FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 6, June, 2025, pp 369 – 374
DOI: https://doi.org/10.33003/fjs-2025-0906-3729

EVALUATING THE PREDICTIVE POWER OF DISCRIMINANT ANALYSIS FOR CORONARY HEART DISEASE RISK ASSESSMENT

***Shazali Umar Madaki, Musa Uba Muhammad, Sani Salihu Abubakar, Zakariyya Ibrahim Musa and Shamsuddeen Ahmad Sabo**

[1]Department of Mathematics and Statistics Kaduna Polytechnic
[2]Department of Statistics, Aliko Dangote University of Science and Technology, Wudil.

*Corresponding authors' email: shazaliumar6@gmail.com

ABSTRACT

Coronary heart disease (CHD) is a major public health concern. This study evaluated the predictive power of discriminant analysis for coronary heart disease (CHD) risk assessment. Using a sample of 1300 patients, the discriminant model identified significant predictors of CHD, including demographic and clinical factors. However, the model's performance was relatively low with an accuracy of 37.6%, specificity of 33.5%, sensitivity of 41.3%, precision of 40.8%, and F1-score of 0.410. Despite identifying key predictors, the model's limited performance suggests that further refinement or alternative models may be necessary for accurate CHD risk assessment.

Keywords: Discriminant, Coronary Heart Disease, Diabetes, Hypertension, Cholosteral, Logistic Regression

INTRODUCTION

Coronary artery disease (CAD) is one of the most serious heart diseases. Coronary arteries include the left anterior descending (LAD), left circumflex (LCX), and right coronary artery (RCA), which is further subdivided in to the left main artery (LMA), this includes LCX and LAD, and the right coronary artery (RCA). Having CAD indicates that at least one of these arteries is more than fifty percent.

Every year, the number of people suffering from cardiovascular disease rises as well. Numerous factors, including age, blood pressure, cholesterol, diabetes, hypertension, heredity, obesity, and bad lifestyle choices, contribute to the development of this disease. Physical indications such as dizziness, shortness of breath, exhaustion, and chest pain can be used to identify a variety of symptoms. Heart disease remains the number one killer worldwide; the world health organization reports that heart disease and stroke are responsible for 17.9 million dealth annually (Anthony et al., 2024).

Ciu et al., (2020), discovered that the logistic regression method was categorized as an efficient and successful algorithm in predicting the primary cause of cardiovascular disease, which was the issue addressed in the investigation. Fourteen cardiovascular performance related characteristics were used as variable to create a logistic regression model. It is discovered that there is a substantial correlation between the variables. As a result, there is often less chance of multicollinearity in the study. (Rajarathinam et al., 2024), investigated the connection between heart disease outcomes and factors using multivariate analysis. It classifies people according to their gender using LDA, finds significant differences using MANOVA, and normalizes data using the Box- cox approach and the finding emphasis how crucial heart disease metrics are to comprehending population traits. Priyadarshini et al.(2021), used binary logistic regression analysis to examine the relationship between the dependent variable (Y) and independent factors (X) of the logit function, with Y as a dependent variable and X as a continuous predictor variable. The evaluation uses logistic regression to model factors influencing coronary heart disease, evaluating its performance using MAPE, RMSE, and SSE. The Major risk factors include of age, gender, obesity, blood stress, ex-smokers, BMI, dyspnea, chest pain, and stenosis. Isnanto et al. (2023), compared the performance of logistic regression (LR) and predictive discriminant analysis (PDA) for the two-group classification problem examined in the Monte Carlo study. The classification process employed prior probability with the assumption that the cost of misclassification would be the same. Three factors were included in the study's fully crossover experimental design: sample size, prior probabilities, and equal/unequal covariance matrices. There were 200 replications in each cell. To give the investigation a replication mechanism, two data patterns were simulated. The principal conclusions are: When two groups have equal prior probability, PDA and LR perform similarly; when two groups have different prior probabilities, LR minimizes the error rate for the smaller group and PDA minimizes the error rate for both the bigger group and the entire sample. Abdulqader (2015) discovered that classification method to classify datasets using linear discriminant analysis. The dataset used was divided into 25 percent tests and 75 percent training. This classification is carried out fewer than two conditions. The first condition is the number of outputs consisting of 5 labels, and the second condition is only the number of labels with 2 outputs. The classification of performance measurement based on accuracy, precision, repeatability, and F1 value shows the results of the performance of the LDA algorithm in classifying heart disease using the two labels used as targets or results. Based on the results, the precision value is 0.82, the repetition value is 0.81, the F1 value is 0.81, and with an accuracy of 81.22 percent, and the confusion matrix that is found in classic heart disease with LDA at 2 targets or outputs. Kiyoshige et al. (2023), used Bayesian age-period-cohort (BAPC) models to estimate future CHD and stroke mortality projections in Japan, focusing on population estimates until 2040.

