@gmail.com

[1] Department of Mathematics and Computer Science, Faculty of Science, Federal University of Kashere, P.M.B. 0182, Gombe, Nigeria

[2] Department of Computer Science and Network Engineering, Jazan University, 828222-6649 Jazan, Kingdom of Saudi Arabia

[3] College of Computer Science and Engineering, University of Hafr Al Batin, Al Jamiah, Hafar Al Batin 39524, Saudi Arabia

[4] Computer Science Department, Yobe State University, Damaturu, Yobe State, Nigeria

[5] Department of Computer Science, Kano University of Science and Technology, Wudil, Kano, Nigeria

well known disease in Nigeria, in year 2014, 2.82% of the total of deaths occurred in country were due to the disease [28]. Furthermore, many recent reports and studies indicate that CAD is now frequently recognized in Nigeria. According World Health Organization (WHO), more than half a million Nigerians died in 2012 from Non-Communicable Diseases and every one Nigerian adult out five Nigerian adults over the age of 30 is likely die from prematurely from Non-Communicable Diseases and coronary artery disease is inclusive [19, 29].

Therefore, it is very paramount and essential to accurately detect or predict the disease form the infected patients at earlier stage in order to prevent and minimize lost lives due to the disease. Many artificial intelligence techniques such as data mining, machine learning, deep learning and expert system are being used in healthcare industry for diagnosis, detection and prediction of many diseases such as diabetes, waterborne, COVID-19, Malaria and typhoid among others.

However, machine learning (ML) is one of the suitable techniques for development models that are being used for diagnosis of diseases in the healthcare industry [24]. ML is an artificial intelligence concept that is being used to build the models or system that can learn the existing dataset to predict future event [14]. Machine learning is automatically discovering useful information and identifying hidden patterns in large data warehouses [16]. ML involves few phases from raw data collection to some interesting patterns, and this process includes data cleaning, transformation, selection, evaluation and knowledge presentations to provide users with explored knowledge [13]. ML algorithms have been used in the health care industry to get the meaningful insights for better diagnosic decision making [3, 5, 12]. Besides, it has helped the systems to learn the diagnosis data, identify useful patterns during the learning process and minimize human interference in making decisions [24].

The study of [2] used machine learning techniques with Iranian patients' data to accurately detect CAD disease. In another work of [4], a personalized treatment system for coronary artery diseases using the machine learning approach was developed and the system provided an interactive interface to health care professionals with accurate, useful, and readily analytics information.

In the study of [22] the obstructive predictive model for coronary artery diseases was developed with machine learning algorithms and result of the study showed that the models perform efficiently. The study of [25], investigated the ensemble of heterogeneous classifiers for diagnosis of CAD. The authors combined three ML methods: K-nearest neighbour (KNN), random forest (RF) and support vector machine (SVM) for diagnosis of CAD. The final results showed that the proposed ensemble method diagnose CAD efficiently. The work of [1], compared the SVM performance and Artificial Neural Networks (ANN) to predict CAD. The researcher concluded that SVM algorithm results were higher in accuracy and better performance than ANN, while a higher sensitivity and power characterized the ANN.

In this work, machine learning predictive models for coronary artery disease have been developed with various ML algorithms which include support vector machine, K nearest neighbor, random tree, Naïve Bayes, gradient boosting and logistic regression.

Methods and materials

Figure 1 shows methods and materials of the study.

Dataset

The medical expert diagnostic dataset for coronary artery disease was obtained in the two General Hospitals in Kano State—Nigeria: Murtala Mohammed General Hospital and Abdullahi Wase General Hospital. The dataset was collected with the approval of the Research Ethical Committee of Kano State Ministry of Health, Nigeria. The medical expert diagnostic dataset for coronary artery disease patients between 2003 and 2017 was considered and collected for the purpose of the study.

