FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 1, January, 2026, pp 111 – 123
DOI: https://doi.org/10.33003/fjs-2026-1001-4511

MACHINE LEARNING–BASED ASSESSMENT OF CLINICAL AND DEMOGRAPHIC PREDICTORS OF CD4 STATUS IN HIV/AIDS PATIENTS

***[1]Samaila Manzo, [2]Abdulhameed Ado Osi, [3]Mannir Abdu, [4]Abba Bello Muhammad and [5]Dauda Usman**

[1]Department of Statistics Binyaminu Usman polytechnic Hadejia, Jigawa state, Nigeria.
[2]Department of Statistics Aliko Dangote University of Science and Technology Wudil, Kano State, Nigeria.
[3]Department of Statistics Binyaminu Usman polytechnic Hadejia, Jigawa state, Nigeria.
[5]Department of mathematics Umar Musa Yar·adua University Katsina, Katsina State, Nigeria.

*Corresponding authors' email: samailaadamu47@gmail.com

ABSTRACT

Human Immunodeficiency Virus (HIV) is the pathogenic organism responsible for acquired immune deficiency syndrome (AIDS) in the human body. The pathogen targets and destroys white blood cells, weakening the body's defenses against infection. HIV is an infectious virus that primarily affects CD4+ T cells. As a result of this infection, the number of these cells steadily decreases, which are essential for protecting the body against foreign antigens, leading to AIDS over time. The main focus of this research is to determine the predictive techniques for evaluating the clinical and demographic factors that affect immune function among individuals living with HIV/AIDS. Clinical and demographic data were collected from the records department of Rasheed Shakoni Teaching Hospital in Dutse, located in the Jigawa Central Senatorial Zone of Nigeria. Data were extracted from HIV/AIDS patients' case files who received antiretroviral treatment (ART) between January 2019 and January 2024. Four classification algorithms were used: Logistic Regression, Artificial Neural Network, Naïve Bayes, and Support Vector Machine. The models' predictive performance was tested using four metrics. A total of 274 HIV/AIDS patients were analyzed. Logistic regression had AUC 0.9342, accuracy 0.9348; SVM had AUC 0.9243, accuracy 0.9565; ANN had AUC 0.8980, accuracy 0.0870; Naïve Bayes had AUC 0.6818, accuracy 0.6957. Logistic regression outperformed all. The results enhance prediction reliability and support better health planning, care, and HIV prevention.

Keywords: HIV/AIDS, SVM, ANN, NB, Logistic Regression

INTRODUCTION

The bacterium that leads to acquired immune deficiency syndrome (AIDS) in the human body is referred to as Human Immunodeficiency Virus, or HIV. The pathogen targets and destroys white blood cells, which weakens the body's defences against infection. HIV is an infectious virus that primarily affects CD4+ T cells. As a result of this infection, the number of these cells steadily decreases, disrupting the organisms that protect the body against foreign antigens and gradually leading to acquired immune deficiency syndrome (AIDS). As the number of CD4+ cells in untreated individuals continuously declines, the CD4+ cell count has emerged as an essential metric for selecting therapies and evaluating the efficacy of antiretroviral therapy (ART). (Février et al. 2011). Blood is an essential component of the body's immune system. White blood cells contribute to illness prevention. The white blood cell type known as T cells is essential to the body. Some T cells function as "supporting cells," ordering the rest of the cells to perform their work. HIV targets and kills dendritic cells, macrophages, and T cells, especially CD4 T cells. Three factors contribute to the reduced amounts of CD4+ cells in HIV infection: direct viral destruction of infected cells, infected cells having higher rates of apoptosis, and CD8 lymphocytes with cytotoxic activity that identify and destroy infected CD4+ T cells. When an individual's CD4+ cell count drops below a specific number, immunity controlled by these cells declines, leaving the body more vulnerable to infectious diseases. As too many cells are eliminated, the immune system malfunctions, and the afflicted individual is diagnosed with acquired immune deficiency syndrome (AIDS) (Anubha 2014)

The CD4 cell count has emerged as a crucial metric for selecting therapies and assessing the effectiveness of antiretroviral therapy (ART). Furthermore, the total number of CD4+ T cells is essential for evaluating the severity of the health condition and determining the patient's prognosis. The CD4 cell count is a type of test conducted in a laboratory to measure the number of white blood cells in the human body. The required range is between 500 and 1,400 calls per cubic millimeter of blood. Physicians use this method to monitor the depletion of CD4 cells and assess the efficacy of antiretroviral treatment (ART). According to the Center for Disease Control and Prevention (CDC), one of the signs for the confirmation of AIDS is when the CD4 cell count falls below 200, which can lead to opportunistic infections and increased mortality.

Two-thirds (25.6 million) of the predicted 39.0 million HIV-positive individuals at the end of 2022 reside within the African WHO Region. In 2022, 1.3 million people were recently infected with HIV, and 000 individuals lost their lives to HIV-related illnesses. Although there are no permanent approved treatments to cure HIV, the condition can be effectively managed, allowing affected patients to live healthy lives for an extended period due to early detection, therapy, and care (WHO 2022).

HIV/AIDS has had a devastating impact on the world's population, particularly in West Africa. Sadly, Nigeria, the continent's most populous country, has only recently become aware of these ramifications. Nigeria's first two AIDS cases were discovered in 1985 and documented in Lagos in 1986, involving a 13-year-old female sex worker from one of the West African countries. Nigeria has a complex HIV epidemic that varies greatly by region. In certain areas, high-risk behaviors are the primary causes of the epidemic. Young people are more susceptible to HIV, with young women being at greater risk than young men.

Jigawa State has the lowest number of HIV cases in Nigeria, with approximately 0.3% of its population affected. The state currently has about 1,700 individuals affected by the disease

who are receiving life-saving ART treatment at various hospitals across the state (JISACA 2024).

The economic effects of AIDS are significant because it removes individuals from society who are at the peak of their careers and parenting lives. Growth and development suffer, productivity declines, earnings decrease, and poverty rises. According to World Bank estimates, AIDS is currently costing 24 African countries between 0.5% and 1.2% of their income annually. These factors deter investment, exacerbating the problem and impacting both business and government.

Machine learning models are used in various application domains and have proven to be highly valuable, particularly in data mining techniques where large amounts of data are utilized to build models and identify patterns in order to make better predictions.

The use of artificial intelligence approaches to clinical and demographic data related to HIV/AIDS is an emerging research area, where data is extracted and utilized to discover various factors affecting immune function among individuals living with HIV/AIDS, focusing on different baseline levels of CD4+ cells, which indicate the disease's status in patients, whether progressive or deteriorative, allowing for effective control, treatment, and planning. Numerous recent studies worldwide have focused on data mining and machine learning techniques, particularly in the healthcare sector, to predict factors contributing to low levels of CD4 cells among HIV patients using various machine learning methods.

Yashik Singh and Maurice (2010) used a classification model based on support vector machines to predict the extent of the shift in CD4 cell counts using several factors. The input variables included genotype, existing viral load level, education, age, and the number of weeks on ART medications. The model's accuracy was 83%. Based on genotype, viral load, and time, this early-stage experiment demonstrates that machine learning can accurately predict changes in CD4 count.

Sameem et al. (2010) proposed the classification and regression tree (CART) for predicting AIDS patient survival. Weight is the main factor contributing to low levels of CD4 and CD8, and CART was utilized for prediction modeling in medical datasets, which included demographic information. One drawback of their work is the low prediction accuracy. The model achieved an accuracy of 77%, which is not very satisfactory for the HIV/AIDS survivability issue. They suggested that incorporating WHO staging and other highly predictive variables could enhance the model's performance accuracy.

