



Predictive Model of Pediatric HIV/AIDS Survival in Nigeria using Support Vector Machine

Olayemi Olufunke C.
Computer Science Department
Joseph Ayo Babalola
University, Ikeji, Osun state,
Nigeria

Olasehinde Olayemi O.
Computer Science Department
Federal Polytechnic, ile- oluji
Ondo State, Nigeria

Agbelusi O.
Computer Science Department
Rufus Giwa Polytechnic, Owo,
Ondo State, Nigeria

ABSTRACT

This paper is focused on the development of a predictive model for the classification of HIV/AIDS survival among Nigerian pediatric patients located in south-western Nigeria using supervised machine learning. Following the identification of the risk factors of HIV/AIDS survival from the review of literature and expert medical physicians, the case files of patients were used to collect information about the distribution of the risk factors and the HIV/AIDS survival status of pediatric patients selected at two hospitals in south-western Nigeria. The predictive model was formulated using the sequential minimal optimization (SMO) algorithm implemented by the support vector machine (SVM) – a binary classification algorithm based on the information collected. The predictive model was simulated using the Waikato Environment for Knowledge Analysis (WEKA) using the 10-fold cross validation technique for model training and testing. The SVM classifier performed well in the classification of the survival of pediatric HIV/AIDS patients with an accuracy of 97.7%. The predictive model developed can be useful to medical practitioners especially in the area of decision support regarding the survival of HIV/AIDS pediatric patients in Nigeria.

Keywords

HIV/AIDS survival, pediatric patients, support vector machines (SVM), sequential minimal optimization (SMO), predictive modeling.

1. INTRODUCTION

Human Immunodeficiency Virus (HIV) is the virus that causes Acquired Immunodeficiency Syndrome (AIDS) in human body. This virus attacks the white blood cells which makes it difficult for a person's body to fight off any infections. These infections would not cause problems for healthy people but for people living with HIV. They may cause severe life threatening problems. Blood plays a vital role in the body's immune system. White blood cells help to protect people from diseases. Certain white blood cells called T cells performs a crucial role in the body. Some of the T cells are "helper cells" that signal other cells to do their jobs. HIV attacks and destroys the T cells (specifically, CD4 T cells), macrophages, and dendritic cells. HIV infection leads to low levels of CD4⁺ T cells through three ways: direct viral killing of infected cells, increased rates of apoptosis in infected cells and killing of infected CD4⁺ T cells by CD8 cytotoxic lymphocytes that recognize infected cells. When CD4⁺ T cell numbers decline below a critical level, cell-mediated

immunity is lost, and the body becomes progressively more susceptible to opportunistic infections. When more than enough cells are destroyed, the immune system no longer works

and the patient is said to have Acquired Immunodeficiency Syndrome (AIDS) (Anubha, 2014)

For more than two decades now, the acquired immune deficiency syndrome / human immunodeficiency virus (HIV/AIDS) has been a challenge that affects all areas of the population worldwide. HIV is transmitted through direct contact of mucous membrane or the blood stream with body fluid containing HIV. This epidemic is highly contagious but it cannot be spread from person to person through the air. It is commonly spread through having unprotected sexual intercourse with an infected partner. It can only spread through contact with infected blood. This can occur when drug users share needles or syringes, blood transfusion and so on. (UNAIDS/ WHO, 2014).

Many people have no symptoms of HIV while some have flu-like symptoms few months after getting the virus. These may include fever, headache, tiredness and enlarged lymph node (these are organs of the immune system in the neck and groin). These symptoms usually disappear before they are confirmed as HIV

2. RELATED WORKS

Sameem et al,(2010) presented classification and regression tree (CART) in prediction of survival of AIDS patients. CART was used in predictive modeling on a medical dataset such as demographic data, weight and factors that deemed to affect the survival of AIDS, namely CD4 and CD8. The limitation of the work is that, there was a low predictive accuracy. He was able to achieve a recognition accuracy of 77% which is not so adequate in HIV/AIDS survivability problem. He recommended that the use of higher predictive variables, like adding WHO staging may increase the performance accuracy of the model.

Ojunga et al (2014) applied logistic regression in modelling of survival chances of HIV-positive patients under highly active antiretroviral therapy (HAART) in Nyakach District, Kenya. The aim of this study was to outline the various social and economic factors affecting survival of HIV patients under highly active antiretroviral therapy (HAART). The study was expected to provide suitable model for predicting the chances of survival among the HIV positives attending ART clinic in

Nyakachi District and also provide information for policy makers on the factors affecting survival of HIV positive ARVs. The strength shows that the survival of infected patient under study can be improved if their access to socio-economic factors is considered. The outcome may only be obtained in services that have smaller numbers of patients. Socioeconomic factors are not enough to predict survival as CD4, CD8, viral load, opportunistic infections and nutritional status were added to the existing study in this paper as predictive factors.

