

## Prediction of heart disease outcomes using machine learning classifier

Kehinde Marvelous Adeniyi<sup>1</sup>, Olasunkanmi James Oladapo<sup>1</sup>, Timothy Oluwaseun Araoye<sup>2</sup>,  
Taiwo Felix Adebayo<sup>3</sup>, Sochima Vincent Egoigwe<sup>2</sup>, Matthew Chinedu Odo<sup>2</sup>

<sup>1</sup>Department of Statistics, Ladoke Akintola University of Technology, Ogbomosho, Nigeria

<sup>2</sup>Department of Mechatronic Engineering/Africa Centre of Excellence for Sustainable Power and Energy Development (ACE-SPED),  
University of Nigeria, Nsukka, Nigeria

<sup>3</sup>Department of Industrial Technical Education, University of Nigeria, Nsukka, Nigeria

### Article Info

#### Article history:

Received Sep 14, 2022

Revised Jan 9, 2023

Accepted Jan 12, 2023

#### Keywords:

Forward and backward method

Heart disease

Logistic regression

Machine learning classifier

Modeling

### ABSTRACT

The responsibility of heart organ is to supply blood to every part of the human body. The method of diagnose heart disease in medical hospital is extremely costly and also consume doctors time of operations. This research work applied forward, backward, and enter method for selection of variables in the logistic regression model, sensitivity, specificity, accuracy, and area under characteristic curve (AUC). The logistic regression model, at 5% level of significance with the enter method is used which denotes that the risk variables associated with heart disease gives accuracy of 87.9%. The preferred model of variable selection method used was the model from forward which has 88.6%. Also using the forward method of variables selection, the process produces 10 models with the best accuracy of 88.6%. The specificity and sensitivity of the analysis model was 91.4% and 85.6%. Also, the misclassification rate was also 11.4%, Positive predicted value is 87% and negative predicted value is 90.5%. Finally, the suitable model to predict the heart disease is from the forward method of variables selection and the positive likelihood ratio is 6 i.e the patients are 6 times likely to have the heart disease and the model has AUC value of 1.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



### Corresponding Author:

Timothy Oluwaseun Araoye

Department of Mechatronic Engineering/Africa Centre of Excellence for Sustainable Power and Energy  
Development (ACE-SPED)

Nsukka, Enugu State, Nigeria

Email: timothy.araoye@unn.edu.ng, timmy4seun@yahoo.com

## 1. INTRODUCTION

The current demographic changes, joined with epidemiological and nutritional changes, are contributing to the continued shift of the cardiovascular disease problem for both the developed and developing countries. Cardiovascular disease cause high rate of death in most countries [1]. Cardiovascular diseases (CVDs), for example coronary heart disease (CHD) and stroke, are major contributors to the overall disease burden, currently ranking as the leading cause of death and causing significant disability and reduced well-being in those who survive. Different primary prevention measures, primarily regular physical activity (PA), a healthy diet, and quitting smoking, have been confirmed as evidence of reducing cardiovascular morbidity and mortality [2]. Cardiac ailments are a broad term that encompasses a wide range of cardiac conditions. The most common type of arterial disease, which can result to a heart attack, is coronary disease artery. Heart valves may be affected by other forms of cardiac sickness, or the heart may be unable to pump properly, resulting in heart failure. Heart disease affects some people from birth. Heart disease can affect anybody, even children, and

happens when a substance called plaque develops in the arteries. The arteries may narrow over time as a result of this effect, limiting blood circulation to the heart. Heart disease is increased by smoking, consuming a poor diet, and not receiving enough and proper exercise. Increased cholesterol, elevated blood pressure, and diabetes can all increase the risk of heart disease [3]. The development of plaque in the arteries that carry blood rich in oxygen to the heart causes CHD, also known as coronary artery disease (CAD). Plaque is a buildup of fat, cholesterol, and calcium deposits in the arteries that can last for years. Atherosclerosis is a disease in which plaque narrows and hardens the coronary arteries in the course of time. The buildup of plaque on the inside wall of an artery is known as atherosclerosis. Although coronary heart disease is often asymptomatic, it increases the risk of angina (chest pain or discomfort), cardiac arrest, heart problems, and cardiac arrhythmias. Lowered or blocked blood flow to the heart causes angina and heart attacks. Stable angina usually gets worse with physical activity and gets better with rest, but a heart attack can kill the heart muscle and necessitates immediate medical intervention [4]. Chronic high blood pressure can really cause artery walls to stiffen, resulting in a decline or reduction in blood flow. The term "silent killer" is often used to describe high blood pressure. People with lower educational attainment, poorer household income, people over 55 years old, retirees and residents unable to work, and people of Native Hawaiian or Japanese heritage are more likely to have high blood pressure [5], [6]. Cardiovascular disease, often known as heart disease, is a primary cause of various disabilities and premature deaths globally, and it greatly influences the growing expenses of healthcare costs [7]. Population-based strategies in line with production of cost-effective medicines for people with established disease and compromised immunity, might prevent a significant portion of this morbidity and death. Kulkarni [8], Prasad *et al.* [9] and Suresh *et al.* [10] developed a hybrid prediction method model based on machine learning algorithms for cardiac disease diagnosis. The suggested system uses a feature selection algorithm and classification algorithms to analyze the heart disease dataset, which comprises the most appropriate features and values for prediction. The accuracy of the outcomes is calculated using the training and testing datasets.