Muhammad *et al.* (2024) assessed the efficiency of six machine learning technique in predicting financial risk in microcredit industry, 5 performance metrics were used (i.e: Precision, Accuracy, Recall, F1-Score and AUC) on credit client's data. The outcome shows that KNN and ANN has outstanding performance in classification of credit applicant. It advocates for the adoption of modern techniques in credit scoring modelling, positioning K-nearest neighbour and Artificial neural networks as a valuable tool in financial institutions' risk assessment processes.

Agbelusi et al. (2015) demonstrated an algorithm for anticipating the survival of HIV/AIDS patients using the Naïve Bayes model by incorporating CD4 cell count as a deterministic variable. The following factors were considered: viral load, nutrition, and occasional infections. One limitation of the work is the large number of false negatives. In other words, the model failed to recognize some AIDS patients. Furthermore, only four variables were used, which may lead to limited accuracy in the results.

Ojunga et al. (2014) modeled the survival prospects of HIV-positive patients receiving highly active antiretroviral therapy (HAART) in Kenya's Nyakach District using logistic regression. Four distinct metrics were used to assess classifier performance: specificity, sensitivity, accuracy, and F1 score. The study utilized two variables, namely social and economic elements, influencing HIV patients' survival. In addition to providing policymakers with information on the variables affecting the survival of HIV-positive individuals on ARVs, the study offers a reasonable model for predicting the probability of survival among those attending the ART clinic in Nyakachi District. This strength demonstrates that considering socioeconomic factors can enhance the survival of infected individuals under study.

Muhammad *et al.* (2025) developed a hybrid ExpAR-FIGARCH-ANN to address the problem of volatility, nonlinearity, and long memory in residuals concurrently. Daily Nigeria All Share Stock Index Data (2001-2019), exhibiting these characteristics was used to assess the forecast performance of the new hybrid Exponential Autoregressive – Fractional Integrated Generalized Autoregressive Conditional Heteroscedasticity – Artificial Neural Network (ExpAR-FIGARCH-ANN) model in comparison to the existing Exponential Autoregressive – Fractional Integrated Generalized Autoregressive Conditional Heteroscedasticity (ExpAR-FIGARCH) and Artificial Neural Network (ANN) models using error-based metrics, viz Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE) and Mean Squared Error (MSE). The empirical findings show that the hybrid ExpAR-FIGARCH-ANN model outperformed the standalone ExpAR-FIGARCH and ANN models.

Oluyemi et al. (2016) employed a support vector machine and a variety of variables, including age, CD4, viral load, opportunistic illness, and nutrition. The research was able to forecast changes in immune adjustment. The results of the experiment show a 97.7% survivorship among pediatric patients in southwestern Nigeria who had HIV/AIDS.

Bingxiang Li et al. (2022) used clinical data from HIV patient records in Yunnan, China. Three algorithms SVM, RF, and MLP were employed to build a model that predicts changes in immune function. The model achieved an accuracy of 80.6%, with RF outperforming the other two models in predicting patients with a CD4 cell count greater than 200.

Singh and Mars (2020) implemented a machine learning model called support vector machine, using extracted facts to predict changes in CD4 cell count for HIV/AIDS patients by utilizing DNA sequencing, existing levels of viral load, and the frequency of hospital follow-ups as prognostic factors. The efficacy was evaluated using four metrics: sensitivity, specificity, accuracy, and ROC. The model's forecasting precision was 74.35%.

Umar *et al.* (2025) study the predictive performance of traditional Probit regression and several machine learning models in predicting Bronchopulmonary Dysplasia (BPD) among preterm infants. The models were evaluated using standard performance metrics, i.e: accuracy, precision, specificity, sensitivity, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Among all models, the Random Forest demonstrated superior predictive performance with the highest accuracy, indicating a strong discriminative ability. The findings suggest that machine learning approaches, particularly the Random Forest algorithm, provide a more robust predictive framework than the conventional Probit regression model for early detection of BPD risk in preterm infants.

Niedja Masristone et al. (2022) applied stepwise multiple logistic regression and the chi-squared test to investigate the

association between the HIV virus and age, gender, residence, education, number of donations, and serological tests for the hepatitis B virus surface antigen (HBsAg), hepatitis C antivirus (anti-HCV), human T-lymphotropic virus types 1 and 2 (anti-HTLV 1 and 2), syphilis (VDRL), and the core hepatitis B antigen (anti-HBc). Among the variables investigated, HIV infection was found to be associated with age, education level, residency, type of donation, and serological status for VDRL and anti-HBc tests. The model's accuracy in predicting HIV survival is 57.1%.

However, the majority of this research incorporates a few factors, such as opportunistic infection, nutrition, sex, age, and viral load, to predict changes in immune function. Moreover, there is a need to include additional factors and algorithms to compare and predict the best contributors to changes in immune function among individuals living with HIV/AIDS, as HIV is associated with the onset of numerous other illnesses. These illnesses can occasionally be relatively mild and intermittent, but in some cases, they may be severe and chronic. Consequently, in this work, we expand our study by incorporating additional factors, including age, weight, frequency of hospital visits, marital status, TB category, WHO clinical stages, HIV status, functional status, residency, viral load, and education to predict changes in immune function among individuals living with HIV/AIDS, using the baseline level of CD4+ cell count for conditions either progressing above 200 or deteriorating below 200.

MATERIALS AND METHODS

Ethics Approval

The research protocol used in the present study has been reviewed by the ethics committee of Rasheed Shekoni Teaching Hospital Dutse. Informed consent was obtained from hospital management as well as head of record department of the teaching hospital before taking the data from patient's case file.

Dataset and attribute

This study uses a dataset collected from Rasheed Shekoni Teaching Hospital in Dutse, located in the Jigawa Central Senatorial Zone, Nigeria. A total of 274 data points were collected from patient case files in the records department of the teaching hospital from 2019 to 2024.The distribution of the respondents is as follows: 143 respondents, representing 52.2%, are male, while the remaining 131 respondents, representing 47.8%, are female. The dataset contains demographic and clinical information, comprising 13 variables described in the table below.

Table 1: Identification of Clinical and Demographic Factors that Affect Immune Function among HIV/AIDs Patients

S/N	Variable name	Description	Type of variable
1	Age	Numeric	Input variable
2	Viral load	Numeric	Input variable
3	Sex	Categorical (M= male, coded as 1 F= female coded as 0)	Input variable
4	Education	Categorical (0= non formal education 1= primary, 2=secondary,3= tertiary,4= Islamic education,	Input variable
5	Marital status	= single coded as 0, M= married coded as 1 an D= divorce coded as 2	Input variable
6	Weight	Numeric	Input variable
7	Residency	Patients reside in his locality or staying outside his locality Binary (yes or no)	Input variable
8	Frequency of hospital follow ups	Number	Input variable
9	TB category	Binary(extra pulmonary=1, pulmonary=0)	Input variable
10	Functional status	Patients level of disease Ambulatory=0, Bedridden=1, Working=2	Input variable
11	WHO clinical stages	Patients stages of disease as describe by WHO Stage=0, stage=1, stage=2, stage=3	Input variable
13	HIV status	Positive=1, negative=0	Input variable
14	CD4 cell count	Coded as Binary (1=CD4>200, 0=CD4<200)	output variable

Data Processing

Some files may exist with missed or abnormal values, so we performed data cleaning to delete them; files with missing record of CD4 cell count were also deleted. In addition, we also performed data transformation in some variables including Gender, TB status, HIV status, WHO clinical stages, functional status, residency and CD4 cell count were also converted to dichotomous.

Training and Testing Dataset

Partitioning the data into training and testing sets is a common machine learning technique. A popular method for better model selection in classification prediction is K-fold cross-validation. The whole dataset was divided into ten folds for this study's stratified tenfold cross-validation method. The model was trained using nine folds, and it was tested using the final fold. With the testing fold, the process is carried out ten times. Additionally, because the trained dataset was small and the model's overall performance is the average of all ten folds, cross-validation was a helpful technique for this investigation.

