Biomedical Statistics and Informatics
2025, Vol. 10, No. 2, pp. 39-45
https://doi.org/10.11648/j.bsi.20251002.13

Research Article

Prediction of Patients' Outcomes in Cardiovascular Disease

Awogbemi Clement Adeyeye[1, *], **Johnson Simeon Adedayo**[2], **Ilori Adetunji Kolawole**[1], **Oyeyemi Gafar Matanmi**[3]

[1]Statistics Programme, National Mathematical Centre, Abuja, Nigeria
[2]Statistics Department, Nasarawa State University, Keffi, Nigeria
[3]Statistics Department, University of Ilorin, Ilorin, Nigeria

Abstract

Cardiovascular Disease (CVD) remains a leading cause of mortality worldwide, necessitating effective prediction methods to improve patient outcomes. The study contributed to knowledge by using Support Vector Machine (SVM) to predict the outcome of patient at risk of CVD. This study explored the application of Logistic Regression (LR) and Support Vector Machine (SVM) models for predicting patient outcomes in CVD. In this analysis, patient medical record data was retrieved online from kaggle.com, comprising a dataset of 1,000 instances with 14 features relevant to cardiovascular health. Both Logistic Regression, implemented in SPSS, and SVM, executed using the R package, were employed for predictive modelling. From this study, SVM emerged as the superior model due to its ability to handle high-dimensional data and complex relationships. It has shown potentials in reducing the severity of cardiac diseases by accurately identifying at-risk individuals, thereby enabling timely intervention. The results indicated that the SVM model achieved an impressive accuracy rate of 98.7%, significantly outperforming the LR model of accuracy rate of 97%. Accurate predictions from the SVM model are vital for healthcare experts in identifying individuals at risk and formulating tailored treatment plans. Leveraging advanced machine learning techniques such as SVM can enhance the predictive capabilities regarding cardiovascular disease outcomes. This study underscores the importance of integrating these models into clinical practice to facilitate proactive healthcare measures and ultimately reduce cardiovascular morbidity and mortality rates.

Keywords

Cardiovascular Disease, Logistic Regression, Support Vector Machine, K-Nearest Neighbours

1. Introduction

According to [13], there are about 80 million deaths yearly, and over the years, cardiovascular disease (CVD) has been a prevalent killer-disease in the global village [9]. It is also known as a heart disease that has been discovered to be a major reason behind the high rate of mortality worldwide. It is a global health disorder triggered by underlying factors such as unhealthy diets, physical inactivity, intake of tobacco, and overweight that requires an immediate attention. There is a need to understand some of the primary causes of CVD through comprehensive analysis, its demographic disparities, prevention strategies, age of onset, mortality rates, global prevalence, therapeutic strategies for affected individuals, etc.

[*]Corresponding author: awogbemiadeyeye@yahoo.com (Awogbemi Clement Adeyeye)

Received: 14 February 2025; **Accepted:** 27 February 2025; **Published:** 6 June 2025

CVD covers a spectrum of conditions affecting the heart and blood vessels, such as coronary artery disease, heart failure, arrhythmias (cardiac irregularity), and cerebral apoplexy (stroke). The major causes of CVD are multi-factorial and include both modifiable and non-modifiable risk factors. The modifiable risk factors include unhealthy diet, physical inactivity, the use of tobacco, excessive alcohol consumption, obesity. These are life-style related factors that can be altered to reduce the risk of developing CVD, while non-modifiable risk factors include genetic predisposition, age, and gender. Other factors that significantly contribute to the development of CVD are diabetes mellitus, hypertension, high cholesterol levels, and chronic stress. However, CVD affects both men and women but there clear differences in prevalence and outcomes between the two aforementioned genders. Men tend to develop CVD at an earlier age compared to women. This later sign in women might be partly attributed to hormonal differences; oestrogen in premenopausal women provides some level of protection against CVD. After menopause, women's risk increases significantly, often surpassing that of men.