To effectively forecast heart illness, an intelligent heart disease prediction system is presented that uses a combination of machine learning approaches. The suggested method was tested using the Kaggle heart disease dataset, which demonstrated that the machine learning techniques used were effective in predicting heart and circulatory disorders. Using optimization techniques, the performance of these predictive classifiers for heart disease prediction could be increased in the future. Shammari *et al.* [11] develop model for examine heart disease using multi-classifier which comprises of 13 attributes and 270 cases. The researcher applied artificial neural network (ANN), J48, Naïve Bayes and REPTree classifiers for selection of most accurate prediction. The result of the research shows that Naïve Bays model classifier is most accuracy with 85% among others model. Qazi *et al.* [12] proposed automatic linear detection classifier for heart disease abnormality using fisher's linear discriminant model. The results of model proposed was compared with other machine learning methods which includes relevance vector machine (RVM), support vector machine (SVM) and linear exponential distribution (LED). Machine Learning model gives best accuracy of 89.6% among others method.

Şajn and Kukar [13] developed machine learning and image processing method for health care Centre. The proposed method shows that there is tremendous improvement through application of diagnostic automatic system which improves the probabilities of posttest diagnostic using image multiresolution parameterization and subset feature arrangements with machine learning methods. The accuracy of the proposed method is 81.3%. Attia *et al.* [14] proposed artificial intelligent (AI) method that allows electrocardiograph (ECG) detect signature of fibrillation atrial system. The result validation was done by operating receiver curve characteristic technique. The results insinuate that the model proposed gives specificity, accuracy and sensitivity of 79%, 87% and 79.5% respectively.

Melgarejo-Meseguer *et al.* [15] applied machine learning method to detect fragment electrocardiography activity. The proposed method used deep learning analysis for data selection of decision tree, SVM and Naïve Gaussian Bayes. The best results gotten from fragmented dataset were 88% specificity, 94% sensitivity, 89% predictive positive value, 91% accuracy and 93% predictive negative value, when SVM-Gaussian kernel was applied. Sureja *et al.* [16] developed new model which focused on support vector machine and salp swarm algorithm for prediction of heart diseases. The algorithm is applied to randomly select best approach from the program database. The proposed method adopted by the researcher gives an accuracy of 98.46% and 98.75% with the system dataset 2 and 1 respectively.

Wiskey *et al.* [17] applied machine learning (ML) using artificial neural network, k-means cluster and decision tree techniques. The research performed is sections into two patterns which includes classification and preprocessing. The accuracy of classification pattern is 99.77% which is more significant compare to second pattern. Also, the classification results show drastically distribution knowledge on the decision tree. Prusty *et al.* [18] study nine classifiers which comprises of deep learning and machine learning techniques for prediction of heart coronary failure. The computation models applied is cheap and simple to operate.

The classifier compared and tested using matrix confusion method. The logistic classifier regression produced optimal outcome accuracy, F1-score and precision of 90.78%, 91.35%, and 90.24% respectively. George and Gaikwad [19] control level of cholesterol in cardiovascular patient through application of system dynamic mathematical models. The model applied recovered the patient faster sporadically. Also, the models reduced the effect of heart stroke and regulate a healthy life. Faieq and Mijwil [20] applied artificial neural network and support vector machine techniques for determining early heart disease diagnosis. The findings confirmed that support vector machine gives maximum execution and high predictive accuracy. Suboh *et al.* [21] analyzed abnormal and normal heart sound response from four pattern of heart nerve disease. The system automation which includes of segmentation, data extraction and heart sound response classification is analyzed in hardware and personal computer (PC) platforms with input electronic stethoscope. The two methods give specificity of 96.3%. Also, portable device only produced sensitivity of 77.78% and accuracy of 87.04% in comparison with PC platform that gives 90% of accuracy and sensitivity.

This research paper applied machine learning based on logistic regression algorithms to determine the outcome of a patient that has cardiac disease in University of Nigeria Teaching Hospital between the years 2018-2021. Also, forward, backward and enter method was used to select variables in the logistic regression model. These findings are crucial not just in terms of diagnosing heart disease but also in terms of reducing the causes of death. The results of this research propose the best model which is suitable for heart disease detection or prediction and also helps physicians and researchers to improve the standard of heart disease diagnosis through machine learning based on logistic regression algorithms.

**2. MATERIALS AND METHOD**

The research data were collected from University of Nigeria Teaching Hospital for descriptive analysis of heart diseases among adult with information of 1,025 respondents. The data were analyzed and coded using statistical software. The explanatory variables used in this research are: age (X<sub>1</sub>), sex (X<sub>2</sub>), chest pain type (X<sub>3</sub>), resting blood pressure (X<sub>4</sub>), cholesterol (X<sub>5</sub>), fasting blood sugar (X<sub>6</sub>), resting electrocardiogram test (X<sub>7</sub>), maximum heart rate (X<sub>8</sub>), exercise induced angina (X<sub>9</sub>), slope (X<sub>10</sub>), vessels colored by floroscopy (X<sub>11</sub>), thalassemia (X<sub>12</sub>), old peak (X<sub>13</sub>) to know the risk factors that contributes to heart disease. Also forward, backward and enter method was used to select variables in the logistic regression model, sensitivity, specificity, accuracy and AUC (Area under receiver operating characteristic curve).