The dataset has been divided into units for testing and training, which is an important aspect of data mining models. Following processing and cleaning, 80% of the data were utilized for training, and the remaining 20% was used for testing. After testing the model on the training data, we compared our training model to the test data, which had never been seen before.

Exploratory Data Analysis

Three approach were used to explore the relationship between CD4 cell count and other factors

Plot of Bar Chart and Histogram

Figures of the bar chart for categorical variables and the histogram for continuous variables were displayed to understand the patterns of the variables.

Chi-square test of independence

We used the chi-square test to assess the statistical significance of categorical variables and their association with immune function among individuals living with HIV/AIDS,

specifically their relationship with CD4 in either a deteriorative or progressive condition.

$$\frac{\sum_{i=1}^{n}(o_j-e_i)^2}{e_i} \quad (1)$$

Where o_j is the observed frequency.

Where e_i is the expected frequency.

Independent samples t-test

We also applied an independent sample t-test to compare the significance of the relationship between CD4 cells and other continuous variables. The t-test statistic is given below

$$t = \frac{\ddot{x}-\bar{y}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t n_1 + n_2 + 2 \quad (2)$$

$$s_p = \sqrt{\frac{1}{n_1+n_2+2}\sum(x-\bar{x})^2 + \sum(y-\bar{y})^2} \quad (3)$$

Where $\bar{x} = \frac{\sum x}{n}$, and $\bar{y} = \frac{\sum y}{n}$

Logistic Regression

Logistic regression is one of the most widely used machine learning techniques for binary classification. The algorithm performs well on a variety of problems. It is utilized when the data is linearly classifiable and the output is binary or dichotomous, but it can be adapted when the predicted variable contains more than two categories.

Logistic regression enables the prediction of a discrete output, such as group membership, from various variables that can be binary, continuous, discrete, or a combination of these types. However, the explained variable in logistic regression is binary, such as yes/no, presence/absence, and success/failure. The explained variable usually takes the value of one (1) with the probability of success (progressive condition), or zero (0) with the probability of failure (deteriorating condition). Problems of this nature are called Bernoulli (binary) variables.

Recall, in linear regression, the target variable is related to the features via the linear relationship:

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 +, \ldots, + \beta_k x_k + e \quad (4)$$

Suppose p(y) is the probability that a patient's health condition is progressive (we could write p(y=1) but stick to the shorter notation) or deteriorative p(y=0) to relate p(y) to the features you might consider the following:

$$p(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 +, \ldots, + \beta_k x_k \quad (5)$$

Unfortunately, the specification can generate the values for p(y) from -∞ to ∞. We need a model that generates probabilities in the 0 to 1 range. This is not guaranteed to be the case if we use equation (1.2). Furthermore, linear regression assumes the values of y are normally distributed. In logistic regression, y takes the values 0 or 1, so this assumption is clearly violated.

We need a more appropriate transformation. This can be achieved using the logistic regression model:

$$\log_e\left(\frac{p(y)}{1-p(y)}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 +, \ldots + \beta_k x_k \quad (6)$$

Log of odds ratio: In logistic regression, a function known as logit is required to map the linear combination of factors that could give any value between 0 and 1 or to link independent variables.

$$\ln(\text{odds}) = \ln\frac{p}{1-p} = \text{logit}\,(P) \quad (7)$$

In logistic regression, we evaluate an unknown probability for any given linear combination of explanatory variables.

We interpret e^{β_i} as the effect of the independent variables or features on the odds ratio. For example, if we postulate the logistic regression;

$$p(y) = \frac{e^{\alpha+\beta_1 x}}{e^{(\alpha+\beta_1 x)}+1} \quad (8)$$

The logistic curve: The logistic curve or sigmoid function captures the relationship between a binary target variable and features. It is calculated as:

$$p(y) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} \quad (9)$$

Because the relationship between p(y) and x is non-linear, the parameter α and β do not have a straight forward interpretation as they do in linear regression. The curve is bounded by 0 and 1,

In logistic regression, the probability model is based on the Bernoulli distribution, where;

$f(x, p) = \theta$, if $y_i = 1$ $\quad 1-\theta$ if $y_i = 0$ Therefore,

$p(y_i = 1) = \theta = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$.

The likelihood equation is given by:

$$L = \prod_{i=1}^{n} p_i^{y_1}\ (1-p_i)^{1-y_1} \quad (10)$$

$$\log g(y_0, \ldots, y_n) = \log(\pi\, p_i^{y_i}(1-p_i)^{1-y_i})$$

$$= \log\{p_i^{\sum y_i}(1-p_i)^{\sum 1-y_i}\}$$

$E(y_i) = p_i = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$ for dichotomous variables.

We therefore obtain $\log\frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_1$ (11)

Support Vector Machines

Support Vector Machines are among the machine learning techniques. It is used for a number of applications, such as regression and classification, as well as outlier detection. This method involves mapping input to a higher-dimensional space. SVM created a hyperplane to partition the data into classes for the classification task, the distance between categorising edges was maximised, and the optimal hyperplane was created. For example, $\{x_i, y_i\}_i^N$ is the dataset with N samples. The attribute of the vector was indicated by $x_i \in R^D$ and y_i is the class label for x_i. To find out the optimal hyperplane, the fundamental formula of support vector machines was expressed as follows:

$$f(x) = w.x + b \quad (12)$$

where w (weight) was the diagonal vector to the hyperplane, which determines its orientation, x was denoted as the training sample and b (bias) was the distance between the origin and the hyperplane. The goal was to maximise the margin. Furthermore, SVM built the two planes called H_1 and H_2 in the manner described below:

$$H_1 \rightarrow w^T x_i + b = +1 \quad \text{for } y_i = +1 \quad (13)$$

$$H_1 \rightarrow w^T x_i + b = -1 \quad \text{for } y_i = -1 \quad (14)$$

The area for the negative class was $w^T x_i + b \leq -1$while that of the positive class was $w^T x_i + b \geq +1$

The following explains how the SVM optimisation problem was formulated:

$$\text{Minimize } \frac{1}{2}|w|^2 \quad (15)$$

$$\text{Subject to } y_i(w^T.x_i + b) \geq 1,\ \forall_i = 1, \ldots, N \quad (16)$$

After solving the problem above, the formula of w and b became:

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \quad (17)$$

$$b = \frac{1}{N_s}\sum_{i\epsilon s}(y_i - \sum_{m\epsilon s}\alpha_m y_m x_m) \quad (18)$$

hence the following constituted the format of the SVM decision equations:

$$f(x) = \text{sign}(w.x + b) \quad (19)$$

Kernel function

In the non-linear transformation of the original input data into a higher-dimensional space, the use of 'kernel trick' is employed to cater for some problems. A kernel type used in this work is radial basis function (RBF) A kernel function was specifically the crucial element required to increase the accuracy of the SVM approach. Additionally, by projecting

the data into a higher-dimensional environment, it was applied to intricate real-world applications. This is how the kernel function is often written:

$k(x_i, x_j) =< \varphi(x_i)\varphi(x_j) > 0$ (20)

Where the Gausian Radial Basis Function Kernel is $K(x_i + x_j) = e^{-\left(\frac{(x_i+x_j)}{2\sigma^2}\right)}$

The problem of SVM optimization became:

Minimize $\frac{1}{2}|w|^2 + c\sum_{i=1}^{N} \varepsilon_i$ (21)

Subject to $y_i(w^T.\varphi(x_i) + b) - 1 + \varepsilon_i \geq 0, \forall_i = 1, \ldots, N$ (22)

The following is the mathematical description of w and b obtained by solving the aforementioned problem:

$w^* = \sum_{i=1}^{N} \alpha_i y_i \varphi(x_i)$ (23)

$b^* = \frac{1}{N_s}\sum_{i \epsilon s}(y_i - \sum_{m \epsilon s} \alpha_m y_m \varphi(x_m))$ (24)

Additionally, below were the SVM decision equations:

$f(x) = \text{sign}(w^*.\varphi(x_i) + b^*)$ (25)

Where ε_i an error that should be minimised and C is the cost, which determines the trade-off between the minimisation of error and the maximisation of the classification margin.