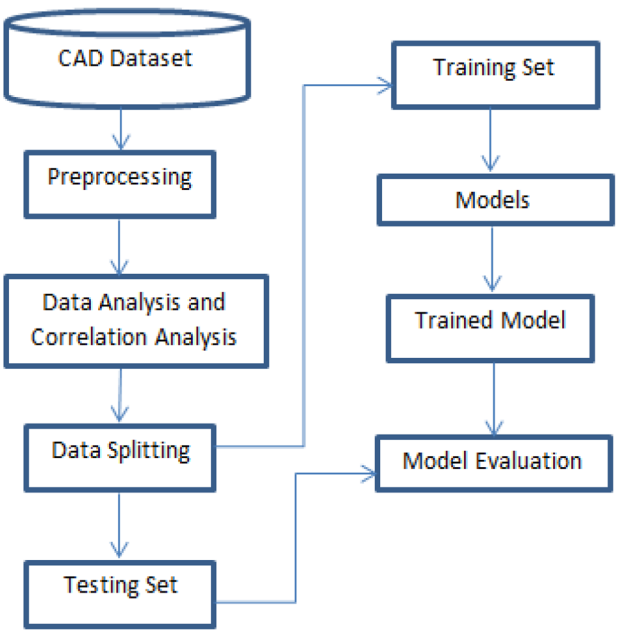

Fig. 1 Study methods and materials

Data Preparation and Analysis

The medical expert diagnostic datasets for coronary artery disease obtained in the two General Hospitals in Kano State—Nigeria were prepared in the appropriate format with the help of medical experts in the hospitals and only data instances of the dataset without missing values were considered and collected. Therefore, there are only 506 data instances of the dataset without missing value. The dataset is labeled one with 18 features including demographic, history and clinical features of the patient's CAD. The feature of the dataset are age, sex, CAD family history, smoking, type of the chest pain, diabetes, glucose, hypertension, blood pressure, cholesterol, Hyperlipidemia, high-density lipoprotein (HDL), Triglyceride, low-density lipoprotein (LDL), Creatinine, BodyMass, HeartRate and Diagnosis. Table 1 shows the description of the dataset features. Figure 2 shows the description of the features of the dataset including data type and count of the non-null of each feature, Figure 3 shows the dataset sample, Figure 4 shows the line graph of the profile information of the features of the dataset. The profile information describes the minimum value, maximum value, mean value and standard deviation values of each feature of the dataset, Figure 5 shows the frequency of the age of the CAD patients, Fig. 6 shows the frequency of CAD family history of the patients, Fig. 7 shows frequency of the body mass of the CAD patients and Fig. 8 shows the frequency of the CAD diagnosis of the patients.

```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             506 non-null    int64
 1   Sex             506 non-null    int64
 2   FamilyHistory   506 non-null    int64
 3   Smoking         506 non-null    int64
 4   ChestPain       506 non-null    int64
 5   Diabetes        506 non-null    int64
 6   Glucose         506 non-null    int64
 7   Hypertension    506 non-null    int64
 8   BLoodPressure   506 non-null    int64
 9   Cholesterol     506 non-null    int64
 10  Hyperlipidemia  506 non-null    int64
 11  HDL             506 non-null    int64
 12  Triglyceride    506 non-null    int64
 13  LDL             506 non-null    int64
 14  Creatinine      506 non-null    float64
 15  BodyMass        506 non-null    int64
 16  HeartRate       506 non-null    int64
 17  Diagnosis       506 non-null    int64
dtypes: float64(1), int64(17)
```

Fig. 2 Data type description of the dataset features

Correlation Analysis of the Dataset features

Correlation coefficient analysis was carried out on the

Table 1 Description of the dataset features

SN	Feature	Units	Range
1	Age	Years	1–150
2	Sex	Male (1), female (0)	0.1
3	Family history	Yes (1), no (0)	0.1
4	Smoking	Yes (1), no (0)	0.1
5	Diabetes	Yes (1), no (0)	0.1
6	Hypertension	Yes (1), no (0)	0.1
7	Hyperlipimedia	Yes (1), no (0)	0.1
8	Blood pressure	mmHg	90–190
9	Glucose	mg/dL	37–295
10	Cholesterol	mg/dL	128–575
11	Triglyceride	mg/dL	40–690
12	HDL	mg/dL	10.6–73
13	LDL	mg/dL	10–220
14	Creatinine	mg/dL	0.6–3.3
15	Body mass index	kg/m^2	20.28–40.25
16	Heart rate	Bpm	42–124
17	Chest pain	Typical angina (4), atypical angina(3), non-anginal pain(2), asymptomatic (1)	1–4
18	Diagnosis of CAD	Positive (1), negative (2)	0,1

NB mmHg millimeters of mercury, *mg/dL* milligrams per deciliter, kg/m^2 kilogram-meter squared, *Bpm* beats per minute

	Age	Sex	FamilyHistory	Smoking	ChestPain	Diabetes	Glucose	Hypertension	BLoodPressure	Cholesterol	Hyperlipidemia	HDL	Triglyceride	LDL	Creatinine	BodyMass	HeartRate	Diagnosis
0	61	1	1	1	4	1	201	1	176	190	0	21	319	57	2.6	32	110	1
1	45	1	0	0	1	0	107	0	120	129	0	11	250	190	0.8	11	78	0
2	32	1	1	1	2	1	181	1	160	401	1	78	78	231	1.0	13	30	1
3	67	0	1	3	1	1	178	1	160	140	0	13	56	78	3.0	13	50	1
4	56	1	1	1	1	0	78	0	140	216	0	43	300	134	2.0	43	50	0