**2.1. Logistic regression model**

In logistic regression, the interaction between the variable outcome and the predictor factors is with regard to logit: the natural logarithm of odds. Take the following scenario for instance: Y is a binary variable outcome with values of "0" and "1," and continuous variable predictor is denoted by X. When a scatter plot is drawn, each outcome variable category will be represented by two parallel lines. Because the relationship does not follow a linear trend, it cannot be described using basic linear regression. By performing a logit transformation on the result variable Y, logistic regression makes this scenario easier. The most basic logistic regression model is [22].

$$\text{Logit}(Y) = \frac{\pi}{1-\pi} = \beta_0 + \beta_1 \tag{1}$$

The above equation represents the chance of the result Y happening, and  $\frac{\pi}{1-\pi}$  represents the odds of success; the ratio of the likelihood of the outcome Y happening to the probability of the outcome Y not happening. The values  $\beta_1$  and  $\beta_0$  indicates the slope and intercept respectively of regression coefficient. The probability of occurrence of result Y may be estimated for a given value of predictor X by taking antilog on both sides of (1):

$$\pi = (Y/X=x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \tag{2}$$

the predictor variable "X" can be categorical or continuous. The logistic model can be stretched to include several predictors.

$$\text{Logic}(Y) = \ln \frac{\pi}{1-\pi} = \beta_0 + \beta_{1x_1} \dots \dots \beta_{px_p} \tag{3}$$

According to (3) indicates the generic representation of a logistic model regression for "p" predictors. The maximum likelihood (ML) approach or the weighted least squares model is applied to analyze regression variables. The  $\beta_1 \dots \beta_p$  which indicates regression coefficients refers to the link between logit of Y and X. When a coefficient variable is larger than 0, it means that the logit of Y increases as X increases, whereas a coefficient

variable less than 0 shows that the logit of X variable increases as Y variable increases. When the coefficient variable is 0, it signifies that there is no significant linear relationship between the logit of Y and the predictor's X. Normally, the odds ratio is presented along with the regression coefficient for simplicity of interpretation. The odds ratio can be computed using the formula below [23], [24]:

$$\text{Odd Ratio} = e^{\beta} \quad (4)$$

Wald's test is applied to determine the statistical application of the regression coefficient, and the likelihood ratio test or pseudo R<sup>2</sup> test can be used to determine the overall model significance.

## 2.2. Forward and backward selection

Forward selection is an instance stepwise regression in which variables are gradually added to a blank model. In each forward step, we add the one variable that provides the most improvement to the model. Backward elimination is nearly the converse of forward elimination; we begin with a model that includes every feasible variable and gradually remove the superfluous variables [25], [26].

## 3. RESULT AND DISCUSSION

From the analysis, the factors that are used in predicting the heart disease outcomes are listed and analyzed based on level of significant (5%). The factors that were considered include; age, sex, type of chest pain, resting blood pressure, intake of cholesterol, fasting blood sugar, resting electrocardiogram test, the maximum heart rate, exercise induced angina, slope, vessels colored by fluoroscopy, thalassemia and the oldpeak.

### 3.1. Interpretation

#### 3.1.1. For age

With all other factors held constant, Exp(B) for age is 1.022. Which means that for every rise in age. There is a 1.022 chance that the patient would develop heart disease.

#### 3.1.2. For sex

Sex (1) which contrasts 'male' with 'female' has an exp(B) of 0.127. Which means that a male patient is only 0.117 less likely to have heart disease than the female patient. Having allowed all other variables to be held constant.

#### 3.1.3. For chest pain type

With all other factors held constant, chest pain type (1), which contrasts 'typical angina' with 'asymptomatic,' has an exp(B) of 0.089. Which suggests that people with typical angina are only 0.089 times less likely to suffer heart disease compared to patients with asymptomatic chest pain type. With all other factors held constant, chest pain type (2), which contrasts 'atypical angina' with 'asymptomatic,' has an exp(B) of 0.206, implying that individuals with atypical angina are only 0.206 times less likely to suffer heart disease compared to patients with asymptomatic chest pain type. With all other variables held constant, chest pain type (3), which contrasts 'non-anginal pain' with 'asymptomatic,' has an exp(B) of 0.730, which means that patients with non-anginal pain are only 0.73 times (i.e. much more) likely to have heart disease compared to patients with asymptomatic chest pain type.

#### 3.1.4. For resting blood pressure

The resting blood sugar coefficient is statistically significant Exp(B) for resting blood pressure is 0.975. Which means for every increase in resting blood pressure. There is 0.975 possibility of the patient having heart disease, having allowed all other variables to be held constant.

#### 3.1.5. For cholesterol

Cholesterol (1) which contrasts 'desirable' with 'high' has an exp(B) of 1.152 which means that patients with desirable are only 1.152 times (i.e. much more) likely to have heart disease compared to patients with high cholesterol, having allowed all other variables to be held constant. Cholesterol (2) which contrasts 'borderline high' with 'high' has an exp(B) of 2.335 which means that patients with borderline high are only 2.335 times less likely to have heart disease than patients with high cholesterol, having allowed all other variables to be held constant.

### 3.1.6. For fasting blood sugar

Fasting blood sugar (1), which contrasts 'more than 120mg/ml' with 'lower than 120mg/ml'. Has an  $\exp(B)$  of 1.593. Implying that patients with greater than 120mg/ml fasting blood sugar are only 1.593 times (i.e. considerably more) likely to suffer heart disease when all other variables are held constant.

### 3.1.7. For rest-ecg

Rest-ecg (1), which compares 'normal' to 'left ventricular hypertrophy,' has an  $\exp(B)$  of 1.959, implying that patients with normal resting electrocardiography are only 1.959 times (i.e. much more) likely to have heart disease than patients with left ventricular hypertrophy resting electrocardiography. Rest-ecg (2), which compares 'ST-T wave abnormality' to 'left ventricular hypertrophy,' has an  $\exp(B)$  of 2.826, implying that patients with ST-T wave abnormality are only 2.826 times (i.e. considerably more) likely to suffer heart disease than patients with left ventricular hypertrophy resting electrocardiography, assuming all other variables are constant.