MATERIALS AND METHODS

Study Design

This study aimed to evaluate the predictive power of discriminant analysis for coronary heart disease (CHD) risk assessment.

Data collection

The data used in this research is a secondary data obtained from 1300 patients at cardiovascular outpatient ward Murtala Muhammad Specialist Hospital Kano.

Method of Data Analysis

The study employed the following statistical methods:

i. Discriminant Analysis: To identify the factors that discriminate between patient with or without CHD.
ii. Performance Metrics: To evaluate the accuracy, specificity, sensitivity, precision, and F1- score of the discriminant model.
iii. Chi-square: To determine the significance of the relationship between individual variable and CHD.
iv. Omnibus Chi-square: To assess the overall significance of the discriminant model.

Statistical Analysis

The discriminant analysis was used to develop a model that predicts the risk of CHD based on various factors. The performance metrics were calculated to evaluate the model`s accuracy. The chi-square test and omnibus chi-square test were used to determine the significance of the relationships and the overall model fit.

Discriminant Aanalysis

Suppose we have two multivariate normal populations with equal variance-covariance matrices, N μ_1, $\sum$) and N (μ_2 , $\sum$); where; μ_1 and μ_2 represents the mean vectors of populations 1 and 2 respectively; and $\sum$ is the common variance- covariance matrices of the two populations. The pdf of ith population ($i = 1,2$) is given as follows (Usman, 2012).

$$P_i X = \frac{1}{(2\pi)^{\frac{1}{2}}|\sum|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu_i)^{|}\sum^{-1}(x-\mu_i)\right] \quad (1)$$

The following represents the ratio of the densities of two multivariate normal populations.

$$\frac{P_1(X)}{P_2(X)} = \frac{\exp\left[-\frac{1}{2}(x-\mu_1)^{|}\sum^{-1}(x-\mu_1)\right]}{\exp\left[-\frac{1}{2}(x-\mu_2)^{|}\sum^{-1}(x-\mu_2)\right]} \geq K \quad (2)$$

$$exp\left[-\frac{1}{2}(x-\mu_1)^{|}\sum^{-1}(x-\mu_1)\right]-\left[-\frac{1}{2}(x-\mu_2)^{|}\sum^{-1}(x-\mu_2)\right]\geq K \quad (3)$$

Using the monotone increasing natural logarithm of the previous equation (3.3), we obtain

$$\left[-\frac{1}{2}(x-\mu_1)^{|}\sum^{-1}(x-\mu_1)\right]-\left[-\frac{1}{2}(x-\mu_2)^{|}\sum^{-1}(x-\mu_2)\right] \geq logK \quad (4)$$

In the preceding formula, the second term denotes the Mahalonobis square distance between $N(\mu_1 \ , \ \sum)$ and $N(\mu_2 \ , \ \sum)$ when appropriately selected (which can of course be one, in which case log k will be zero), the left hand side of the equation:

$$X^{|}\sum^{-1}(\mu_1-\mu_2)-\frac{1}{2}(\mu_1+\mu_2)^{|}\sum^{-1}(\mu_1-\mu_2) \geq \log k \quad (5)$$

The first expression of the equation (3.4) above is known as fisher's linear discriminant function which is linear in the component of the observation vector.

Let,

$$\bar{X}_i = \begin{bmatrix} \bar{X}_{i1} \\ \bar{X}_{i2} \\ \bar{X}_{13} \\ \vdots \\ \bar{X}_{ip} \end{bmatrix} \quad (6)$$

Where, $\hat{X}_i$ represent the sample mean vector and i denotes two groups (affected and not unaffected).

Let $\hat{X}_{i1,}$ $\hat{X}_{i2,}$ … , $\hat{X}_{ip}$ represent the individual mean vectors for the 13 variables that is $P = 13$ for example:

$$\bar{X}_{i1} = \frac{1}{k}\sum_{i=1}^{k} x_{i1} \quad (7)$$

Where, $\bar{X}_{11}$ is the mean of the first variable in group one, while $\bar{X}_{12}$ represent the mean of the first variable in group two, k is the number of the case and n is the sum of all observation in a particular group. The sample variance covariance matrix is given as:

$$S_i = \begin{bmatrix} s_{ii} & s_{ip} \\ s_{pi} & s_{pp} \end{bmatrix} \quad (8)$$

Where, s_i denote variance-covariance matrix, for $i = 1,2,\dots,s_{ii}$ denotes an individual variance and $s_{ip} = s_{pi}$ denotes an individual covariance for $p = 1,2,\dots,13$.