Fig. 3 Sample of the dataset

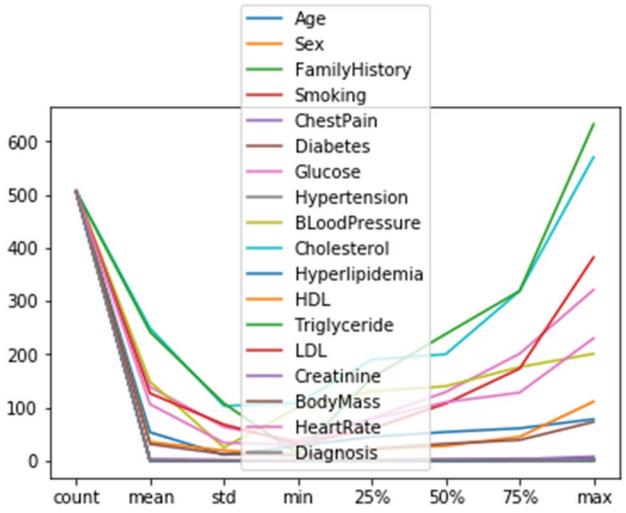

Fig. 4 Profile information of the dataset

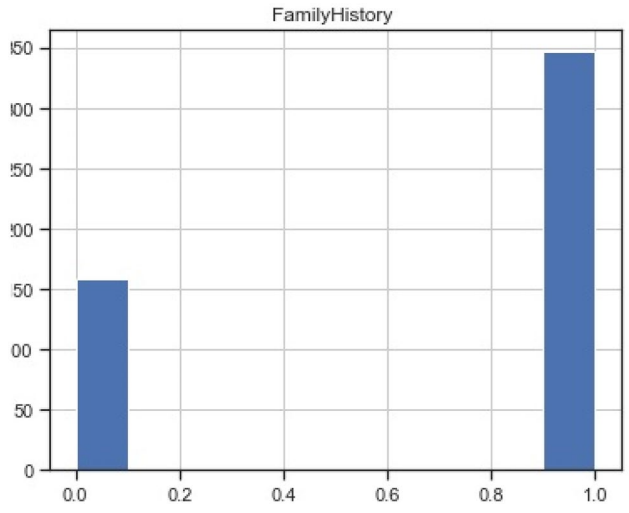

Fig. 6 Frequency of CAD family history of the patients

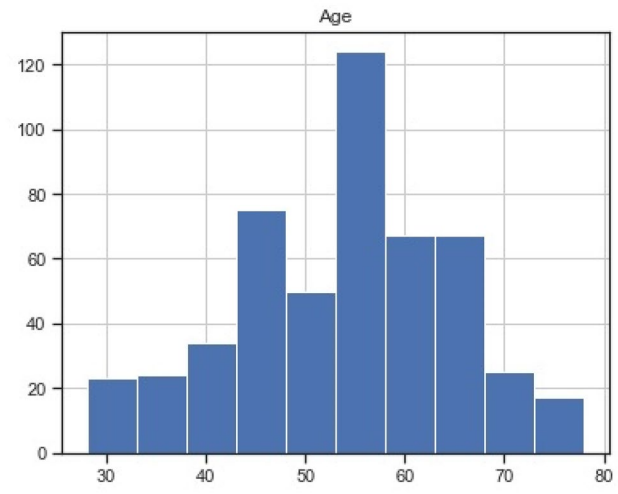

Fig. 5 Frequency of age of the CAD patients

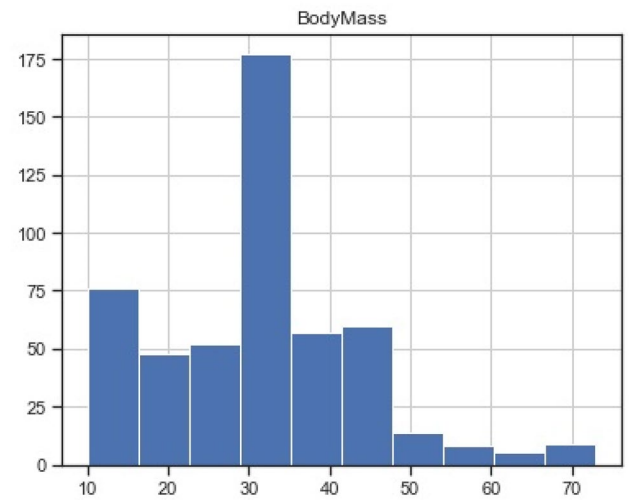

Fig. 7 Frequency of the body mass of the patients

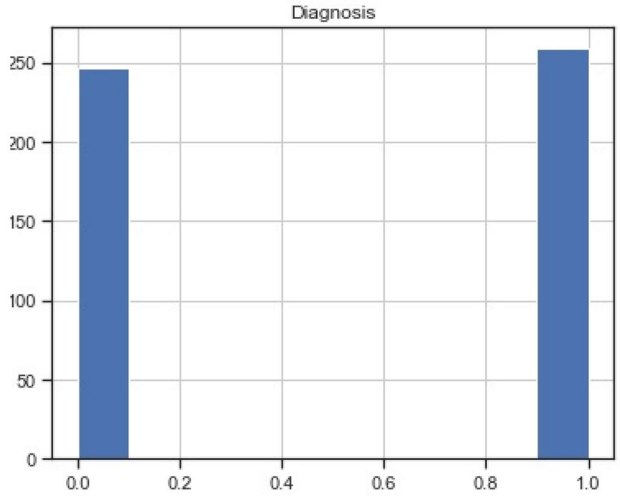

Fig. 8 Frequency of the CAD diagnosis of the patients

dependent features against independent feature of the CAD Dataset [16]. Correlation coefficient is used to determine the strength relationship that exists between the dependent features against independent feature which can either be positive or negative. The *r* value is a set of infinite number between − 1 to + 1 which shows the existing relationship either positive or negative between the dependent features against independent feature [7]. Therefore, if the *r* number is positive, it shows the relationship exists between the dependent and independent feature is positive while if the r number is negative, it shows the relationship exists between the dependent and independent feature is negative. A feature set is considered good for ML model if the dependent features are correlated with the independent features [26]. The feature can be evaluated by Eq. (1) as follows:

$$\text{Importance} = \frac{\overline{k\text{avg}\left(\text{corr}_{\text{fc}}\right)}}{\sqrt{k + k(k-1)\overline{\text{avg}\left(\text{corr}_{\text{ff}}\right)}}}, \tag{1}$$

where the importance is the correlation coefficient between dependent feature set and independent feature and is the ranking criteria for evaluating the set of feature, $(\text{avg}(\text{corr}_{\text{fc}}))$ is the average of the correlation between the dependent feature and the independent feature, $\text{avg}(\text{corr}_{\text{ff}})$ is the average of the correlation between feature set and k is the number of features.