### 3.1.8. For max heart rate

The maximum heart rate coefficient is statistically significant.  $\exp(B)$  for maximum heart rate is 1.021. Which means for every increase in heart rate, there is 1.021 possibility of the patient having heart disease, having allowed all other variables to be held constant.

### 3.1.9. For exercise induced angina

Exercise caused angina (1), which contrasts 'no' with 'yes,' has an  $\exp(B)=2.022$ . Implying that patient with 'no exercise induced angina' are only 2.022 times (i.e. less) likely to suffer heart disease than patients with exercise induced angina'. Assuming all other variables remain constant.

### 3.1.10. For slope

Slope (1), which compares 'downsloping' to 'flat,' has an  $\exp(B)=4.101$ , implying that patients with downsloping are only 4.101 times less likely to develop heart disease than those with flat slope when all other variables are held constant. Slope (2), which compares 'upsloping' to 'flat,' has an  $\exp(B)=2.317$ . Implying that patients with upsloping are only 2.317 times (i.e. considerably more) likely to develop heart disease than patients with flat slope, assuming all other variables remain constant.

### 3.1.11. For vessels colored by flourosopy

Vessels colored by flourosopy (1) which contrasts 'zero' with 'four' has an  $\exp(B)$  of 0.219 which means that patients with 'zero' are only 0.219 times (i.e. much more) likely to have heart disease than patients with 'four', having allowed all other variables to be held constant. Vessels colored by flourosopy (2) which contrasts 'one' with 'four' has an  $\exp(B)$  of 0.018 which means that patients with 'one' are only 0.018 times less likely to have heart disease than patients with 'four', having allowed all other variables to be held constant. Vessels colored by flourosopy (3) which contrasts 'two' with 'four' has an  $\exp(B)$  of 0.006 which means that patients with 'two' are only 0.006 times less likely to have heart disease than patients with 'four', having allowed all other variables to be held constant. Vessels colored by flourosopy (4) which contrasts 'three' with 'four' has an  $\exp(B)$  of 0.024 which means that patients with 'three' are only 0.024 times less likely to have heart disease than patients with 'four', having allowed all other variables to be held constant.

### 3.1.12. For thalassemia

Thalassemia means a blood disorder. Thalassemia (1) which contrasts 'no' with 'reversible defect' has an  $\exp(B)$  of 0.237 which means that patients with no thalassemia are only 0.237 times (i.e much more) likely to have heart disease than patients with reversible defect, having allowed all other variables to be held constant. Thalassemia (2) which contrasts 'normal' with 'reversible defect' has an  $\exp(B)$  of 6.499 which means that patients with normal thalassemia are only 6.499 times (i.e less) likely to have heart disease than patients with reversible defect, having allowed all other variables to be held constant. Thalassemia (3) which contrasts 'fixed defect' with 'reversible defect' has an  $\exp(B)$  of 4.390 which means that patients with fixed defect thalassemia are only 4.390 times less likely to have heart disease than patients with reversible defect, having allowed all other variables to be held constant.

### 3.1.13. For old peak

The old peak coefficient is statistically significant.  $\exp(B)$  for the old peak is 0.624, which means for every increase in heart rate, there is 0.624 possibility of the patient having heart disease, having allowed all other variables to be held constant.

**3.2. Forward Stepwise (likelihood ratio) regression**

The regression line for this method shows the factors that predict heart disease outcomes at 5% level of significant:

**The Regression Line:**

$$\begin{aligned} \text{Logit}(Y) = & 6.484 - 1.705x_1 - 2.588x_2 - 1.556x_3 - 0.387x_4 - 0.021x_5 + 0.826x_6 + 1.574x_7 \\ & + 0.814x_8 - 1.547x_9 - 3.902x_{10} - 4.789x_{11} - 3.950x_{12} - 0.956x_{13} \\ & + 1.788x_{14} + 1.440x_{15} - 0.487x_{16} \end{aligned}$$

where X<sub>1</sub>, X<sub>2</sub>, X<sub>3</sub>, X<sub>4</sub>, X<sub>5</sub>, X<sub>6</sub>, X<sub>7</sub>, X<sub>8</sub>, X<sub>9</sub>, X<sub>10</sub>, X<sub>11</sub>, X<sub>12</sub>, X<sub>13</sub>, X<sub>14</sub>, X<sub>15</sub>, and X<sub>16</sub> are male, typical angina chest pain type, atypical angina chest pain type, non-anginal pain chest pain type, resting blood pressure, no exercise induced angina, downslopping, upslopping, zero vessels colored by flourosopy, one vessels colored by flourosopy, two vessels colored by flourosopy, three vessels colored by flourosopy, no thalassemia, normal thalassemia, fixed defect thalassemia and old peak respectively. The variables in the equation for the forward Stepwise (likelihood Ratio) approach in Model 8 are shown in the regression line above. Model 8 is the suitable model for this method. Resting blood pressure, sex, exercise induced angina, chest pain type, vessels colored by flourosopy, slope, thalassemia, old peak are the variables that are significant. From the Table 1, variables that are significant at 5% are: Chest pain type, sex, resting blood pressure, cholesterol, exercise induced angina, max heart rate, slope, vessels colored by flourosopy, thalassemia, and old peak. This table shows the percentage accuracy of heart disease, with the accuracy of 87.9%, sensitivity of 91.4% and the specificity to be 84.2%.