The illustrations are given below:

$$s_{ij} = \frac{1}{k}\sum_{i=1}^{n}\left(X_{ij}-\bar{X}_i\right)^2 \quad (9)$$

$i = 1,2 \quad j = 1,\dots,13$

$$s_{12} = \frac{1}{k}\sum_{i=1}^{n}(X_{i1}-\bar{X}_1)(X_{i2}-\bar{X}_2) \quad (10)$$

Let π_1 denotes group one (unaffected) and π_2 denotes group two (affected). The Euclidean distance of the unaffected is defined:

$$l_1 = \bar{X}_1^1 s_p^{-1}(\bar{X}_1-\bar{X}_2) \quad (11)$$

The Euclidean distance of the affected is:

$$l_2 = \bar{X}_2^1 s_p^{-1}(\bar{X}_1-\bar{X}_2) \quad (12)$$

Where, s_p denotes the pooled variance matrix.

The mean Euclidean distance used in this study for the two groups is given as:

$$\hat{M} = \frac{1}{2}(l_1+l_2) \quad (13)$$

And the discriminant function is calculated by:

$$Y = X^1 s_p^{-1}(\bar{X}_1-\bar{X}_2) \quad (14)$$

Where, $\hat{Y}$ denotes the estimate of the discriminant function, and $\bar{M}$ denotes the mean Euclidean distance for affected and unaffected groups.

$$X^1 = (X_1 \ X_2) \quad (15)$$

$$S_P = \frac{n_1 s_1 + n_2 s_2}{n_1+n_2} \quad (16)$$

Since $n_1 \neq n_2$ equation (3.16) will be used, but if $n_1 = n_2$ then estimated pooled variance S_P above becomes:

$$S_P = \frac{s_1+s_2}{2} \quad (17)$$

Where, s_1 and s_2 are respectively sample variance covariance matrices for the two groups and n_1

And n_2 is the sample size of the two groups respectively.

Where, s_1 and s_2 are respectively sample variance covariance matrices for the two groups and n_1 and n_2 are the sample size of the two groups respectively.

The fisher's linear discriminant model to be use is:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_P X_P \quad (18)$$

Y: is the dependent variable.

$X_{1,}$ $X_{2,}$ …,X_P: are independent variables.

β_1, $\beta_{2,}$ …, β_p: are the parameters or coefficients to be estimated.

Therefore, the classification rule is that:

$$if\ \hat{Y} \geq \hat{M} \quad classified\ as\ group\ one(\pi_1)\ \text{And} \quad (19)$$

$$if\ \hat{Y} < \hat{M}\ classified\ as\ group\ two\ (\pi_2) \quad (20)$$

Where $\hat{Y}$ denote the estimate of the discriminant function, and $\hat{M}$ denoted the mean Euclidean distance for unaffected and affected groups.

Performance evaluation metrics

Sensitivity (True Positive Rate): The proportion of actual coronary heart disease patients correctly identified by the model.

$$sensitivity = \frac{TP}{TN+FN} \times 100 \quad (21)$$

Specificity (True Negative Rate): The proportion of non-coronary heart disease patients correctly identified by the model.

$$specificity = \frac{TN}{TN+FP} \times 100 \quad (22)$$

Precision: This measure the proportion of patients correctly predicted to have CHD among all patients predicted to have CHD.

$$Precision = \frac{TP}{(TP+FP)} \quad (23)$$

Accuracy: The overall proportion of correct predictions (both CHD and non-CHD patients):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (24)$$

F1-score :(Harmonic mean of sensitivity and specificity): A balanced measure of sensitivity and specificity.

$$F1 - Score = \frac{(SE \times SP)}{(SE+SP)} \times 100 \quad (25)$$

Chi-square Test

In this study, we employed the chi-square test to investigate the association between independent variable and coronary heart disease risk. The chi-square test is a non-parametric statistical method used to determine if there is significant association between two categorical variables.

Test of hypothesis

H_o: There is no significant association between the independent variables and CHD risk

H_1: There is significant association between the independent variables and CHD risk

Test statistic

$$x^2 = \sum\sum \left(\frac{o-e)^2}{e}\right. \quad (26)$$

Where O_{ij} is observed value and e_{ij} is expected value.