Agbelusi et al., (2015) presented Naïve Bayes model for predicting HIV/AIDS survival in paediatric patients. The variables used were CD₄ count, viral load, opportunistic infections and nutritional status. The limitation of the work is that, there were high false negatives. That is, there were AIDS patients which the model could not detect. Also, the variables used were only limited to four, which may limit the accuracy of the result.

3. MATERIALS AND METHODS

In this section of the chapter, the methodology applied in the development of the predictive model for the prediction of the survival of pediatric HIV/AIDS patients was presented. The section covers areas including data identification and collection, the supervised machine learning algorithm used, model formulation and simulation environment and the model validation process. They are elaborated in the following paragraphs.

a. Data identification and collection

Following the review of related works in the area of the risk factors associated with the survival of pediatric HIV/AIDS patients and the application of data mining techniques in developing predictive models, the risk factors associated with the survival of HIV/AIDS were identified and presented. Figure 1 shows a description of the variables that were identified from literature to be predictive of the survival of HIV/AIDS patients including pediatrics. These variables were the CD4 count, viral load, nutritional status and opportunistic infection. They are explained as follows:

- a. The CD4 count is a lab test that measures the number of CD4 T lymphocytes (CD4 cells) a type of white blood cells that play a major role in protecting the body from infection) in a sample of blood. In people with HIV, it is the most important laboratory indicator of how well your immune system is working and the strongest predictor of HIV progression. The normal range for CD4 count is between $600 - 1500 \times 10^6/L$ and any value equal and more than is High else Low.;
- b. The viral load is a measure of the number of viral particles present in an organism or environment, especially the number of HIV viruses in the bloodstream. The more HIV there is in blood (and therefore the higher the viral load), then the faster the CD4 cell count will fall, and the greater the risk of becoming ill because of HIV. The viral load was allocated values: High and Low;
- c. An opportunistic infection is an infection caused by pathogens (bacteria, viruses, fungi, or protozoa) that take advantage of an opportunity not normally available, such as a host with a weakened immune system, an altered micro-biota (such as a disrupted

gut flora), or breached integumentary barriers. The opportunistic infection was assigned labels Yes and No for the presence of opportunistic infection or not respectively; and

- d. Nutritional status is the condition of the body in those respects influenced by the diet; the levels of nutrients in the body and the ability of those levels to maintain normal metabolic integrity.

Table 1: Identification of Risk Factors

Type	Variable	Labels
INPUT	CD4 count	High, Low
	Viral Load	High, Low
	Opportunistic Infection	Yes, No
	Nutritional Status	High, Low
OUTPUT	HIV/AIDS Survival	Yes, No

In order to develop the predictive model for the development of the predictive model for HIV/AIDS survival, there was the need to gather historical dataset containing the values of the four risk factors alongside their respective risk of HIV/AIDS survival from hospitals where pediatrics were receiving treatment and follow-up. Thus, the Federal Medical Centre, Owo, Ondo State and the Obafemi Awolowo University Teaching Hospitals Complex (OAUTHC), Ile-Ife, Osun State both in south-western Nigeria were considered as the location from which the historical dataset was gathered. Thus, case-files of pediatric HIV/AIDS patients receiving follow-up was used to gather the needed information which was converted to electronic copy in comma separated variable (.csv) format. This dataset had to be later converted to the acceptable format needed by the simulation environment to perform the training and testing procedure required for the predictive model development process.

a. Model formulation

Following the identification of the risk factors necessary for predicting the survival of HIV/AIDS in pediatric Nigerian patients and the collection of historical data explaining the relationship between the identified risk factors and the HIV/AIDS survival, the predictive model for the risk of depression was formulated using a supervised machine learning algorithms called the support vector machines (SVM) using the sequential minimal optimization (SMO) algorithm. For any supervised machine learning algorithm proposed for the formulation of a predictive model, a mapping function can be used to easily express the general expression for the formulation of the predictive model for HIV/AIDS survival—this is as a result that most machine learning algorithms are black-box models which use evaluators and not power series/polynomial equations.