Table 1. Classification table for forward stepwise (likelihood ratio) method

Model	Specificity (%)	Sensitivity (%)	Accuracy (%)	Variables insert
1	75.2	76.8	76.0	X <sub>3</sub>
2	82.0	71.7	76.7	X <sub>3</sub> , X <sub>11</sub>
3	81.8	88.8	85.4	X <sub>3</sub> , X <sub>11</sub> , X <sub>12</sub>
4	85.4	87.6	86.5	X <sub>3</sub> , X <sub>10</sub> , X <sub>11</sub> , X <sub>12</sub>
5	87.8	85.2	86.4	X <sub>2</sub> , X <sub>3</sub> , X <sub>10</sub> , X <sub>11</sub> , X <sub>12</sub>
6	86.4	88.4	87.4	X <sub>2</sub> , X <sub>3</sub> , X <sub>10</sub> , X <sub>11</sub> , X <sub>12</sub> , X <sub>13</sub>
7	87.2	87.6	87.4	X <sub>2</sub> , X <sub>3</sub> , X <sub>4</sub> , X <sub>10</sub> , X <sub>11</sub> , X <sub>12</sub> , X <sub>13</sub>
8	85.6	91.4	88.6	X <sub>2</sub> , X <sub>3</sub> , X <sub>4</sub> , X <sub>9</sub> , X <sub>10</sub> , X <sub>11</sub> , X <sub>12</sub> , X <sub>13</sub>
9	84.0	90.3	87.2	X <sub>2</sub> , X <sub>3</sub> , X <sub>4</sub> , X <sub>5</sub> , X <sub>9</sub> , X <sub>10</sub> , X <sub>11</sub> , X <sub>12</sub> , X <sub>13</sub>
10	84.2	90.9	87.6	X <sub>2</sub> , X <sub>3</sub> , X <sub>4</sub> , X <sub>5</sub> , X <sub>8</sub> , X <sub>9</sub> , X <sub>10</sub> , X <sub>11</sub> , X <sub>12</sub> , X <sub>13</sub>

Figure 1 shows the overall proportion of predicted cases. Model 8 has the best accuracy of 88.6% and sensitivity of 91.4%, according to this classification table. Figure 2 shows the model analysis from 1 to 4. It shows that model 1 has the best sensitivity of 91.4. The overall percentage of correctly classified cases is 87.9%. The overall percentage of cases accurately classified is 88.6%.

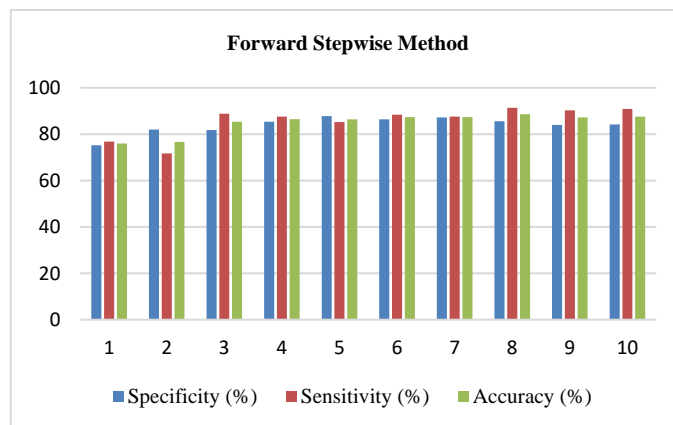


Figure 1. Backward stepwise (likelihood ratio) regression

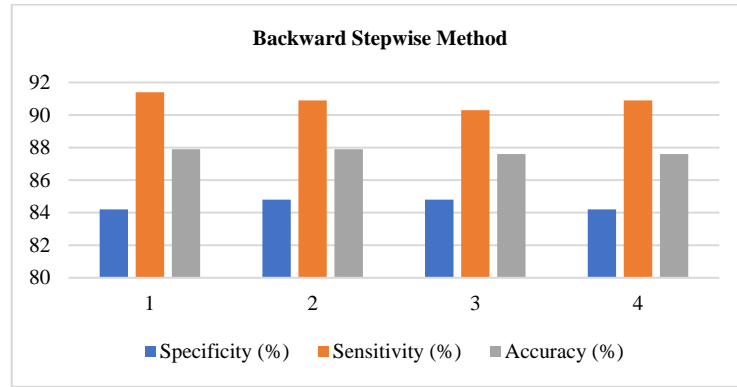


Figure 2. Forward stepwise (likelihood ratio) regression

**3.3. Backward stepwise (likelihood ratio) regression**

The regression line for this method shows the factors that predict heart disease outcomes at 5% level of significant:

**The Regression Line**

$$\begin{aligned}
 \text{Logit}(Y) = & 1.633 + 0.022x_1 - 2.067x_2 - 2.424x_3 - 1.582x_4 - 0.314x_5 - 0.025x_6 + 0.141x_7 \\
 & + 0.848x_8 + 0.465x_9 + 0.672x_{10} + 1.039x_{11} + 0.021x_{12} + 0.704x_{13} + 1.411x_{14} \\
 & + 0.840x_{15} - 1.517x_{16} - 4.007x_{17} - 5.173x_{18} - 3.735x_{19} - 1.438x_{20} + 1.872x_{21} \\
 & + 1.479x_{22} - 0.471x_{23}
 \end{aligned}$$

where  $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{22}$  and  $x_{23}$  are: age, male, typical angina chest pain type, atypical angina chest pain type, non-anginal pain chest pain type, resting blood pressure, desirable cholesterol, borderline high cholesterol, greater than 120mg/ml fasting blood sugar, normal rest ecg, ST-T wave abnormality, max heart rate, no exercise induced angina, down-slopping, up-slopping, zero vessels colored by flourosopy, one vessels colored by flourosopy, two vessels colored by flourosopy, three vessels colored by flourosopy, no thalassemia, normal thalassemia, fixed defect thalassemia and old peak respectively. The variables in the equation for the forward Stepwise (likelihood Ratio) approach in model 1 are shown in the regression line below. Table 2 shows that models 1 and 2 are ideal for this procedure; both have an accuracy of 87.9%. Sex, kind of chest pain, resting blood pressure, cholesterol, maximum heart rate, exercise-induced angina, Slope, vessels colored by flourosopy, thalassemia, and old peak are the variables that are significant. The results finding shows an agreement with the previous studies that Machine learning were effective in predicting heart and circulatory disorders [15]-[26].

Table 3 compares the forward and backward methods of variable selection, and it can be seen that the forward technique is more efficient, it shows that the forward method has the suitable model with the accuracy of 88.6%, sensitivity of 91.4% and specificity of 85.6% and it's the suitable model that fits the data. This table compares the forward and backward methods of variable selection, and it can be seen that the forward technique is more efficient, it shows that the forward method has the suitable model with the accuracy of 88.6%, sensitivity of 91.4% and specificity of 85.6% and it's the suitable model that fits the data.

Table 2. Classification table for backward stepwise (likelihood ratio) method

Model	Specificity (%)	Sensitivity (%)	Accuracy (%)	Variables insert
1	84.2	91.4	87.9	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}$
2	84.8	90.9	87.9	$x_1, x_2, x_3, x_4, x_5, x_6, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}$
3	84.8	90.3	87.6	$x_2, x_3, x_4, x_5, x_6, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}$
4	84.2	90.9	87.6	$x_2, x_3, x_4, x_5, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}$

Source: SPSS output

Table 3. Comparison between forward and backward method

Forward stepwise (likelihood ratio)			
Model	Specificity (%)	Sensitivity (%)	Accuracy (%)
8	85.6	91.4	88.6
Backward stepwise (likelihood ratio)			
1	84.2	91.4	87.9

And for the suitable model, the following were calculated:

- Misclassification Rate=1-Accuracy=1-0.886=11.4%. This shows the error rate of the model, which is 11.4%.
- Positive predicted value= $\frac{TP}{TP+FP}$ , shows that if the test result is positive, what is the likelihood that the patient will have heart disease.  $=\frac{481}{481+72} = 87.0\%$ .
- Since the result is positive, we conclude that the likelihood that the patient will have heart disease is 87.0%.
- 3. Negative predicted value= $\frac{TN}{TN+FN}$ , shows that if the test result is negative, what is the likelihood that the patient will be healthy (i.e. will not have heart disease)  $=\frac{427}{427+45} = 90.5\%$ .
- Since the result is positive, we conclude that the likelihood that the patient will be healthy is 90.5%
- Positive likelihood ratio= $\frac{Sensitivity}{1-Specificity} = \frac{0.914}{1-0.856} = 6.347 \approx 6$
- These show that the patients are 6 times more likely to have the heart disease.
- Negative likelihood ratio =  $\frac{1-Sensitivity}{Specificity} = \frac{1-0.914}{0.856} = 0$ . The result shows that the patients are not likely to have heart disease.
- $AUC = \frac{Sensitivity}{Specificity} = \frac{0.914}{0.856} = 1.068$ .
- This shows that the AUC of the model has an outstanding discrimination.

#### 4. CONCLUSIONS

The data used was collected from University of Nigeria Teaching Hospital, with information of 1,025 respondents. The objectives of this research work are to show the variables that contribute to heart disease and to also know the model with the suitable classification. For age, resting blood pressure, maximum heart rate, and old peak, the descriptive analysis summary are as follows: their minimum value are 29, 94, 71, and 0.00 respectively, their maximum value are 77, 200, 202, and 6.20 respectively and their average mean are 54 years, 131.61, 149.11, and 1.0715 respectively. For the logistic regression analysis, at 5% level of significance, the enter method is used and the variables that are significant are: resting blood pressure, sex, chest pain type, cholesterol, max heart rate, exercise induced angina, slope, vessels colored by flourosopy, thalassemia, old peak. also using the forward method of variables selection, the following variables are significant: sex, resting blood pressure, chest pain type, exercise induced angina, slope, vessels colored by flourosopy, thalassemia, and old peak. Using the backward method of variables selection, the following variables are significant: Resting blood pressure, sex, chest pain type, vessels colored by flourosopy, slope, and exercise induced angina, thalassemia, old peak. In the forward method, the best model for specificity is model 7 with 87.2%. The best model for sensitivity and accuracy is model 8 with 96.4% and 88.6% respectively. For the backward method the best model for specificity is model 2 with 84.8%, the best model for both sensitivity and accuracy is model 1 with 91.4% and 87.9% respectively. The two comparison shows that the overall best model is from the forward selection with model 8 with 88.6% accuracy, 91.4% sensitivity and 85.6% specificity; the AUC is 1.068 which shows that the model has an outstanding discrimination.

#### ACKNOWLEDGMENT

The authors acknowledge the support received from the Department of Statistics, LAUTECH Ogbomoso and Africa Centre of Excellence for Sustainable Power and Energy Development (ACE-SPED), University of Nigeria, Nsukka for their support and enabled the timely completion of this research.