Naïve Bayes

The foundation of the naive Bayes algorithm is the Bayes probability theory, which postulates that every attribute of a certain class in a dataset is independent. Accordingly, it is assumed that each of them makes an equal contribution to the classification task's result when applied to the dataset. This is referred to as the strong (naive) assumptions of independence. According to Pandey and Pal (2011), the Naive Bayes algorithm is a descriptive and predictive method for predicting a target tuple's class membership.

Naive Bayes is a straightforward stochastic classifier that works by adding together a set of probabilities. Model is particularly useful in medical science for patient diagnosis because it is easy to build and doesn't involve intricate iterative parameter estimates, It generally performs better than more intricate categorization methods. The Naïve Bayesian classifier is extensively used despite its simplicity in terms of performing very well. The posterior probability can be computed using the Bayes theorem.

$p(c|x)$, from $p(c), P(x)$ (26)

The following describes how the Naive Bayes Classifier operates:

i. A training set of strings A and the class labels C1, C2, …. Cm that correspond to them is provided. if X stands for each of the characteristics in A that need to be categorized.

ii. (Ci) – class prior probabilities is the probability of each class.

iii. Determine the conditional probabilities for every single component of X in relation to the classes Ci, that is $P(X|Ci)$.

iv. For each classification, we get the posterior probability $P(Xi|C)$ by utilizing the probabilities in (iii) above. The following is the Bayes theorem:

$P(Ci|X) = \frac{p(X_i/c)p(c)}{p(X_i)}$ (27)

where P(X) is the prior probability of each of the components of X.

v. As $P(C)$ is constant for all classes, the Naïve Bayes forecasting will favor the class whose posterior likelihood $p(X \mid Ci)\, p(C)$ is largest. If and only if, the classifier predicts that tuple X's label belongs to class Ci.

$P(X|Ci)\, P(Ci) > P(X|Cj)\, P(Cj)$ for $1 \leq j \leq m, j \neq I$ (28)

Hence, it is regarded as the most significant posteriori probability outcome. (Han et al., 2012).

Artificial Neural Network

An artificial neural network is a machine learning method that mimics the brain's neuronal system to model how people learn. It makes use of interconnected neurons that are stacked, including input, hidden, and output neurons. The computing units used in this approach are coupled by weights. The weights serve the same purpose as the brain's synapses, where neurons are linked from one to another. The signal intensity link is represented by the weight (wi). Signals having activation functions like sigmoid, tansig, softmax, and ReLU are typically activated by the stronger networks. Additionally, in this investigation, we employ two hidden layers as well as a sigmoid activation function.

Consequently, the ANN method can easily be defined mathematically as:

$\overline{Y} = f(\overline{X}, \overline{W})$ (29)

Where $\overline{Y}$ and $\overline{X}$ are denoted by output and input matrices. $\overline{W}$ is a vector of weight parameters expressing the connections inside the ANN.?

The input layer uses attribute arrays to gather data, while the hidden layer receives the input values. The output values of the j^{th} neuron y_j of vector $\overline{Y}$ are calculated using a weighted average of the input elements x and w,

$y_i = \theta(\sum_{i=1}^{N_i} W_{ij} X_i)$ (30)

The θ is the activation function (transfer function), N_i is the aggregate amount of i^{th} links to the j^{th} neuron and X_i is the output value from the previous layer of i^{th} neuron. The activation function (θ) is sigmoid activation function which is used to transfer the value of weighted sum of inputs to the output layer. The resultant activated node for the next input layer is therefore:

$f(\theta) = \frac{1}{1+e^{-\theta}}$ (31)

The sigmoid activation function outputs a value in the range (0; 1) which thus can be interpreted as a probability.

$X_j = \theta(y_j)$ (32)

When the ANN model is given input and output variables, the BP-based supervised learning procedure is utilized. The ANN model employs BP as the training rule with two hidden layers, which increases the weights of neurons w_{ij} depending on the computed errors to ultimately produce the expected outputs. The sum of squares difference between the target values and the expected outcomes is used to compute the error function (E) of the computed BP-based ANN:

$E = \frac{1}{2}\sum_{j}^{N_j}(y_j - t_j)^2$ (33)

Where t_j is the predicted number for neuron i in the output layer and N_j is the aggregate amount of output neurons. In ANN weights are modified sequentially. The BP-based Levenberg–Marquardt optimisation method is use in the ANN training.

Artificial Neural Network Structure

The network consists of two hidden layers with thirteen nodes and one predicted variable, the learning rate of the network was 0.01. The activation function is sigmoid, the dataset for ANN model was divided into two folds, 80% of the data was used for training the remaining 20% to evaluate the model.

Confusion Matrix

A table of data that demonstrate a classification model's performance is called a confusion matrix. Both binary and multiclass classifications are handled. Furthermore, it displays the model error, FP (False Positive), and FN (False Negative), as well as the precise projections, TP (True Positive) and TN (True Negative).

A detailed description of the confusion matrix is shown in Table 2 below

Table 2: Confusion Matrix

		Predicted value	
Actual value		**Positive**	**Negative**
	Positive	True positive count (TP)	False positive count (FP)
	Negative	False negative count (FN)	True negative count (TN)

The proposed model's ability to separate samples across different n classes where n ≥ 2, is clearly shown in the confusion matrix table. While False Positive (FP) and False Negative (FN) represent samples that were incorrectly classified, True Positive (TP) and True Negative (TN) indicate samples that were correctly classified. Employing the confusion matrix, the widely used evaluation model metrics of accuracy, sensitivity, specificity, and Kappa can be defined as follows:

i. The number of properly diagnosed HIV/AIDS patients (positive and negative) divided by the total number of HIV/AIDS patients is the accuracy.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (34)$$

ii. Sensitivity (also known as true positive rate or recall) shows how many positive classes (CD4>200 progressive conditions or not at risk of HIV) were accurately categorised, and it is calculated using the following formula:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (35)$$

iii. Specificity (also called as true negative rate) shows how many negative class (CD4<200 deteriorative condition or at risk of HIV/AIDS) were correctly classified, and it is computed using the following formula:

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (36)$$

iv. Kappa is used to estimate O and E represent observed and expected accuracy, they can be obtained from the confusion matrix marginal totals. This statistic has a value between 0 and 1: a value of 0 indicates that there is no agreement between the observed and anticipated classes, while a value of 1 is shows that the model prediction and the observed classes are perfectly concordant.

$$kappa = \frac{o-E}{1-E} \quad (37)$$

Area under the Receiver Operating Characteristics (ROC) curve

AUC is the area of the curve plotted by the graph of the true positive rate (sensitivity) against the true negative rate (specificity) for the different instances of test datasets used for testing the predictive model.