Correlation coefficient analysis was carried out on the dependent features of the CAD dataset which include age, sex, family history, smoking, chest pain, diabetes, glucose, hypertension, blood pressure, cholesterol, hyperlipidemia HDL, triglyceride, LDL, creatinine, body mass and heart rate and diagnosis feature which is an independent features of the CAD dataset. Table 2 and shows the r value of dependent feature against the independent feature of the dataset while Fig. 9 shows entire the correlation coefficient analysis matrix of the dataset features.

The result of the correlation coefficient analysis of the dataset shows that age feature has *r* value of 0.42 of moderate uphill positive correlation coefficient relationship, sex feature has *r* value of 0.50 of moderate uphill positive correlation coefficient relationship, family history feature has *r* value of 0.48 of moderate uphill positive correlation

Table 2 *r* value of the correlation coefficient analysis

SN	Dependent feature	Independent feature	*r* value	Correlation coefficient relationship
1	Age	Medical diagnostic result	0.42	Moderate uphill positive correlation coefficient relationship
2	Sex	Medical diagnostic result	0.50	Moderate uphill positive correlation coefficient relationship
3	Family history	Medical diagnostic result	0.48	Moderate uphill positive correlation coefficient relationship
4	Smoking	Medical diagnostic result	0.24	Weak uphill positive correlation coefficient relationship
5	Chest pain	Medical diagnostic result	0.58	Moderate uphill positive correlation coefficient relationship
6	Diabetes	Medical diagnostic result	0.61	Strong uphill positive correlation coefficient relationship
7	Glucose	Medical diagnostic result	0.55	Moderate uphill positive correlation coefficient relationship
8	Hypertension	Medical diagnostic result	0.65	Strong uphill positive correlation coefficient relationship
9	Blood pressure	Medical diagnostic result	0.53	Moderate uphill positive correlation coefficient relationship
10	Cholesterol	Medical diagnostic result	0.44	Moderate uphill positive correlation coefficient relationship
11	Hyperlipidemia	Medical diagnostic result	−0.50	Moderate uphill negative correlation coefficient relationship
12	HDL	Medical diagnostic result	−0.20	Weak uphill negative correlation coefficient relationship
13	Triglyceride	Medical diagnostic result	0.28	Weak uphill positive correlation coefficient relationship
14	LDL	Medical diagnostic result	0.35	Moderate uphill positive correlation coefficient relationship
15	Creatinine	Medical diagnostic result	0.40	Moderate uphill positive correlation coefficient relationship
16	Body mass	Medical diagnostic result	0.50	Moderate uphill positive correlation coefficient relationship
17	Heart rate	Medical diagnostic result	0.53	Moderate uphill positive correlation coefficient relationship

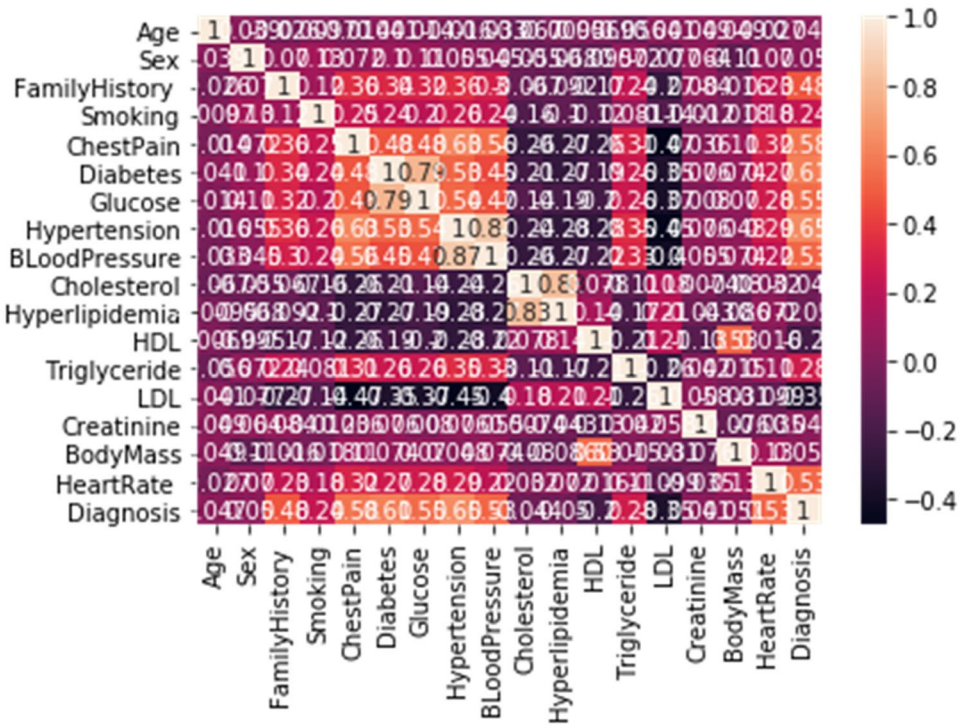

Fig. 9 The correlation coefficient analysis matrix of the dataset features

coefficient relationship, smoking feature has *r* value of 0.24 of moderate uphill positive correlation coefficient relationship, chest pain feature has *r* value of 0.58 of moderate uphill positive correlation coefficient relationship, diabetes feature has *r* value of 0.61 of strong uphill positive correlation coefficient relationship, glucose feature has *r* value of 0.55 of moderate uphill positive correlation coefficient relationship, hypertension feature has *r* value of 0.65 of strong uphill positive correlation coefficient relationship, blood pressure feature has *r* value of 0.53 of moderate uphill positive correlation coefficient relationship, cholesterol feature has *r* value of 0.44 of moderate uphill positive correlation coefficient relationship, hyperlipidemia feature has *r* value of −0.50 of moderate uphill negative correlation coefficient relationship, HDL feature has *r* value of −0.20 of moderate uphill weak negative correlation coefficient relationship, triglyceride feature has *r* value of 0.28 of weak uphill positive correlation coefficient relationship, LDL feature has *r* value of 0.35 of moderate uphill positive correlation coefficient relationship, creatinine, feature has *r* value of 0.40 of moderate uphill positive correlation coefficient relationship, body mass feature has *r* value of 0.50 of moderate uphill positive correlation coefficient relationship and heart rate feature has *r* value of 0.53 of moderate uphill positive correlation coefficient relationship.