A logistic regression (LR) technique for predicting the risk of CVD was presented by [10]. This was to create an LR algorithm and build a prediction model that would foretell the development of CVD. Using some data characteristics, this dataset was employed to train the LR technique; a robust model that was created to accurately predict the existence of CVD in new patients. With an accuracy of 81%, a precision of 83%, and a recall score of 76%, the accuracy, precision, and recall key metrics were used to evaluate the model's efficacy. The model's accuracy was compared to alternative methods, such as K-Nearest Neighbours (KNN) and Decision Tree Classifiers (DTC), which yielded accuracy of 81% and 76%, respectively. The obtained results are of great significance for healthcare providers, the proposed model can assist in identifying those who are at high risk of heart diseases and allow for early implementation of prophylactic measures.

Patient medical record information was investigated and used in conjunction with an algorithm for logistic regression in order to make heart disease diagnoses [1, 7]. The results of the logistic regression have been utilized to achieve a high level of accuracy in the prediction of heart disease. To get the model coefficients needed for the equation, the experiment used an iterative form of the logistic regression test. Iteration 14 produced the best results, with an accuracy of 81.3495% and an average calculation time of 0.020 seconds. The percentage of space that lies beneath the Receiver Operating Characteristics (ROC) curve is 89.36%.

The use of logistic regression algorithm was adopted to generate predictions. The approach for feature selection from the dataset that is suggested in this study is information gain. Machine learning was used as a method to reduce the dimensions of the data. Five features of dataset description: time, serum creatinine, ejection fraction, age, and serum sodium were used to determine the outcome of information gain. Predictions were also made using logistic regression, and a data sharing ratio of 70% training data and 30% test data resulted in an accuracy of 0.8556. This demonstrates how feature selection with information gain can improve the accuracy of the logistic regression model [2].

A consideration of a categorical data of both male and female was made by [3]. This ratio may vary according to regions, and it is considered for the people of age group 25-69. This does not indicate that the people with other age group will not be affected by heart diseases. The researchers suggested that the problem may start in early age group, predicted the cause of heart disease, and discussed various algorithms and tools used for prediction of heart diseases.

Classification modelling techniques were used by digging up the information contained in cardiovascular disease data [11]. The model is created using training data and tested with data testing as a form of evaluation of results. The researchers used the lift chart and matrix method to evaluate the effectiveness of the model. Based on the research that has been done, they concluded that the most effective model for prediction of heart disease is Naive Bayes which found that the use of the Naive Bayes algorithm produces a better level of accuracy than the Decision Tree.

The application of support vector machine (SVM) to healthcare improved performance metrics on benchmark datasets was reported by [6]. This included hybrid classification methods that combined optimization algorithms with SVMs.

A CVD detection model based on the quantum-behaved particle swarm optimization (QPSO) algorithm and SVM classification model, named QPSO-SVM was proposed by [5] to analyse and predict CVD risk. The experimental results showed that the model QPSO-SVM out-performs other state-of-the-art prediction models considered in the research findings in terms of sensitivity (96.13%), specificity (93.56%), precision (94.23%) and F1 score (0.95%).

SVM was utilized by [8] to classify the presence of CVDs to possibly decrease the rate of misdiagnosis. They developed a model capable of accurately forecasting CVDs to minimize the deaths associated with these conditions. In the study, two types of SVM models were used; linear SVM and polynomial SVM. Accuracy, precision, recall, and F1 score have been evaluated for comparing linear SVM and polynomial SVM. The result clearly shows that polynomial SVM provides a better accuracy than linear SVM.

Lucid and easy-to-understand details of SVM algorithms along with applications in virology and viral biology were provided by [12]. However, machine learning techniques are useful for predicting heart disease. Implementing the machine learning method may be more effective in terms of cost.