When evaluating the range of acceptable values for decision-making processes, the area under the ROC curve (AUC) approach gives an in-depth assessment of a predictor's accuracy. The diagnostic test's effectiveness increases with the size of the area. The following formula can be used to determine the ROC curve's AUC: Where t = (1 – specificity) and ROC (t) is sensitivity.

$$\text{AUC} = \int_0^1 \text{ROC(t)dt} \quad (38)$$

To put it straightforwardly, high precision means that an algorithm obtained considerably more appropriate outcomes than inappropriate ones, and high recall means that an algorithm produced most of the appropriate outcomes. Accuracy is another statistical assessment of how well an ordinal test finds or excludes a condition. To put it another way, accuracy is the proportion of accurate results—including true positives and true negatives—among all the cases examined. The F1-measure is the test's accuracy, AUC is used for model comparison, and sensitivity is the test's ability to correctly identify people who are not at risk of HIV/AIDS. It considers both precision and recall.

RESULTS AND DISCUSSION

This study uses a dataset collected from Rasheed Shekoni teaching hospital Dutse, located at Jigawa central senatorial zone Nigeria. A total of 274 data were collected from patients' case file at record department of the teaching hospital, the distribution of the respondents is as follows. 131 respondents are female representing (47.8%) while male has 143 respondents representing (52.2%), 39 of all of the respondent had CD4<200, while 235 of them had CD4 >200 22 (56.4%) out of 143 of the male patients had CD4 less than 200 and 121(51.5%) out of 143 had CD4 greater than 200 apart from the female side 17(43.6) out of 131 had CD4 less than 200 and 114(48.5) out of 131 had CD4 greater than 200. The mean ages of those with CD4 <200 is 43.85 with standard deviation of 14.56 and those with CD4>200 had a mean age of 39.38 with a standard deviation of 15.75. Table 3 and 4 show the characteristics of both categorical and continuous variables.

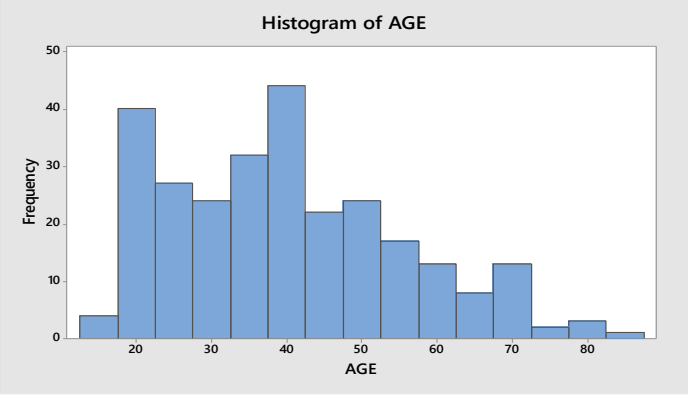

Figure 1: Distribution of Age Among Individuals Living with HIV/AIDS

Figure 1 depicts the distribution of age among individual patients. It is clear that individuals aged 20 to 50 years are among the patients affected by HIV/AIDS.

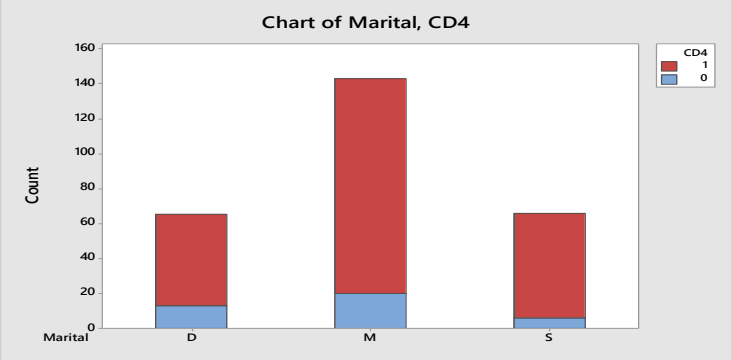

Figure 2: Patient Marital Status

The bars in Fig 2 represent the count of marital status of a patient with CD4 >200 and <200, we see that the majority of patients enrolled in the study were married.

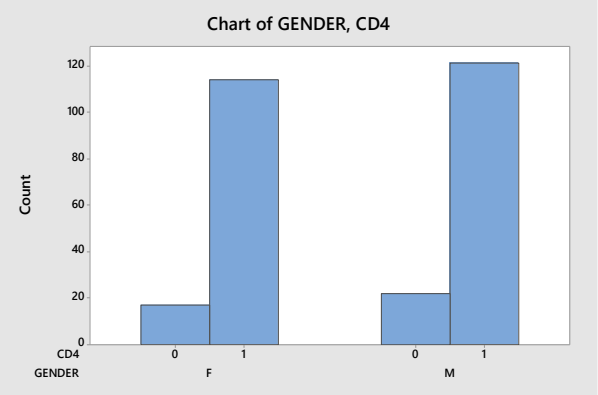

Figure 3: Distribution of Gender and CD4

Table 3: Characteristic of Study Participants (Categorical Variable) By HIV/AIDS Patients

S/NO	characteristics		All(n=274), n(%)	CD4< 200 (n=39) n(%)	CD4>200 (n=235) n(%)	P-value
1.	Gender	Female	131(47.80)	17(43.60)	114(48.50)	0.568
		Male	143(52.20)	22(56.40)		
2.	Marital status	Divorce	65(23.72)	13(37.14)	52(22.13)	0.200
		Married	143(52.18)	20(51.28)	123(50.34)	
		Single	66(24.09)	6(15.38)	60(25.53)	
3.	TB	Extrapulmonary	233(85.04)	35(89.80)	198(51.48)	0.354
		Pulmonary	41(14.96)	4(10.26)	37(15.74)	
4.	HIV status	Nagative	62(22.63)	3(7.69)	59(25.06)	0.008
		Positive	212(77.37)	36(92.31)	176(74.04)	
5.	Functional status	Ambulatory	138(58.73)	16(41.03)	122(51.91)	0.325
		Bedridden	87(37.02)	13(33.33)	74(31.49)	
		Working	49(20.85)	10(25.64)	39(16.60)	
6.	WHO clinical stages	Stage=0	124(45.23)	10(25.64)	114(48.51)	0.001
		Stages=1	61(22.26)	3(7.69)	58(24.68)	
		Stages=2	52(18.61)	3(7.69)	48(20.43)	
		Stages=3	38(13.86)	23(58.97)	15(6.38)	
7.	Residency	Urban=0	117(42.70)	16(41.03)	101(42.98)	0.819
		Rural=1	157(57.30)	23(58.69)	134(57.02)	
8.	Education	Illiterate =0	88(32.12)	15(38.46)	73(31.06)	0.041
		Primary=1	42(15.33)	7(17.95)	35(14.89)	
		Secondary=2	65(23.92)	12(30.77)	53(22.55)	
		Tertiary=3	57(20.80)	5(12.82)	52(22.13)	
		University=4	22(8.03)	0(0)	22(9.36)	

We used the chi-square test to determine the statistical significance of categorical variables used to identify evaluate clinical and demographic factors affecting immune function among individuals living with HIV/AIDS, specifically their association with CD4 for either in deteriorative or progressive condition. Table 3 display the result at 0.05 level of

significance was considered. The findings revealed a highly significant relationship between some demographic, clinical variables and CD4 cell (CD4 is used to measured changes in immune function) these factors include HIV status, WHO clinical stages and education. The results for patients' condition variables show a link between education, HIV status and WHO stages.

Table 4: Characteristic of Study Participants (Continuous Variable) By HIV/AIDS Patients

S/no	Characteristic	All (n=274)	CD4<200 (n=39)	CD4>200 (n=235)	p- value
1	Age years mean(SD)	40.000(15.60)	43.846(14.560)	39.380(16.000)	0.001
2	Weight mean(SD)	42.100(12.50)	43.410(10.818)	41.706(12.762)	0.001
3	Time mean(SD)	30.900(18.40)	23.282(14.681)	32.166(18.689)	0.001
4	Viral load mean(SD)	14405(14979)	14101(15763)	14456(14874)	0.001

The p-value of the independent t-tests performed for each of the continuous predictors are shown in table 4 the p-value for all of the continuous variables are less than the level of significance used in the test (=0.05), indicating that the variables was highly significance in determine the changes in immune function among individual living with HIV/AIDS.