#### REFERENCES





- [1] A. Maharani, Sujarwoto, D. Praveen, D. Oceandy, G. Tampubolon, and A. Patel, "Cardiovascular disease risk factor prevalence and estimated 10-year cardiovascular risk scores in Indonesia: The SMARThealth Extend study," *PLOS ONE*, vol. 14, no. 4, p. e0215219, Apr. 2019, doi: 10.1371/journal.pone.0215219.
- [2] Centers for Disease Control and Prevention, "Division for heart disease and stroke prevention," *Trends & maps*, 2009. [http://nccd.cdc.gov/DHDSP\\_DTM/LocationSummary.aspx?state=Louisiana%5Cnhttp://www.cdc.gov/dhdsp/](http://nccd.cdc.gov/DHDSP_DTM/LocationSummary.aspx?state=Louisiana%5Cnhttp://www.cdc.gov/dhdsp/) (accessed Jan. 28, 2022).
- [3] J. Li and J. Siegrist, "Physical activity and risk of cardiovascular disease-a meta-analysis of prospective cohort studies," *International Journal of Environmental Research and Public Health*, vol. 9, no. 2, pp. 391–407, Jan. 2012, doi: 10.3390/ijerph9020391.
- [4] S. S. Virani *et al.*, "Heart disease and stroke statistics—2021 update," *Circulation*, vol. 143, no. 8, pp. E254–E743, Feb. 2021, doi: 10.1161/CIR.0000000000000950.






- [5] R. Devol and A. Bedroussian, "An Unhealthy America : the economic burden of chronic disease," *Milken Institute*, no. October, pp. 1–252, 2007, [Online]. Available: <http://health-equity.pitt.edu/id/eprint/847>.
- [6] World Health Organization, "WHO | Prevention of recurrent heart attacks and strokes in low and middle income populations," *WHO*, 2013.
- [7] R. Lafta, J. Zhang, X. Tao, Y. Li, M. Diykh, and J. C.-W. Lin, "A structural graph-coupled advanced machine learning ensemble model for disease risk prediction in a telehealthcare environment," in *Studies in Big Data*, vol. 44, 2018, pp. 363–384.
- [8] N. Kulkarni, "Support vector machine based alzheimer's disease diagnosis using synchrony features," *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 9, no. 1, p. 57, Apr. 2020, doi: 10.11591/ijict.v9i1.pp57-62.
- [9] R. Prasad, P. Anjali, S. Adil, and N. Deepa, "Heart disease prediction using logistic regression algorithm using machine learning," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 3 Special Issue, pp. 659–662, 2019.
- [10] P. Suresh, P. Keerthika, K. Logeswaran, D. Myvizhi, K. Shobika, and G. S. Raja, "A hybrid heart disease prediction system using machine learning techniques," *International Journal of Advanced Science and Technology*, vol. 29, no. 5, pp. 7262–7275, 2020.
- [11] A. Al Shammari, H. Al, and H. Zardi, "Prediction of heart diseases (PHDs) based on multi-classifiers," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 5, pp. 230–236, 2020, doi: 10.14569/IJACSA.2020.0110531.
- [12] M. Qazi, G. Fung, S. Krishnan, J. Bi, R. Rao, and A. S. Katz, "Automated heart abnormality detection using sparse linear classifiers," *IEEE Engineering in Medicine and Biology Magazine*, vol. 26, no. 2, pp. 56–63, Mar. 2007, doi: 10.1109/EMEMB.2007.335591.
- [13] L. Šajn and M. Kukar, "Image processing and machine learning for fully automated probabilistic evaluation of medical images," *Computer Methods and Programs in Biomedicine*, vol. 104, no. 3, pp. e75–e86, Dec. 2011, doi: 10.1016/j.cmpb.2010.06.021.
- [14] Z. I. Attia *et al.*, "An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction," *The Lancet*, vol. 394, no. 10201, pp. 861–867, Sep. 2019, doi: 10.1016/S0140-6736(19)31721-0.
- [15] F.-M. Melgarejo-Meseguer *et al.*, "Electrocardiographic fragmented activity (II): A machine learning approach to detection," *Applied Sciences*, vol. 9, no. 17, p. 3565, Aug. 2019, doi: 10.3390/app9173565.
- [16] N. Sureja, B. Chawda, and A. Vasant, "A novel salp swarm clustering algorithm for prediction of the heart diseases," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 25, no. 1, pp. 265–272, 2022, doi: 10.11591/ijeecs.v25.i1.pp265-272.
- [17] I. A. Wisky, M. Yanto, Y. Wiyandra, H. Syahputra, and F. Hadi, "Machine learning classification of infectious disease distribution status," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 27, no. 3, p. 1557, Sep. 2022, doi: 10.11591/ijeecs.v27.i3.pp1557-1566.
- [18] S. Prusty, S. Patnaik, and S. K. Dash, "Comparative analysis and prediction of coronary heart disease," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 27, no. 2, pp. 944–953, Aug. 2022, doi: 10.11591/ijeecs.v27.i2.pp944-953.
- [19] J. P. George and S. M. Gaikwad, "Simulation modeling for heart attack patient by mapping cholesterol level," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 1, p. 16, Apr. 2020, doi: 10.11591/ijeecs.v18.i1.pp16-23.
- [20] A. K. Faieq and M. M. Mijwil, "Prediction of of heart diseases utilising support vector machine and artificial neural network," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 26, no. 1, p. 374, Apr. 2022, doi: 10.11591/ijeecs.v26.i1.pp374-380.
- [21] M. Z. Suboh *et al.*, "Portable heart valve disease screening device using electronic stethoscope," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 15, no. 1, p. 122, Jul. 2019, doi: 10.11591/ijeecs.v15.i1.pp122-132.
- [22] D. Kelley, "Heart Disease: Causes, Prevention, and Current Research," *JCCC Honors Journal*, vol. 5, no. 2, p. 1, 2014.
- [23] M. Rahman, K. Nakamura, K. Seino, and M. Kizuki, "Sociodemographic factors and the risk of developing cardiovascular disease in Bangladesh," *American Journal of Preventive Medicine*, vol. 48, no. 4, pp. 456–461, Apr. 2015, doi: 10.1016/j.amepre.2014.10.009.
- [24] T. C. Turin *et al.*, "Time lag to hospitalisation and the associated determinants in patients with acute myocardial infarction: The Takashima AMI Registry, Japan," *Emergency Medicine Journal*, vol. 28, no. 3, pp. 239–241, Mar. 2011, doi: 10.1136/emj.2009.087676.
- [25] I. P. Bhatti, . H. D. Lohano, Z. A. Pirzado, and . I. A. Jafri, "A logistic regression analysis of the ischemic heart disease risk," *Journal of Applied Sciences*, vol. 6, no. 4, pp. 785–788, Feb. 2006, doi: 10.3923/jas.2006.785.788.
- [26] R. L. Lafta, M. S. AL-Musaylh, and Q. M. Shallal, "Clustering similar time series data for the prediction the patients with heart disease," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 2, p. 947, May 2022, doi: 10.11591/ijeecs.v26.i2.pp947-954.