Decision Criterion:

Reject H_0 if P<0.05 otherwise accept H_1 at the 5% level of significance.

Omnibus chi-square

In this study, we employed the omnibus chi-square test to examine the association between independent variable and CHD risk. The omnibus chi-square test is a statistical method used to determine if there is a significant association between a categorical independent variable and a binary dependent variable CHD risk. This test is particularly useful when dealing with multiple categories in the independent variable.

Test of hypothesis

H_0: There is no significant association between the independent variables and CHD risk

H_1: There is significant association between the independent variables and CVD risk

Test Statistic

$$x^2 = 2[\sum_{i=1}^{r}\sum_{j=1}^{c} (o_{ij}\ In\ \frac{O}{eij}] \quad (27)$$

Where;

χ^2_0--- Chi-square calculated

χ^2_v---- Chi-square value with v DF

Decision Criterion

Reject H_0 if P<0.05 otherwise accept H_1 at the 5% level of significance.

RESULTS AND DISCUSSION

Discriminant Analysis Results

Table 1: Group means

Response	DB (Yes)	HPT (Yes)	FMH (Yes)	SMK1(Yes)	GND (Male)
No	0.3306	0.8322	0.0526	0.1348684	0.5049342
Yes	0.4032	0.8309	0.0347	0.1272	0.4827

The table shows probabilities or risk scores for coronary heart disease based on various risk factors. For individuals with a "No" response, the probability is: 33.06% for DB (likely diabetes), 83.22% for HPT (hypertension), 5.26% for FMH (family history), 13.49% for SMK (smoking), and 50.49% for being male. For "yes" response, the probabilities are slightly different, with increased risk for DB (40.32%) and similar risks for other factors. These probabilities suggest a potential predictive model for CHD risk based on these factors.

Table 2: Group Means

Response	Divorced	Married	Married	Single	Single
No	0.03453947	0.7072368	0.1217105	0.06250000	0.01644737
Yes	0.06791908	0.4855491	0.2962428	0.06069364	0.02456647

The table shows CHD risk probabilities by marital status. Individuals with CHD (yes) are more likely to be divorced (6.79% vs 3.45%) and remarried (29.62% vs 12.17%) while less likely to be married (48.55% vs 70.72%). This suggests an association between CHD risk and marital status, particularly for divorced and remarried individuals.

Table 3: Group Means

Response	Widowed	Hausa	Igala	Igbo	Yoruba
No	0.02796053	0.3865132	0.01151316	0.09868421	0.1694079
Yes	0.03179191	0.4436416	0.01878613	0.08670520	0.1676301

The table shows probabilities of CHD risk by marital status and ethnicity. For "No" CHD, the probabilities are 2.8% for widowed, 38.65% for Hausa, 1.15% for Igala, 9.87% for Igbo, and 16.94% for Yoruba. For "Yes" CH, the probabilities are 3.18% for widowed, 44.36% for Hausa, 1.88% for Igala, 8.67% for Igbo, and 16.76% for Yoruba. This suggests a potential association between CHD risk and ethnicity, particularly increased risk for Hausa and widowed individuals.

Table 4: Group Means

Response	Secondary	Tertiary	Uneducated	Retired	Self employe
No	0.3849	0.5214	0.0411	0.0724	0.2599
Yes	0.3757	0.5217	0.0434	0.0737	0.2890

CHD risk probabilities show slight increases for uneducated (4.34% vs 4.11%) and self-employed (28.90% vs 25.99%) individuals, suggesting a potential link between CHD risk and lower education or certain occupations, particularly self-employment.

Table 5: Group Means

Response	Unemployed	Unemployed	Age	Weight	Cho
No	0.001644737	0.4967105	52.90625	68.39474	184.3257
Yes	0.005780347	0.4653179	53.54046	69.82225	172.6113

The table shows probabilities of CHD risk by employment status, age, weightm and cholesterol. For "Yes" CHD, individuals are more likely to be unemployed (0.58% vs 0.16%), slightly older (53.54 vs 52.91 years), and heavier (69.82 kg vs 68.39 kg), but have lower cholesterol levels (172.61 vs 184.33). This suggests a potential association between CHD risk and unemployment, age, and weight.