Following the implementation of the four machine learning algorithm on the training dataset sample used for this study using the 10- fold cross validation, the results of the predictions made by the models on the 274 datasets used were plotted onto a confusion matrix in order to plot the true and false positives/negatives. Figures and tables shows the confusion matrix of the results of the implementation of the four different algorithms on the dataset used in this study. The predicted values are summed up along the vertical for both progressive and deteriorative condition while the actual values are summed up along the horizontal for both progressive and deteriorative condition in the dataset.

Logistic Regression

Model Building

The model was built in R software, and the performance was evaluated using test data. Table 5 shows the result of the confusion matrix, and Table 6 displays the results of the classification metric; AUC, Accuracy, Sensitivity, specificity and kappa are presented below.

Table 5: Logistic Regression Confusion Matrix.

	0	**1**
0	7	2
1	1	36

From table 5, the confusion matrix of logistic regression algorithm for changes in immune function among individual living with HIV/AIDS, 7 patients were correctly predicted with deteriorative condition, while 2 patients were misclassified with progressive condition. The model also misclassified 1 patient with deteriorative condition and correctly predicted 36 patients with progressive condition with a predictive accuracy of 0.9348%

Table 6: Logistic Regression Classification Matric

METRICS	SCORES
AUC	0.9342
ACCURACY	0.9348
SENSITIVITY	0.8750
SPECIFICITY	0.9474
KAPPA	0.7837

From table 6, the AUC=0.9342 shows a high discrimination, with good accuracy, sensitivity, specificity and kappa, the model shows a good performance.

Figure 3.2 Logistic Regression ROC curve

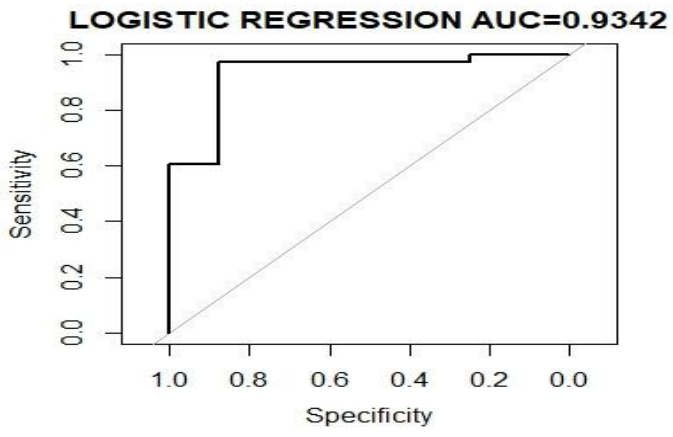

Figure 4: The AUC-ROC Shows 0.9342, Indicating A High Discrimination

Support Vector Machine

Model Building

The model was built in R software, and the performance was evaluated using test data. Table 7 shows the result of the confusion matrix, and Table 8 shows the result of the classification metric; AUC, Accuracy, Sensitivity, specificity and kappa are presented below.

Table 7: Support Vector Machine Confusion Matrix

	0	**1**
0	7	1
1	1	37

From table 7, the confusion matrix of the support vector machine algorithm for change in immune function among individuals living with HIV/AIDS, 7 patients were correctly predicted with deteriorative health condition, while one patient was misclassified with progressive condition. The model also misclassified one patient with a progressive condition and correctly classified 37 patients with progressive conditions with a predictive accuracy of 0.9565%.

Table 8: Support Vector Machine Classification Metrics

METRICS	**SCORES**
AUC	0.9243
ACCURACY	0.9565
SENSITIVITY	0.8750
SPECIFICITY	0.9737
KAPPA	0.8487

From Table 8, the AUC = 0.9243 shows good discrimination, with good accuracy. Sensitivity, specificity, and kappa all show excellent performance.

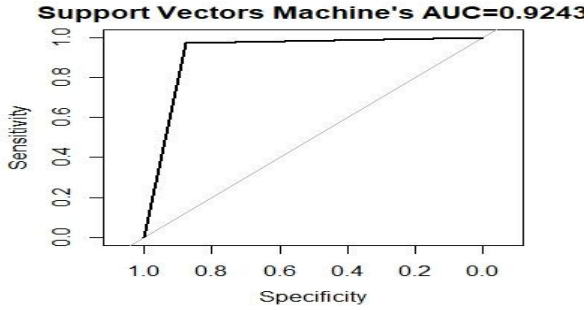

Figure 5: Support Vector Machine AUC Curve

From Figure 5, the result of the AUC shows a high discrimination.

Artificial Neural Network

Model Building

The model was built in R software, the performance was evaluated using test data and the result of the confusion matrix, together with the classification metrics: AUC, Accuracy, Sensitivity, specificity and kappa, are presented below

Table 9: Artificial Neural Network Confusion Matrix

	0	**1**
0	1	35
1	7	3

From Table 9, the confusion matrix of the Artificial neural network model for changes in immune function among individuals living with HIV/AIDS, 1 patient were correctly predicted with a deteriorative condition. While 35 were misclassified as a progressive condition. The model also misclassified 7 patients with deteriorative conditions and correctly classified 3 patients with progressive conditions with a predictive accuracy of 0.087%.

Table 10: Artificial Neural Network Classification Matric

METRICS	**SCORES**
AUC	0.8980
ACCURACY	0.087
SENSITIVITY	0.12500
SPECIFICITY	0.07895
KAPPA	-0.3343

From Table 10, the AUC= 0.8980 shows a moderate discrimination, with poor accuracy score. While precision shows a perfect performance.

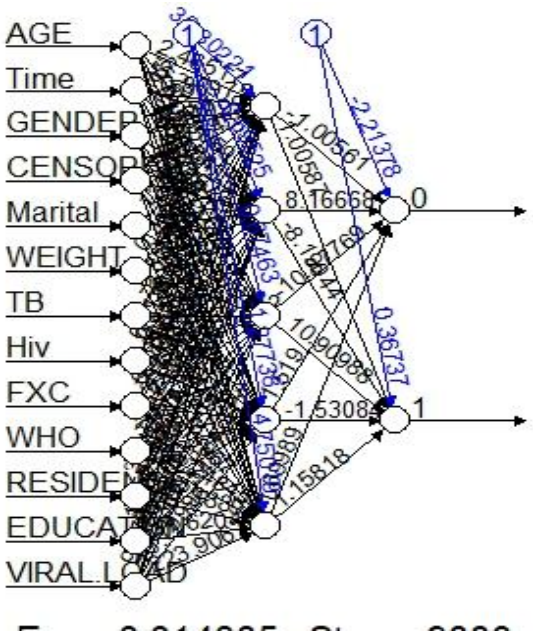

Figure 6: Network Structure of ANN

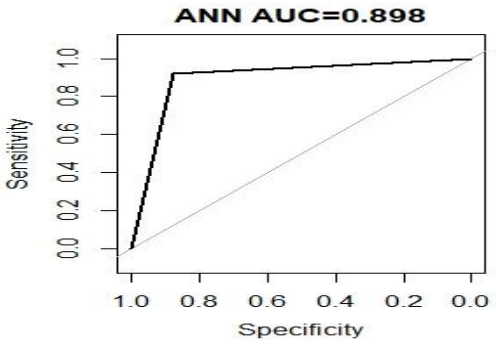

Figure 6: AUC-ROC OF ANN

From Figure 7, the result of the AUC shows no discrimination

Naïve Bayes

Model Building

The model was built in R software, and the performance was evaluated using test data. Table 11 shows the result of the confusion matrix, and Table 12 shows the result of the classification metric; AUC, Accuracy, Sensitivity, specificity and kappa are presented below.