The historical dataset S which consists of the records of individuals containing fields representing the set of risk factors (i number of input variables for j individuals), X_{ij} alongside the respective target variable (HIV/AIDS survival) represented by the variable Y_j – the HIV/AIDS survival for the j th individual in the j records of data collected from the hospital selected for the study. Equation 1 shows the mapping function that describes the relationship between the risk

factors and the target class – HIV/AIDS survival. The mapping φ was used to represent the predictive model for mapping the risk factors of each pediatric patient to their respective HIV/AIDS survival according to equation 2.

$$\varphi: X \rightarrow Y; \quad (1)$$

defined as: $\varphi(X) = Y$

$$\varphi(X) = \begin{cases} YES \\ NO \end{cases} \quad (2)$$

The following paragraph discusses the supervised machine learning algorithm used in formulating the predictive model for the survival of HIV/AIDS in pediatric patients in Nigerian hospitals.

b. Support Vector Machines

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples for HIV/AIDS survival, each marked for belonging to one of two categories (HIV/AIDS survival), an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on (Figure 2). In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

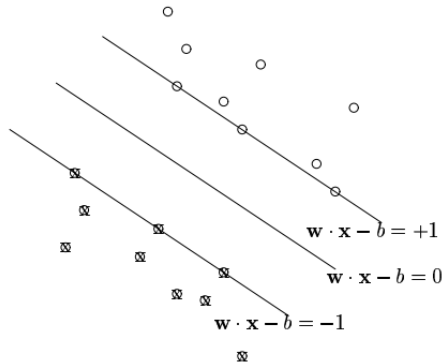


Figure 2: A linearly separable hyperplane using SVM

Assuming the dataset used in the study containing N training datasets $\{(x_1, y_1), (x_2, y_2) \dots (x_N, y_N)\}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$. The motivation is to learn a linear separating hyperplane classifier:

$$f(x) = \text{sgn} \langle w \cdot x - b \rangle \quad (3)$$

Furthermore, there is the need for the hyperplane to have maximum separating margin with respect to the two classes. Specifically, there is a need to find the hyperplane $H: y = w \cdot x - b = 0$ and two hyperplanes, H_i for $i = 1, 2$ parallel to it and with equal distance to it (see Figure 2) defined as in equation (4) while the problem can be formulated as in equation (5).

$$y_i(w \cdot x_i - b) \geq 1 \text{ for } y_i = \pm 1 \quad (4)$$

$$\min_{w,b} \frac{1}{2} w^T w \text{ subject to } y_i(w \cdot x_i - b) \geq 1 \quad (5)$$

This equation is a convex, quadratic programming problem (in w, b), in a convex set. By introducing Lagrange multipliers $\alpha_1, \alpha_2, \dots, \alpha_N \geq 0$, we have the following Lagrangian:

$$\mathcal{L}(w, b, \alpha) \equiv \frac{1}{2} w^T w - \sum_{i=1}^N \alpha_i y_i (w \cdot x_i - b) + \sum_{i=1}^N \alpha_i \quad (6)$$

Sequential minimal optimization (SMO) is an algorithm for solving the quadratic programming (QP) problem that arose during the training of support vector machines. It was invented by John Platt in 1998 at Microsoft Research. SMO is widely used for training support vector machines and is implemented by the popular LIBSVM tool. For this study, the SMO algorithm was implemented using the classifier.functions.SMO class available in the WEKA simulation environment.

c. Model Simulation Environment

The Waikato Environment for Knowledge Analysis (WEKA[®]) software – a suite of machine learning algorithms was used as the simulation environment for the development of the predictive model. The software is freely available at <http://www.cs.waikato.ac.nz/ml/weka>. The system was written using object-oriented language, Java. There are several different levels at which Weka can be used. Weka provides implementations of state-of-the-art data mining and machine learning algorithms. Weka contains modules for data preprocessing, classification, clustering and association rule extraction for market basket analysis. The main features of Weka include:

- 49 data preprocessing tools;
- 76 classification/regression algorithms;
- 8 clustering algorithms;
- 15 attribute/subset evaluators + 10 search algorithms for feature selection;
- 3 algorithms for finding association rules; and
- 3 graphical user interfaces, namely:
 - The Explorer for exploratory data analysis;
 - The Experimenter for experimental environment; and
 - The Knowledge Flow, a new process model inspired interface.