Various approaches have been used to predict CVD accurately and with maximum accuracy. The methods used range from simple to hybrid methods with other methods aimed at increasing the accuracy of the classifier model. These methods include Bayes Naïve (BN), Random Forest (RF), Machine

Learning Predictor (MLP), Support Vector Machine (SVM), k-Nearest Neighbour (KNN), Linear Regression (LR), Decision Tree (DT), and Deep Convolutional Neural Network (DCNN). The method for pre-processing uses classification error rate and chi-square. Thus, the aim of this study is to predict high-risk patients living with CVD using the logistic regression and support vector machine models.

2. Methodology

Two machine learning models were used to predict patient outcome with CVD, and the models are LR model and SVM model. Logistic regression is a predictive model utilized to assess the correlation between the dependent variable (target), which comprises categorical data with a nominal or ordinal scale, and the independent variable (predictor). LR is a supervised machine learning algorithm widely used for binary classification tasks. It encompasses categorical data with an interval or ratio scale. This algorithm is also applicable for time series modelling to ascertain the interrelation among the involved variables. Logistic regression serves as a tool to forecast the probability of categorical dependent variables. Within logistic regression, the dependent variable is delineated as a binary variable, denoted by “1” for “yes” or “0” for “no”. The logistic regression model forecasts as a function of X. The assumptions underpinning logistic regression entail the following: binary logistic regression necessitates binary dependent variables, wherein the factor 1 level of the dependent variable should epitomize the desired outcome; the independent variables must exhibit independence from each other. Consequently, the model should exhibit minimal or negligible multicollinearity and a linear relationship with log odds.

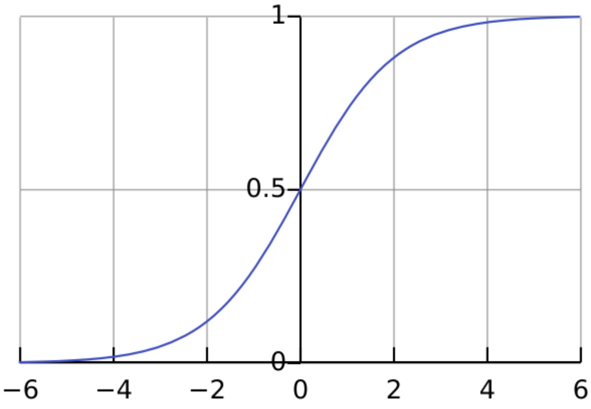

Figure 1. The standard logistic function of σ(t); σ(t) ∈ (0, 1) ∀t.

The logistic function is described mathematically as:

$$f(x)=\frac{L}{1+e^{-k(x-x_0)}}, \quad (1)$$

where L =1, k =1, $x_0 = 0$.

The standard logistic function: $\mathbb{R} \longrightarrow (0,1)$ is defined as:

$$f(x)=\frac{1}{1+e^{-x}} \quad (2)$$

Support vector machine is a supervised machine learning algorithm that classifies data by finding an optimal line or that maximizes the distance between each class in an N-dimensional space. It is also used for classification and regression tasks.

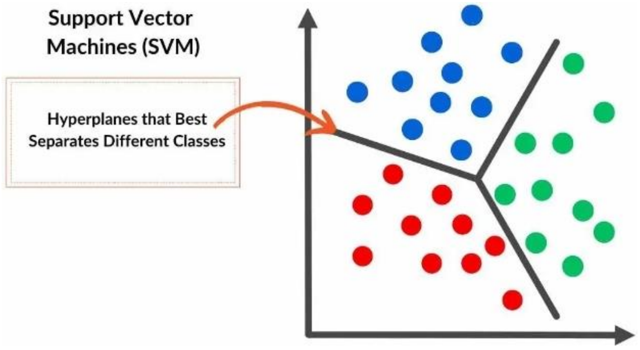
Support Vector Machines (SVM)
Hyperplanes that Best Separates Different Classes

Figure 2. Support Vector Machine Model Representation.

The points closest to the hyper-plane are called support vectors, which are important for defining the decision boundary. SVMs are widely applied in fields like image recognition, text classification, and bioinformatics due to their high accuracy and robustness against over-fitting.