Table 11: Naïve Bayes Confusion Matrix

	0	**1**
0	8	14
1	0	24

From table 11, the confusion matrix of the Naïve Bayes algorithm for changes in immune function among individual living with HIV/AIDS, 8 patients were correctly predicted with deteriorative condition, while 14 patients were misclassified with progressive condition. The model also misclassified 0 (no patients were misclassified) with a deteriorative condition, and lastly, the model correctly classified 24 patients with a progressive condition with a predictive accuracy of 0.6957%

Table 12: Naïve Bayes Classification Metrics

METRICS	**SCORES**
AUC	0.6818
ACCURACY	0.6957
SENSITIVITY	1.0000
SPECIFICITY	0.6316
KAPPA	0.3735

From Table 12, the AUC= 0.6818 shows a moderate discrimination, with moderate accuracy and poor F1 Score, recall, and precision show a very poor performance

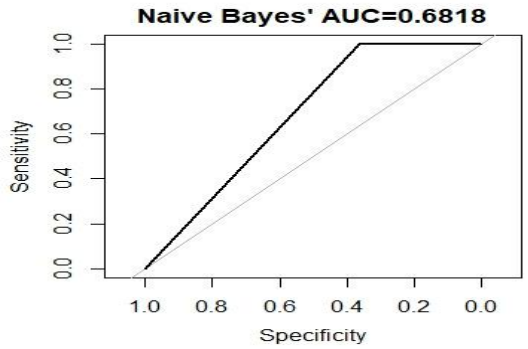

Figure 8: AUC-ROC of Naïve Bayes

From Figure 8, the result of the AUC shows discrimination on negative side, which indicates.

Table 13: Predictive Performance of Various Classification Techniques

ALGORITHMS	AUC	ACCURACY	SENSITIVITY	SPECIFICITY	Kappa
LOGISTIC REGRESSION	0.9342	0.9348	0.8750	0.9474	0.7837
SVM	0.9243	0.9565	0.8750	0.9737	0.8487
ANN	0.8980	0.0870	0.1250	0.0790	0.3735
NAÏVE BAYES	0.6818	0.6957	1.0000	0.6316	-0.3343

In order to compare the models and choose the one with the most accurate result, four algorithms were compared: LGR, SVM, ANN, and NV. The classifier's prediction results were summarised. Table 13 shows the sensitivity, specificity, and F1 values for each of the four classifiers. The summary statistics were generated using cross validation procedure. SVM achieved 0.9565 classification accuracy, 0.8750 sensitivity, 0.9737 specificity, and 0.8487 kappa. The classification accuracy of LGR was 0.9348, with a sensitivity of 0.8750, a specificity of 0.9474, and a kappa of 0.7837. In all matrices, SVM and LGR outperformed the remaining models. It is critical to have a high sensitivity to properly identify patients with a specific potentially progressive condition. Aside from accuracy, these two models have a higher sensitivity.

Discussion

In this study, we created models to predict and evaluate factors affecting immune function among HIV/AIDS patients. Data from Rasheed Shekoni Teaching Hospital were used. Our study predictive model can distinguish between deteriorative and progressive conditions of patients living with HIV/AIDS using CD4 cell count, which indicates the level of immune function in the body. The following models are compared in this paper: Logistics Regression (LGR), Support Vector Machine (SVM), Naïve Bayes (NB) and Artificial Neural Network (ANN). We used four metrics to assess their predicted performance. Based on our data set results, three of the predictive models in this study have a very good predictive ability. The result, however, revealed that the SVM classifier is the best model, followed by the LGR.

Many scholars have recently shown their interest in developing and comparing conditions of patients living with HIV/AIDS using predictive models, data mining and machine learning methods (Sameem et al 2010, Agbulusi et al 2015, Oluyemi et al 2015) reported that SVM as the best performing models based on their result. Also (Bingxiang et al 2022) also reported that RF as the best performing model based on their result, (Singh and Mars 2020, Niedja Masristone et al 2022) have been reported that LG as the best performing model based on their result. Lastly (Saurav et al. 2024, Zeming Li 2020, Basavarajiah 2020) reported that ANN is the best model in predicting factors affecting immune function among HIV/AIDS patients. None of these studies were achieved more accurate model than ours. Many studies have found that Decision tree, RF and linear discriminant analysis give a best predictive performance but our best model support vector machine outperforms the results of these models by predicting crucial factors affecting the immune function using clinical and demographic factors. Therefore, this study discovered that WHO clinical stages, HIV stage, education are significance factors that progress the condition of the disease. Table 3 and table 4 investigated the impact of categorical and continuous covariates on CD4 progressive condition Age, weight, Time, and Education are among the demographic factors that contribute to the progressive condition of the disease, while WHO clinical stages and HIV status were discovered to be significant among clinical factors affecting immune function. Several studies, including Bingxiang et al. 2022, Agbulusi et al. 2015, and Sameem et al. 2010, discovered that a patient's age, weight, education, viral load and nutritional status were all significantly related to immune function among HIV/AIDS patients, but did not include a WHO clinical, HIV stages and time to visit the hospital.

CONCLUSSION

The study concentrate on developing a predictive model for changes in immune function among individuals living with HIV/AIDS using clinical and demographic factors to identify the condition of an individual patients after the collection of historical dataset on the distribution of various factors that lead to changes of immune function among individual patients using CD4 cell count as response variable, with two different categories i.e. CD4>200 shows the condition of a patients is getting progressive while CD4<200 shows the condition of a patients is deteriorate. The data obtained in their respective case file include Age, weight, marital status, HIV status, WHO clinical stages, viral load, education, residency, TB, gender and functional status. Four algorithms were applied (logistic regression, support vector machine, artificial neural network and Naïve Bayes).

The anticipated techniques developed for the prediction of changes in immune function among individuals living with HIV/AIDS from Rasheed Shekoni teaching hospital, Dutse, Jigawa state, Nigeria. The outcomes of the study show that logistic regression models indicate an outstanding performance among four different models, with an AUC of 0.9342 and predictive accuracy of 0.9565 compared to other models. The predictive model of logistic regression is expected to give very promising results when used on other HIV/AIDS patients to determine their respective conditions. Also, it was discovered that the logistic regression algorithm was able to infer as much information from the historical dataset used for this study about the relationship between the risk factors affecting immune function among individuals living with HIV/AIDS. The model can also be integrated into existing Health Information Systems (HIS), which capture and manage clinical information which can be fed to the HIV/AIDS survival classification models, thus improving the clinical decisions affecting HIV/AIDS survival and the real-time assessment of clinical and demographic information. It is advised that a continual assessment of other variables that have a relationship with HIV/AIDS survival be made to increase the amount of information relevant to creating improved prediction models for HIV/AIDS.

REFERENCES

Abbass, A. K., Ali Yasir. (2008). "Congenital and Acquired Immunodeficiency" *Basic Immunology. Data status,* 209–223. (2018).

Anubha (2014). "Binary Text Classification Using an Ensemble of Naïve Bayes and Support Vector Machines". *GESJ: Computer Science and Telecommunications, 2(52), 37–45.*

Agbelusi O, Oluyemi Olufunke C, Olashinde O. (2015) "Evaluation of Immune Survival Factors in Pediatric HIV-1 Infection". *Annual National Academic Journal.* Vol. 91, No. 8. pp 298-312

Chen, M. jia, Yang C, J. (2018) "The Changing HIV-1 Genetic Characteristics and Transmitted Drug Resistance among Recently Infected Population in Yunnan China". *Epidiomol infection.* 146(6), 775-781 (doi 10.1017/5095026).