Machine Learning Algorithms

The machine learning algorithms used for the development of the predictive models for coronary artery disease are explained in the subsequent subsections.

Logistic Regression

Logistic Regression is a supervised machine learning algorithm that is used to model the probability of an event or certain classes [7, 24]. The algorithm uses logistic function to model a binary dependent features or variables and against the independent one. The following equation (2) is used for model ling binary dependent features or variables and against the independent one:

$$i = \text{Logistic Regression}(p) = 1n\left(\frac{p}{1-p}\right). \quad (2)$$

Support Vector Machine

Support vector machine (SVM) is also a supervised machine learning algorithm which was developed by Vladimir Vapnik and his colleagues at AT&T Bell Laboratories. The algorithm is used to model data for classification and regression analysis [12, 31]. SVM builds a set of hyper plane in infinite dimensional space which might be used for regression or classification or other even tasks like detection of outliers.

K-Nearest Neighbor

K-nearest neighbor is a supervised machine learning algorithm that uses existing cases and classifies new cases based on the of similarity measure [13, 32]. The algorithm classifies a new case by votes of the majority of its neighbors with the case that is being given the most common category among its closest neighbors measure by a distance function. If $K = 1$, then the new case is assigned to its nearest neighbor [16, 20–23, 27].

Random Forest

Random forest is a supervised machine learning algorithm for regression and classification and it builds a multitude of decision tree at training time and produces the class that is the mode of the classes of the individual decision trees [14]. The algorithm consist of number of individual decision trees that work as an ensemble where each decision tree in the random forest give out a class of prediction and class with majority votes becomes model for prediction [8, 23, 24, 27, 30].

Naive Bayes

Naive Bayes is also a supervised machine learning algorithm for probabilistic classification by applying Bayes Theorem with strong independence assumption between dataset features [15]. The Eq. (3) below shows Bayes Theorem as follows:

$$P(A|C) = \frac{P(C|B)P(B)}{P(C)}. \tag{3}$$

The algorithm is highly scalable which requires a number of parameters linear in the number of features in the learning problem [9].

Gradient Boosting

Gradient boosting is a supervised machine learning algorithm for regression and classification tasks and builds a prediction model in a form of an ensemble of weak prediction model, usually in decision tree form [10, 11]. A weaker decision is usually called gradient boosted tree, which often outperforms random forest. The algorithm builds a model in a stage-wise fashion like boosting method does and it generalizes it by enabling optimization of arbitrary differentiate loss function.

Experimental Setup

Laptop HP computer System Corei7 with 8 GB of RAM and 2.8 GHz processor speed was used as the environment for data analysis, model development and evaluation. Python Programming language which is one of the most powerful open source programming was used as the programming tools for the data analysis, correlation analysis and model development and evaluation.

Predictive Models for Diagnosis of Coronary Artery Diseases

The medical expert diagnostic labeled dataset for coronary artery disease was applied on machine learning algorithm including support vector machine, K nearest neighbor, random tree, Naïve Bayes, gradient boosting and logistic regression algorithms to build predictive models for diagnosis of CAD. The python programming language was used for the development and evaluation of the models. Before, the development of the models, correlation analysis between all the dependent features which include age, sex, family history, smoking, chest pain, diabetes, glucose, hypertension, blood pressure, cholesterol, hyperlipidemia, HDL, triglyceride, LDL, creatinine, body mass and heart rate against diagnosis independent feature of the dataset was carried out. The result of the correlation analysis shows that all dependent variables have positive correlation relationship with independent feature of the dataset except hyperlipidemia and HDL features.

The CAD dataset was partitioned into 80% training set and 20% testing set, respectively, where the models were trained with 80% dataset and tested with 20% dataset. Therefore, machine learning predictive models for diagnosis of CAD with support vector machine, K nearest neighbor, random tree, Naïve Bayes, gradient boosting, logistic regression algorithms were developed for diagnosis of CAD patients. The performance evaluation of the models was carried based on accuracy, sensitivity, specificity and receiver operating curve (ROC). Table shows the performance evaluation of the models.

(i) The accuracy is an evaluation technique that is used to determine the percentage of the instances predicted correctly by the models. The accuracy can be expressed using below equation:

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fn} + \text{fp}}. \tag{4}$$

(ii) The specificity shows the percentage of CAD negative patients correctly by the models and it is being expressed using the following equation:

$$\text{Specificity} = \frac{\text{tp}}{\text{tn} + \text{fp}}. \quad (5)$$

(iii) The sensitivity shows the percentage of CAD positive patients correctly by the models and it is being expressed using the following equation:

$$\text{Sensitivity} = \frac{\text{tp}}{\text{tp} + \text{fn}}, \quad (6)$$

where *tp* is the true positive, tn is the true negative, fp is the false positive, while fn is the false negative.

iv. Receiver operating curve (ROC) is also one machine learning model evaluation technique that is used to determine the degree or percentage of how much the model is capable to differentiate between classes of the dataset instances [16, 17, 24]. ROC shows the relationship between the specificity and sensitivity of the machine learning model. Table 3 shows the result of the performance evaluation carried out on the predictive models.