Human cardiovascular system is examined in this study using some classification techniques variables that affect its performance. As shown on Figure 1, the process is started from: retrieve data, analyze the correlation between variables, split data, prediction with logistic regression algorithm, and finished with data validation.

This cardiovascular disease dataset is retrieved online from one of the multispecialty hospitals in India, via www.kaggle.com. Over 14 common features which make it one of the heart disease dataset available so far for research purposes. This dataset consists of 1000 subjects with 12 features. This dataset will be useful for building early-stage heart disease detection as well as to generate predictive machine learning models [4].

The logistic regression model was used for data analysis via SPSS to test the performance of the model in predicting high-risk patient outcome with CVD, while SVM was also used to test the effectiveness of the machine learning model via R package in forecasting patient that tested were diagnosed with CVD. Comparison was made between the two machine learning models in this study.

The preparation of the dataset for analysis and the features in the dataset are shown in Table 1.

Table 1. *Cardiovascular Disease Dataset Description.*

S. No.	Attribute	Assigned Code	Unit	Type of Data
1.	Patient Identification Number	Patient ID	Number	Numeric
2.	Age	Age	In Years	Numeric
3	Gender	Gender	1,0(0 = female, 1 = male)	Binary
4.	Chest pain type	Chestpain	0,1,2,3 (Value 0: typical angina Value 1: atypical angina Value 2: non-anginal pain Value 3: asymptomatic)	Nominal
5.	Resting blood pressure	Resting BP	94-200 (in mm HG)	Numeric
6.	Serum cholesterol	Serum cholestrol	126-564 (in mg/dl)	Numeric
7.	Fasting blood sugar	Fasting blood sugar	0,1 > 120 mg/dl (0 = false, 1 = true	Binary
8.	Resting electrocardiogram results	Resting relectro	0,1,2 (Value 0: normal, Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria)	Nominal
9.	Maximum heart rate	Max heart rate	achieved 71-202	Numeric
10.	Exercise induced angina	Exercise angia	0,1 (0 = no, 1 = yes)	Binary
11.	Oldpeak = ST	Oldpeak	0-6.2	Numeric
12.	Slope of the peak exercise ST segment	Slope	1,2,3 (1-upsloping, 2-flat, 3- down sloping)	Nominal
13.	Number of major vessels	No of major vessels	0,1,2,3	Numeric
14.	Classification	Target	0,1 (0 = Absence of Heart Disease, 1= Presence of Heart Disease)	Binary

The second stage creates a LR classification model that uses probability estimation for each class. LR is one of the supervised learning methods. However, the LR does not require a lot of parameter optimization and it is easy to implement.

The LR model operates as linear regression as seen in equation (3). However, the difference lies in the function used. In LR, the sigmoid function in equation (4) is employed within the equation. By substituting the sigmoid function into equation (1), equation (5) is derived. Equation (6) represents the formulation of logistic regression as the log probability function. The term inside the brackets is referred to as the odds, representing the ratio of the probability of success to the probability of failure. The LR coefficients are estimated using the iteratively reweighted least squares (IRLS) method. In each iteration, the dependent variable is adjusted to obtain the optimal LR coefficient:

$$\hat{y} = E(y/x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \in \tag{3}$$

$$\sigma = (z) = \frac{1}{1+e^{-z}} \tag{4}$$

$$E(y/x) = \sigma(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n) \tag{5}$$

$$E(y/x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_n x_n)}} \tag{6}$$

where,

$\hat{y}$ represents the predicted value of the dependent variable y given the independent variables x_1, x_2, ..., x_n;

β_0, β_1, ..., β_n are estimated parameters that determine the relationship between the independent variables and the dependent variable.

$\in$ represents the error term or residual;

Z represents the linear combination of the coefficients and independent variables.