Center for Disease and Control (2024), "General Overview of HIV, Including Transmission Prevention, Testing and Control" https://cdc.gov/hiv/index.

Degninou Yehadji, Geraldine Gray, Carlos Arias Vicente, Petros Isaakidis. (2025) "Development of machine learning algorithms to predict Viral load suppression among HIV patient in Conakry (Guinea)". DOI: https://doi.org/10.3380/frial.2025.144687.

Erel orel, Rachel Estra, Janne Estill. (2022). "Prediction of HIV status based on socio- behavioural characteristics in East and Southern Africa" https://doi.org/10.1371/journal.pone.0264429.

Jantawan, B., Tsai, C. (2014). "A Classification Model on Graduate Employability Using Bayesian Approaches". *International Journal of Innovative Research in Computer and Communication Engineering.* Vol 2 Issue 6 Juni 2014.

Jessica P. Eri J and Jeria J.(2022). "Machine Learning and Clinical Informatics for Improving HIV care Continuum Outcomes". (doi: https://doi.org/10.1007/sl1904.021.00552-3

JISACA (2024) www. Punch newspaper on World HIV Day "Jigawa to Roll out Drugs at PHC" [Accessed on 17 September 2024].

Jaiteh Musa, Edith Phalane,Yeganew A, Shiferaw. (2025) *The Application of Machine Learning Algorithms to Predict HIV Testing in Repeated Adult Population–Based Surveys in South Africa: Protocol for a Multiwave Cross-Sectional Analysis* **(**IRRID): DERR1-10.2196/59916

Jialu Li, Yiwei Hao, Ying Liu, Liang Wu. (2024). "Supervised machine learning algorithms to predict the duration and risks of long term hospitalization in HIV-infected individuals: A retrospective study". DOI: https://doi.org/10.3389/fpubh.2023.1282324

Mao Q, Yingiao Dong, Shangbin Liu, Danni Xia. (2025). "Predicting disease progression in people living with HIV using machine learning and a nomogram: a 10- year cohort study based in Xinjiang, China" doi: https://doi.org/10.1136/bmjopen-2025-105026.

Miedema, F., et al. (2000). *The Dominant Source of CD4+ and CD8+ T-Cell Activation in HIV Infection is Antigenic Stimulation.*

Mutai Charles K, Patrick E Mcsharry, Innocent Nganye. (2021) Use of Machine Learning Techniques to Identify HIV Predictors for Screening in Sub-Saharan Africa**. (** https://doi.org/10.1186/s12874-021-01346-2)

Muhammad, A. B., Ishaq, O. O., Janet, B. B., Alhassan, M. A., Samaila Manzo, and Shehu, S. (2025)."A Hybrid ExpAR-FIGARCH-ANN Model for Time Series Forecasting". *Journal of Statistical Sciences and Computational Intelligence, 1*(4), 283–293. https://doi.org/10.64497/jssci.68

Muhammad, A. B., Olawoyin, I. O., Yahaya, A., Gulumbe, S. U., Muhammad, A. A., & Salisu, I. A. (2024). "Credit Risk Analysis: An Assessment of the Performance of Six Machine Learning Techniques in Credit Scoring Modelling**".** *FUDMA Journal of Sciences*, Vol. 8 No. 6, December 2024 (Special Issue), pp. 163 – 173. DOI*: https://doi.org/10.33003/fjs-2024-0806-2893*.

N. Cristianini and J. Shawe-Taylor (2019). *An Introduction to Support Vector Machines.* Cambridge University Press, 2000. Neidja M Barreto Q (2012) *logistics regression model for determining factors associated with HIV infection among blood donor*s www.rbhh.org (doi 10:5581/1516-8484.20120053)

Nigerian Bulletin, (2014). World HIV/AIDS Day: 10 Facts about HIV/AIDS in Nigeria You Probably Didn't Know. Available from www.nigeriabulletin.com/threads world HIV/AIDS day 10-fact-about-HIV-AIDS-IN- NIGERIA-you- probably-didn't-know.24303/. [Accessed 28 July, 2014].

Ngcobo Sanele, Edith Madela Mntla, Jonathan Shock. (2024) *Artificial intelligence for HIV care: a global systematic review of current studies and emerging trends* https://doi.org/10.1002/jia2.70045

Ojunga, N., Olaruwanju, and Olasihende.(2014) . The Application of Logistic Regression in Modeling of Survival Chances of HIV-Positive Patients under Highly Active Antiretroviral Therapy (HAART): *A Case of Nyakach District, Kenya. Journal of Medicine and Clinical Sciences*. Vol. 3, No 3. pp. 14-20.

Joshi, Kavita Sanjeev, Pranav Milind Ambardeka, Rushabh Yatish Gujarath.(2022) Logistic Regression-Based Parametric Analysis Of HIV-Associated Dementia Using A Screening Tool In A Tertiary Care Hospital In Mumbai DOI: https://doi.org/10.4103/ijstd.ijstd_80_21

Raper, J. L. (2007). Complete Blood Cell Count as a Surrogate CD4 Cell Marker for HIV Monitoring in Resource-Limited Settings.

Sameem, A., et al., (2010). Classification and Regression Tree in Prediction of Survival of AIDS Patients, *Malaysian Journal of Computer Science. Vol. 23. No 3. pp 153-165*

Singh, and Yashik, (2010) predicting immune risk in treatment of naïve HIV patient using machine learning algorithms (doi: https://doi.org/10.3389/fnut.2024-1443076)

Sendila Ernesy Asari, Rahmi Susanti, Ismail AB, rfansyah Baharuddin. (2025) Prediction Model of Human Immunodeficiency Virus Status at Abdoel Wahab Sjahranie Hospital, Samarinda Indonesia. DOI: https://doi.org/10.14710/jphtcr.v8i1.24302

Seboka Binyam Tariku, Delelegn Emwodew Yehualashet & Getanew Aschalew Tesfa (2023) Artificial Intelligence and Machine Learning Based Prediction of Viral Load and CD4 Status of People Living with HIV (PLWH) on Anti-Retroviral Treatment in Gedeo Zone Public Hospitals. https://doi.org/10.2147/IJGM.S397031.

Song, L., Biang Xi and Zan Yo. (2015). Multiple Introduction and Naturally Occuring Drug Resistance of HCV Among HIV Infected Intravenous Drug Users in Yunnan.

T. Joachims. (2018) Making large-scale support vector machine learning practical. In Advances in Kernel Methods: Support Vector Learning. B.

Umar Madaki, S., Bello Muhammad, A., and Ahmad Hamisu, H. (2025). "Predicting Bronchopulmonary Dysplasia in Infants: A Comparative Evaluation of Probit and Machine Learning Models**".** *Proceedings of 2025 International Conference on Data Science and Official Statistics (ICDSOS). 2025*(1), 660–665. *https://doi.org/10.34123/icdsos.v2025i1.617*

UNAIDS, 2012. Together We Will End AIDS. Available fromwww.unaids.org. [Accessed 12 June, 2013] UNAIDS, 2013. Global Report on HIV. Available from http://www.unaids.org/en/resources/documents/20 13/ name, 85053, en.asp. [Accessed 2nd August, 2014]

WHO, (2012). Towards Universal access: Scaling up Priority HIV/AIDS interventions in the Health Sector. Progress Report 2010.Available from http://whqlibdoc.who.int [Accessed January 15, 2014]

WHO, (2022). www.who. Int >Data>GHO HIV and Statistics. Data on the Size of the HIV Epidemic

Yashik Singh and Maurice Mars. (2010) Support vector machines to forecast changes in CD4 count of HIV-1 positive patients. ISSN 1992-2248 ©2010 Academic Journals

Zeming Li and Yanning Li (2020) journal *of medical informatics and decision making*. https://doi.org/10.1186/s12911-020-01157-3.