In terms of accuracy random forest based predictive model happened to be the best model with 92.04%, followed by gradient boosting based predictive model which has 90.90% accuracy, then followed by support vector machine based predictive model which has 88.68% accuracy, then Naive Bayes based predictive model which has 87.50% accuracy, then K-nearest neighbor based predictive model which has 82.35% accuracy and finally logistic regression predictive model which has 80.68% accuracy. While in terms of specificity Naive Bayes based predictive model happened to be the best model with 92.40%, followed by gradient boosting based predictive model which has 91.12% specificity, then support vector predictive model which has 34% specificity, then K-nearest neighbor based predictive model which has 83.76% specificity, then random forest based predictive model with 83.34% specificity and then logistic regression based predictive model which has 81.20% specificity. Whereas, in terms of sensitivity support vector predictive model happened to be the best model with 87.34% sensitivity, followed by gradient boosting based predictive model which has 87.20% sensitivity, then random forest based predictive model which has 86.50% sensitivity, then K-nearest neighbor based predictive model which has 84.30% sensitivity, then Naive Bayes based predictive model which has 83.30% and then logistic regression based predictive model which has 83.22% sensitivity. Though, in terms of ROC, random forest based predictive model happened to be the best model with 92.20% ROC, followed by gradient booting based predictive model which has 90.28% ROC, then support vector predictive model which has 88.63% ROC, then K-nearest neighbor based predictive model which has 82.95% ROC, then logistic regression based predictive model which has 80.68% ROC and then Naive Bayes based predictive model which has 77.43% ROC.

Results and Discussion

The random forest predictive model achieved 92.04% accuracy and 92.20% ROC respectively, which make to be the best model among other models built with support vector machine, K nearest neighbor, Naïve Bayes, gradient boosting and logistic regression algorithms as shown in Fig. 11. The decision tree generated with random forest algorithm is shown in Fig. 10. The heart rate appeared to be the first splitting attribute of the decision tree of random forest model, which indicated that the heart rate is the most important attribute or feature to predict whether a patient is CAD positive or CAD negative. This corroborates moderate uphill positive correlation coefficient relationship of 0.53 with the diagnosis feature of the dataset that was earlier found out from correlation analysis conducted in the study. Other important features for the prediction according to the model are hypertension, glucose and chest pain which have 0.58, 0.55 and 0.65 positive correlation coefficient relationship with diagnosis feature of the dataset, respectively. The decision tree generated with random forest algorithm can be converted into production rules and below are some of the production rules generated from the decision tree.

Table 3 Performance evaluation result of the models

S/N	Machine learning model	Accuracy (%)	Specificity (%)	Sensitivity (%)	ROC (%)
1	Logistic regression	80.68	81.2	83.22	80.68
2	Support vector machine	88.68	86.34	87.34	88.63
3	K-nearest neighbor	82.35	83.76	84.30	82.95
4	Random forest	92.04	83.34	86.50	92.20
5	Naive Bayes	87.50	92.4	83.30	77.43
6	Gradient booting	90.90	91.12	87.20	90.28