SVM generally works by splitting data class based on the hyper-plane, as discussed above. The SVM function is shown in equation (7).

$$\text{LD} = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_j y_j x_i^T x_j \qquad (7)$$

L_D represents the SVM function;

α_i and α_j are the weights assigned to the data points;

y_i and y_j are the class labels, and x_i and x_j are the feature vectors.

The objective of SVM is to find the optimal weights that maximize the margin between the classes.

Performance metrics are used as reference for a benchmark in comparing the results. The formula for calculating accuracy is shown in equation (6) below. It also uses PPV (Positive Predicted Value) and FPR (negative predicted value) to get the Receiver Operating Curve (ROC) value. ROC here is valid for modelling errors, or errors from the built classification model. PPV and NPV can be seen in equations (9) and (10) below for the accuracy formula. True Positive (TP) means that it is correct and predicted correctly, True Negative (TN) is correct, but the prediction is wrong, False Positive (FP) is wrong but predicted correctly, and False Negative (FN) is wrong and predicted wrongly.

$$Acuracy = \frac{TP+TN}{Total\ \ Data} \qquad (8)$$

$$PPV = \frac{TP}{TP+FN} \times 100 \qquad (9)$$

$$NPV = \frac{TP}{TP+FN} \times 100 \qquad (10)$$

3. Data Analysis and Discussion

The research findings were from the results of logistic regression performance and student vector machine with the appropriate decision taken from the best model.

Table 1 shows the variables that are both significant and non-significant. The variables that are significant are gender, chest pain, resting bp, resting result, and old peak, with .000 respectively, while the variables that are non-significant are age, serum cholesterol, fasting blood, heart rate, exercise induced angina, and number of major vessel with .889, .646, .048, 0.009, .779, .578 respectively.

Table 2*. Parameter Estimate.*

Variables in the Equation						
	B	**S. E.**	**Wald**	**Df**	**Sig.**	**Exp(B)**
Age	-.002	.012	.019	1	.889	.998
Gender	3.146	.562	31.314	1	.000	23.246
Chest pain	1.083	.228	22.639	1	.000	2.953
Resting bp	.039	.008	24.420	1	.000	1.040
Serum cholesterol	.001	.002	.211	1	.646	1.001
Fasting blood	.964	.487	3.914	1	.048	2.622
Resting result	1.181	.300	15.461	1	.000	3.257
Heart rate	.016	.006	6.874	1	.009	1.016
Exercise induced angina	-.114	.405	.079	1	.779	.893
Old peak	-1.283	.205	39.141	1	.000	.277
Slope of the peak exercise	7.309	.830	77.521	1	.000	1493.345
Number of major vessel	.122	.219	.309	1	.578	1.129
Constant	-18.976	2.385	63.303	1	.000	.000

Table 2 shows that the number of correctly classified (right diagonal) items were 405 and 565 respectively, while the number of wrongly classified (left diagonal) items were 15 and 15 respectively. The percentages of the correctly and wrongly classified cases are 97% and 3% respectively. From the binary logistic regression model, the analysis of the model of Cox and Snell R-square is 0.694 and Nagelkerke R-square is 0.934.

Table 3. Logistic Regression Model Confusion Matrix Table.

Predicted group Cross Tabulation			
	Predicted group		**Total**
	Healthy	**Unhealthy**	
Healthy	405	15	420
Unhealthy	15	565	580
Total	420	580	1000

Table 4. Support Vector Machine Model Confusion Matrix Table.

Predicted Group Cross Tabulation			
	Predicted group		**Total**
	Unhealthy	**Healthy**	
Unhealthy	413	7	420
Healthy	6	574	580
Total	419	581	1000

Table 3 reveals that the number of correctly classified (right diagonal) items were 413 and 574 respectively, while the number of wrongly classified (left diagonal) items were 7 and 6 respectively. The percentages of the correctly and wrongly classified cases are 98.7% and 1.3% respectively. From the two models, it can be justified that SVM outperforms LR model to predict patients' outcome in cardiovascular disease.