## BIOGRAPHIES OF AUTHORS






**Kehinde Marvelous Adeniyi**     is a graduate of Statistics in the Department of Statistics, Ladoko Akintola University of Technology (LAUTECH), Ogbomoso, Oyo State. Her research interest includes econometrics, biometrics, biostatistics, linear modeling and regression analysis. She can be contacted at email: [kennymarvel15@gmail.com](mailto:kennymarvel15@gmail.com).






**Olasunkanmi James Oladapo**    is a Lecturer in the Department of Statistics, LAUTECH. He obtained his Bachelor degree of Technology (B.Tech) and Master's degree (M.Tech) in Statistics at the Ladoko Akintola University of Technology (LAUTECH), Ogbomoso, Oyo State. He is presently pursuing a Ph.D degree at the Federal University of Technology Akure (FUTA). His research interest includes econometrics, biometrics, biostatistics, linear modeling and regression analysis. He can be contacted at email: [ojoladapo@lautech.edu.ng](mailto:ojoladapo@lautech.edu.ng).






**Timothy Oluwaseun Araoye**    is a Lecturer in the Department of Mechatronics Engineering, University of Nigeria, Nsukka, Nigeria. He is a Registered Member of the Nigeria Society of Engineers (NSE) and council for the regulation of engineers in Nigeria (COREN). He obtained his Bachelor degree of Technology (B.Tech) in Electrical and Electronics Engineering, at the Ladoko Akintola University of Technology (LAUTECH), Ogbomoso, Oyo State, Master's Degree (M.Eng) at the Enugu State University of Science and Technology (ESUT) Agbani, Enugu State and Ph.D degree at the University of Abuja, Nigeria in the same course. His research interest includes artificial intelligence, machine learning, power system and machine, renewable energy, power system reliability, microgrid, modeling and simulation, distributed generation, control system, and power electronics. He has attended many conferences where he presented papers. He can be contacted at email: [timothy.araoye@unn.edu.ng](mailto:timothy.araoye@unn.edu.ng).






**Taiwo Felix Adebayo**    is a Lecturer in the department of Industrial Technical Education (Electrical Electronics Technology), University of Nigeria Nsukka. He holds both Bachelor and Masters in Industrial Technical Education from University of Nigeria Nsukka. He has published several Journals to his credit. His research interest includes; artificial intelligent, machine learning and electrical sensors. He can be contacted at email: [taiwo.adebayo@unn.edu.ng](mailto:taiwo.adebayo@unn.edu.ng).



**Sochima Vincent Egoigwe**    holds a Bachelor degree in Engineering (B.Eng) in Electronics Engineering from the University of Nigeria, Nsukka, Enugu State, a Master's degree (M.Eng) from Enugu State University of Science and Technology (ESUT) in Electrical and Electronic Engineering and presently pursuing Ph.D. degree in the same course from Enugu State University of Science and Technology (ESUT), Nigeria. His research interest includes automatic control systems and modeling. He has attended conferences where he presented papers. He can be contacted at email: [sochima.egoigwe@unn.edu.ng](mailto:sochima.egoigwe@unn.edu.ng).



**Matthew Chinedu Odo**    is a Lecturer in the Department of Mechatronic Engineering, University of Nigeria Nsukka, Nigeria. He has both B.Eng. and M.Eng. Degrees from the same University of Nigeria Nsukka, Nigeria. He has published in peer-reviewed journals and presented papers in refereed conferences. He is a COREN registered Engineer and a member of the Nigerian Society of Engineers (NSE) and the Nigerian Institute of Electrical and Electronic Engineers (NIEEE). His research interests are in the areas of instrumentation and automatic control systems, artificial intelligence, mechatronics, and control of power electronic converters. He can be contacted at email: [matthew.odo@unn.edu.ng](mailto:matthew.odo@unn.edu.ng).