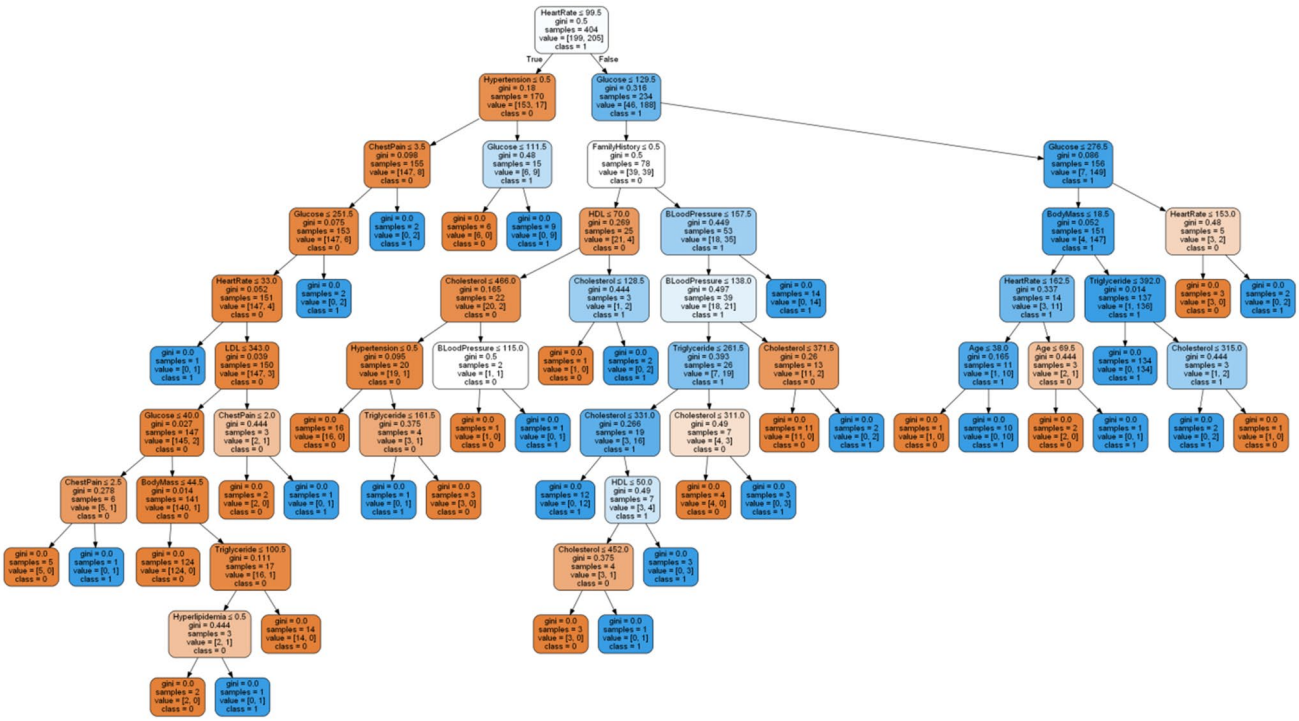

Fig. 10 Decision tree generated with random tree

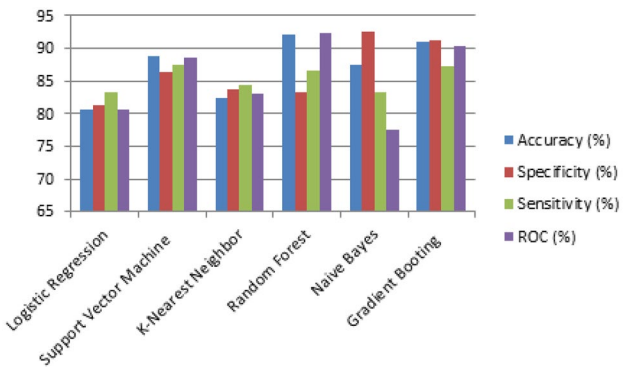

Fig. 11 Models performance evaluation result

- IF (heart rate ≤ 99.5 mg/dl and Hypertension ≤ 0.0.5 mg/dl and Glucose ≤ 251.5) THEN Positive.
- IF (Heart Rate ≤ 99.5 mg/dl and Hypertension ≤ 0.5 mg/dl and Glucose ≤ 251.5 and Glucose ≤ 251.5 and Chest Pain is 1 or 2 or 3) THEN Positive.
- IF (Heart Rate ≤ 99.5 mg/dl and Hypertension ≤ 0.0.5 mg/dl and Glucose ≤ 251.5 and Glucose ≤ 251.5 and Chest Pain is 1 or 2 or 3 and LDL ≤ 343) THEN Negative.
- IF (Heart Rate ≤ 99.5 mg/dl and Hypertension ≤ 0.5 mg/dl and Glucose ≤ 111.5 and Glucose ≤ 251.5) THEN Negative.
- IF (Heart Rate ≤ 99.5 mg/dl and Hypertension ≤ 0.5 mg/dl and Glucose ≤ 111.5) THEN Positive.

The production rules could be used to develop expert system for diagnosis of CAD patients (Fig. 11).

Conclusion

According WHO, more than half a million Nigerians died in 2012 from non-communicable diseases and every one out of five Nigerian adults over the age of 30 is likely to die prematurely from non-communicable diseases, including coronary artery disease. Therefore, in this study, a machine leaning predictive model for prediction of coronary artery diseases has been developed with the medical expert diagnostic dataset CAD which was obtained in the two General Hospitals in Kano State—Nigeria. The dataset was partitioned into 80% training set and 20% testing set, respectively, where the models were trained with the 80% and tested with 20% dataset The dataset was applied on machine learning algorithms including support vector machine, K nearest neighbor, random tree, Naïve Bayes, gradient boosting and logistic regression algorithms to build the predictive models and the models were evaluated based on accuracy, specificity, sensitivity and receiver operating curve performance evaluation

techniques. In terms of accuracy, random forest-based machine learning model emerged to be the best model with 92.04% accuracy, In terms of specificity, Naive Bayes-based machine learning model emerged to be the best model with 92.40% specificity, In terms of sensitivity, support vector machine-based machine learning model emerged to be the best model with 87.34% while in terms of ROC, random forest-based machine learning model emerged to be the best model with 92.20% sensitivity. The decision tree generated with random forest machine learning algorithm can be converted into production rules and could be used develop expert system for diagnosis of CAD patients in Nigeria.