4. Conclusion

The study made an attempt to select the best and accurate model between logistic regression model and support vector machine model. It also revealed that SVM is a better model that can forecast cardiovascular disease premised on the patients' health details. A dataset of 1000 was utilized, and it has a total of 14 features pertaining to cardiovascular health. It was found that the classification error rate from each model's confusion matrix shows that the SVM error rate of 98.7% is better than that of LR error rate of 97%. This could be regarded as template for accuracy and precision. It was established that splitting data class based on the hyper-plane can be more reliable in making predictions with relatively high precision and speedy computation. In order to get a better and robust model, further research should be carried out by comparing some other machine learning models such as Deep Convolutional Neural Network (DCNN) and Bayesian Network (BN).

Abbreviations

CVD	Cardiovascular Disease
SVM	Support Vector Machine
LR	Logistic Regression
DTC	Decision Tree Classifiers
DCNN	Deep Convolutional Neural Network
BN	Bayesian Network
ROC	Receiver Operating Characteristics (ROC)

Acknowledgments

The authors wish to express their gratitude to the Management of National Mathematical Centre (NMC), Abuja, Nigeria for the opportunity given to them to participate in the Research Oriented Course (ROC) in Multivariate Analysis & Applications.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Anshori M., & Haris, M. S. (2022). Predicting heart disease using logistic regression. Knowledge Engineering and Data Science (KEDS), 5(2), 188-196.

[2] Anshori M., Haris, M. S., & Wahyudi, A. (2024). Logistic regression's effectiveness in feature selection with information gain in predicting heart failure patients. Journal of Enhanced Studies in Informatics and Computer Applications, 1(2), 35-39.

[3] Bharani B. R., Manjunatha S., Vijayalakshmi R., & Preethi S. (2024). Heart disease prediction using effective machine learning techniques. International Journal for Multidisciplinary Research (IJFMR), 6(2), 1-8.

[4] Doppala P. B., & Bhattacharyya, D. (2021). Cardiovascular disease dataset, Mendeley Data, V1, https://doi.org/10.17632/dzz48mvjht.1

[5] Elsedimy, E. I., AboHashish, S. M. M. A & Algarni F. (2024). New cardiovascular disease prediction approach using support vector machine and quantum-behaved particle swarm optimization. Multimed Tools Application, 83, 23901-23928.

[6] Guido, M., De Santis, F., & Pappalardo, G. (2024). Application of support vector machine to healthcare: Improved Performance Metrics. Journal of Healthcare Engineering, Article ID 123456. https://doi.org/10.1155/2024/123456

[7] Harrell J., & Frank E. (2015). Regression modelling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis (2nd ed). Springer International Publishing. https://doi.org/10.1007/978-3-319-19425-7

[8] Hoque, R., Billah M., Debnath A., Hossain, S. M. S., & Sharif, N. B. (2024). Heart disease prediction using support vector machine. International Journal of Science and Research Archive, 11(2), 412-420.

[9] Maghdid, S. S., & Rashid, T. A. (2022). An extensive dataset for the heart disease classification system, Mendeley Data.

[10] Nwohiri, A. M., Laguda, A. A., Olanite, A. A., & Olabamire, D. D. (2024). Logistic regression technique for cardiovascular disease prediction. FUDMA Journal of Sciences (FJS), 8(4), 266-275.

[11] Palaniappan, S., & Awang R. (2008). Intelligent Heart Disease Prediction System Using Data Mining Techniques. International Conference on Computer Systems and Application, Doha, 108-115.

[12] Shapshak, P., Balaji, S., Kangueane, P., Chiappelli, F., Somboonwit C., Menezes, L. J., Sinnott, J. T. (2019). Application of Support Vector Machines in Viral Biology. Global Virology III: Virology in the 21st Century. Springer, Cham. https://doi.org/10.1007/978-3-030-29022-1_12

[13] World Health Organization (2023). Cardiovascular diseases, https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 accessed on March 14, 2024.