Table 6: Coefficients of linear discriminants

Factors	Coefficients
Intercept	-0.0032333
DB (Yes)	0.225731700
HPT (Yes)	-0.0436819383
FMH (Yes)	-0.5302514586
SMK1 (Yes)	-0.0978724559
Male	-0.1737646152
Married	-0.6263642936
Single	0.7182727460
Widowed	0.2392286713
Hausa	0.3800318695
Igala	0.9985913382
Igbo	-0.0799466420
Yoruba	0.3164300464
Tertiary	0.1339541566
Secondary	-0.1665115822
Uneducated	-0.1069369124
Retired	0.0055783299
employed	0.1757741390
Unemployed	0.5957952795
Age	0.0008248313
Weight	0.0127516526
Cholesterol	-0.0034784140

From table 6, the variables Igala (0.9985913382), single (0.7182727460) and unemployed (0.5957952795) have strongest positive associations, suggesting potential protective effects. Family History (yes) (-0.5302514586), marriage (-0.6263642936) have strong negative associations, suggesting protective effects. Diabetes (yes) (0.225731700), Hausa (0.3800318695), and employed (0.1757741390) have positive associations, indicating potential risk factors

Diagnostic Measures of Discriminant Analysis Model

Table 7: Confusion Matrix

Response	No	Yes
No	401	207
Yes	282	410

Individuals with CHD are more likely to have a "Yes" response (410 vs 207 without CHD), indicating a strong association between the response and CHD risk.

Table 8: Accuracy classification Measures

Accuracy	Specificity	Sensitivity	Precision	F1-Score
0.376	0.335	0.413	0.408	0.410

The accuracy of 37.6% indicates proportion of correctly classified instances out of all instances; this value represents the model's overall correctness. The specificity of 33.5% indicates the proportion of true negatives correctly identified by model. This value indicates the model's ability to detect negative case. The sensitivity of 41.3% indicates the proportion of true positives correctly identified by the model. This value represents the model's ability to detect positive cases. The precision of 40.8% indicates the proportion of true positives among all predicted positive instances. This value indicates the model's accuracy when predicting positive cases. And the F1- score of 41.0% indicates the harmonic mean of precision and sensitivity, providing a balanced measure of both. This value represents the model's overall performance in detecting positive case.