Funding There is no funding source.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval The ethical approval has been granted by Ministry of Health, Kano State—Nigeria.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Ayatollahi H, Gholamhosseini L, Salehi M. Predicting coronary artery disease: a comparison between two data mining algorithms. BMC Public Health. 2019;9(1):1–9.
2. Abdar M, Książek W, Acharya UR. A new machine learning technique for an accurate diagnosis of coronary artery disease. Comput Methods Progr Biomed. 2019;179:104992.
3. Ali MA. Fuzzy expert system for coronary artery disease diagnosis in Jordan, Heath Technology. Berlin: Springer; 2017.
4. Bertsimas D, Orfanoudaki A, Weiner RB. Personalized treatment for coronary artery disease patients: a machine learning approach. Health Care Manag Sci. 2020;23(4):482–506.
5. Debabrata P, Mandana KM, Sarbajit P, et al. Fuzzy expert system approach for coronary artery disease screening using clinical parameters. J Knowl Based Syst. 2012;36:162–74.
6. Faro, A. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study. Lancet. 2016;388(10053):1459–544.
7. Haruna AA, Muhammad LJ, Yahaya BZ, et al. An improved C4.5 data mining driven algorithm for the diagnosis of coronary artery disease. In: International conference on digitization (ICD), Sharjah, United Arab Emirates. 2019. p. 48–52.
8. Hussain S, et al. Performance evaluation of various data mining algorithms on road traffic accident dataset. In: Satapathy S, Joshi A, editors. Information and communication technology for intelligent systems. Smart Innovation, Systems and Technologies; 2019. p. 106–12.
9. Ishaq FS, Muhammad LJ, Yahaya BZ, et al. Data mining driven models for diagnosis of diabetes mellitus: a survey. Indian J Sci Technol. 2018;11:42.
10. Muhammad LJ, Algehyne EA. Fuzzy based expert system for diagnosis of coronary artery disease in Nigeria. Health Technol. 2021;11:319–29. https://doi.org/10.1007/s12553-021-00531-z.
11. Muhammad LJ, Garba EJ, Oye ND, et al. On the problems of knowledge acquisition and representation of expert system for diagnosis of coronary artery disease (CAD). Int J u e-Serv Sci Technol. 2018;11(3):49–58.
12. Muhammad LJ, Haruna AA, Mohammed IA, et al. Performance evaluation of classification data mining algorithms on coronary artery disease dataset. In: 2019 9th international conference on computer and knowledge engineering (ICCKE), Mashhad, Iran. 2019. p. 1–5.
13. Muhammad LJ, Besiru JM, Yahaya BZ, et al. An improved C4.5 algorithm using principle of equivalent of infinitesimal and arithmetic mean best selection attribute for large dataset. In: 2020 10th international conference on computer and knowledge engineering (ICCKE), Mashhad, Iran. 2020. p. 006–10. https://doi.org/10.1109/ICCKE50421.2020.9303622.
14. Muhammad LJ, Islam MM, Usman SS. Predictive supervised machine learning models for diabetes mellitus. Springer Nat Comput Sci. 2020;1:240.
15. Muhammad LJ, Islam MM, Usman SS, et al. Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery. Springer Nat Comput Sci. 2020;1:206. https://doi.org/10.1007/s42979-020-00216-w.
16. Muhammad LJ, Algehyne EA, Usman SS, et al. Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. Springer Nat Comput Sci. 2021;2:11. https://doi.org/10.1007/s42979-020-00394-7.
17. Muhammad LJ, Ali AG, Iliya IS. Security challenges for building knowledge based economy in Nigeria. Int J Secur Appl. 2015;9:1.
18. Muhammad LJ, Garba EJ, Oye ND, et al. Fuzzy rule-driven data mining framework for knowledge acquisition for expert system. In: Translational bioinformatics in healthcare and medicine. Elsevier, Academic Press. 2021. p. 201–14.
19. Nwaneli CU. Changing trend in coronary heart disease in Nigeria. Afr Med J. 2010;1(1):1–4.
20. Noor AS, Venkatachalam PA, Ahmad FH. Diagnosis of coronary artery disease using artificial intelligence based decision support system. In: Proceedings of the international conference on man-machine systems (ICoMMS), BatuFerringhi, Penang. 2009; p. 45–54.
21. Pathak AK, Arul VJ. A predictive model for heart disease diagnosis using fuzzy logic and decision tree. In: Elçi A, Sa P, Modi C, Olague G, Sahoo M, Bakshi S, editors. Smart computing paradigms: new progresses and challenges. Advances in intelligent systems and computing, Springer Singapore. 2020. p. 767–74.
22. Polonsky TS, McClelland RL, Jorgensen NW, et al. Coronary artery calcium score and risk classification for coronary heart disease prediction. JAMA. 2010;303(16):1610–6.
23. Tsipouras MG, Exarchos TP, Fotiadis DI, et al. A decision support system for the diagnosis of coronary artery disease. In: Proceedings of the IEEE symposium on computer-based medical systems. 2008. p. 67–78.
24. Thabtah F. Machine learning in autistic spectrum disorder behavioural research: a review and ways forward. Inform Health Soc Care. 2019;44(3):278–97.
25. Velusamy D, Ramasamy K. The ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset. Comput Methods Progr Biomed. 2020;198:105770.
26. Yahaya BZ, Muhammad LJ, Abdulganiyyu N, et al. An improved C45 algorithm using L' hospital rule for large dataset. Indian J Sci Technol. 2018;11:47.
27. Nigeria Health Watch Confronting Nigeria's growing epidemic of heart disease. 2016. https://nigeriahealthwatch.com/confr

onting-nigerias-growing-epidemic-of-heart-disease/. Accessed 2 Jan 2021.
28. Nigeria Health Watch. Confronting Nigeria's growing epidemic of heart disease. 2016. https://nigeriahealthwatch.com/confronting-nigerias-growing-epidemic-of-heart-disease/. Accessed 2 Jan 2021.
29. World Health Organization. Cardiovascular diseases (CVDs). 2017. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed 4 Jan 2021.
30. Center for Disease Control and Prevention. Coronary Artery Disease. 2019. https://www.cdc.gov/heartdisease/coronary_ad.htm. Accessed 20 Jan 2021.
31. Haruna AA, Jung LT, Arputharaj V, et al. Incentive-scheduling algorithms to provide green computational data center. SN Comput Sci. 2021;2:252. https://doi.org/10.1007/s42979-021-00633-5.
32. Islam M, Mahmud S, Muhammad LJ, et al. Wearable technology to assist the patients infected with novel coronavirus (COVID-19). SN Comput Sci. 2020;1:320. https://doi.org/10.1007/s42979-020-00335-4.