ROC Curve and Histogram of Discriminant Analysis Model

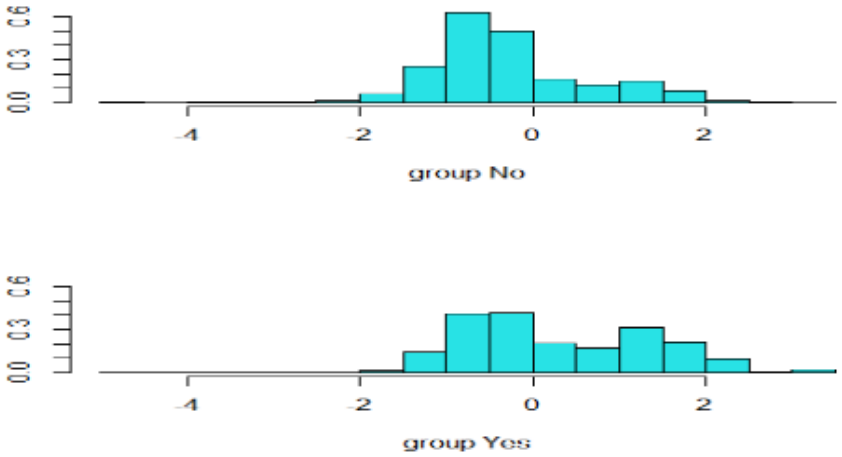

Figure 1: Histogram of CHD

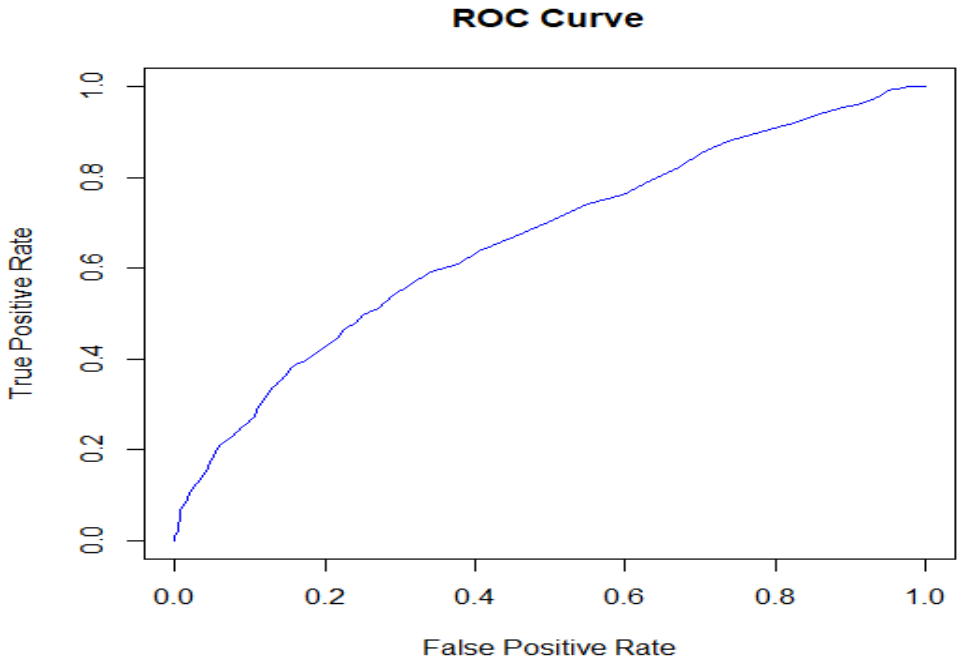

Figure 2: ROC curve of discriminant analysis model

The histogram displays the distribution of discriminant scores for individuals grouped by CHD status "No" and "Yes". The "No" group is centered around a mean slightly less than zero, with most values clustering near the center, indicating a relatively symmetric distribution. The "Yes" group, while also centered near zero, shows a slightly wider and more right –skewed distribution. This separation between the two distributions suggests that the discriminant function has some ability to differentiate between individuals with and without CHD, though there is considerable overlap, indicating that classification accuracy may be moderate.

The ROC curve for the Discriminant Analysis model predicting coronary heart disease risk indicates modest predictive performance, as the curve lies above the diagonal but lacks a strong bend toward the top left corner. This suggests the model performs better than randomguessing but with only fair discrimination ability. Visually, the estimated AUC appears to be between 0.6 and 0.7, reflecting limited effectiveness in distinguishing between individuals at risk and not at risk. Overall, while the model shows some utility, it may require improvement or supplementation for reliable clinical decision-making.

CONCLUSION

The discriminant analysis model identified key predictors of coronary heart disease, but its performance was relatively low, with an accuracy of 37.6%, specificity of 33.5%, and sensitivity of 41.3%, precision 40.8%, and F1-score of 0.410. These findings suggest that the model may not be effective in predicting CHD accurate, highlighting the need for further refinement or exploration of alternative models. Future research should focus on improving the model's performance and exploring other machine learning techniques to enhance CHD risk assessment and inform targeted interventions.

REFERENCES

Abdulqader, Q.M. (2015). Comparison of discriminant analysis and logistic regression analysis: An application on caesarean births and natural births data. *Yuzuncu Yil Universitesi Fen Bilimleri Enstitusu Dergisi,* 20(1-2):34-46.

Anthony M. Nwohiri, Adeyemi A. Laguda, Abidemi A. Olanite, Damilare D. Olabam-Ire (2024). Logistic Regression Technique for Cardiovascular Disease Prediction. *FUDMA Journal of sciences (FJS) ISSN online: 2616-1370 ISSN print: 2645-2944, Vol. 8 No 4, August, 2024, pp 266-275.*

Ciu, T. and Oetama, R.S. (2020). Logistic regression prediction model for cardiova Scular disease. *IJNMT (International journal of new media Technology),7(1):33-38.*

Isnanto, R. R., Rashad, I., and Widodo, C.E. (2023). Classisication of heart disease Using linear discriminant analysis algorithm. In *E3S Web of Conferences*,Volume 448, page 02053. EDP Sciences.

Kiyoshige, E., Ogata, S., O'Flaherty, M., Capewell, S., Takegami, M., lihari, K., Kypridemos, C., and Nishimura, K. (2023). Projections of future coronary heart disease and stroke mo- Mortality in japan until 2040: a bayesian 42 age-period cohort analysis. *The Lancet Regional Health-Western pacific*

Priyadarshini, E., Gayathri, G. R., Chakravarthy, S. E., Vidhya, M., and Memala, W. A. (2021). Analysis of heart disease using statistical techniques. *In Journal Of physics:* Conference series, volume 1770, page 012105. IOP Publishing.

Rajarathinam, A. (2024). Discriminant analysis for heart disease attributes

Usman, A. (2012). Statistical Methods for Biometric and Medical Research. *Mil- Lennium Printing and Publishing Company Limited, Kaduna, Nigeria, pages* 486-495.