The datasets were subjected to 10-fold cross validation using the SMO algorithm for SVM which was selected for this study. Before subjecting the historical datasets containing the values of the risk factors alongside the survival of HIV/AIDS in the original dataset; there was the need of storing the dataset according to the format for data representation needed for data mining tasks on the Weka environment. The default file type is called the attribute relation file format (.arff) although the initial dataset was collected and stored using the csv format. The dataset collected was divided into two parts: training and testing data – the training data was used to

formulate the model while the test data was used to validate the model. The arff format required the identification of three (3) parts of the dataset, namely:

- a. **The relation section:** was used to identify the name of the file identified which in this case is AIDS-all for the data containing all patients' records. The relations tag is identified using the name @relation before the relation name;
- b. **The attribute section:** was used to identify the fields/attributes (risk factors) identified as the input variables for the HIV/AIDS survival class where the last attributes describes the target class. The attribute tag for each attribute is identified using the name @attribute before each attribute name; and
- c. **The data section:** was used to identify the dataset values for each respondents collected in the same order as the attributes were listed. Each respondent's record of data is represented as the set of values on each line with the HIV/AIDS survival shown on the last portion of each line. The data containing the values of the attributes for each respondent is listed on the line following the name tag identified as @data.

d. Model Validation

For this classification problem, it was natural to measure a classifier's performance in terms of the error rate. The classifier predicted the class of each instance – the pregnant women's record containing values for each risk of depression: if it is correct, that is counted as a success; if not, it is an error. The error rate being the proportion of errors made over a whole set of instances, and thus measured the overall performance of the classifier. The error rate on the training data set was not likely to be a good indicator of future performance; because the two classifiers were learned from the very same training data. In order to predict the performance of a classifier on new data, there was the need to assess the error rate of the predictive model on a dataset that played no part in the formation of the classifier. This independent dataset was called the test dataset – which was a representative sample of the underlying problem as was the training data.

The process of leaving a part of a whole dataset as testing data while the rest is used for training the model is called the holdout method. The challenge here is the need to be able to find a good classifier by using as much of the whole historical data as possible for training; to obtain a good error estimate and use as much as possible for model testing. It is a common trend to holdout one-third of the whole historical dataset for testing and the remaining two-thirds for training. For this study the cross-validation procedure was employed, which involved dividing the whole datasets into a number of folds (or partitions) of the data. Each partition was selected for testing with the remaining $k - 1$ partitions used for training; the next partition was used for testing with the remaining $k - 1$ partitions (including the first partition used or testing) used for training until all k partitions had been selected for testing. The error rate recorded from each process was added up with the mean the mean error rate recorded. The process used in this study was the stratified 10-fold cross validation method which involves splitting the whole dataset into ten partitions.

Figure 3 shows a representation of the 10-fold cross validation process.

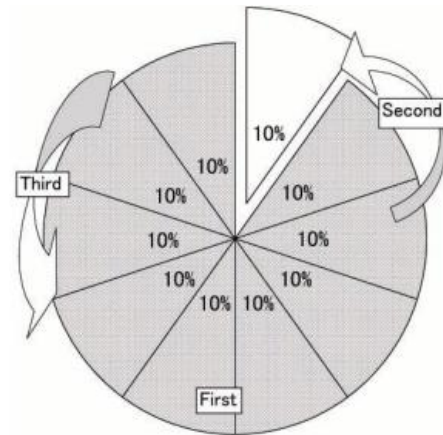


Figure 3: 10-fold cross validation process

During the course of evaluating the predictive model, a number of metrics were used to quantify the model's performance. In order to determine these metrics, four parameters must be identified from the results of predictions made by the classifier during model testing. These are: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). In this study which involves a binary classification, either of survived and not survived can be considered as positive while the other negative.

True positives are the correct prediction of positive cases, true negatives are the correct prediction of negative cases, false positives are the negative cases predicted as positives while false negatives are positive cases predicted as negatives. These results are presented on confusion matrix (Figure 4).

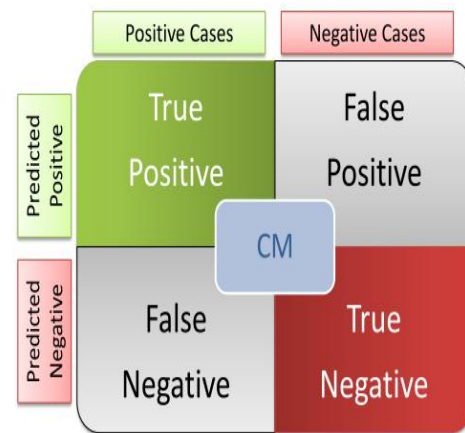


Figure 4: Confusion matrix for model performance

The four parameters were used to formulate the metrics discussed as follows. The performance metrics are thus defined as follows:

- **Sensitivity/True positive rate/Recall:** is the proportion of actual positive cases that were correctly predicted positive by the model.

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

- **False Positive rate/False alarm:** is the proportion of actual negative cases that were predicted as positive by the model.

$$FP = \frac{\text{False alarm}}{FP + TN} \quad (8)$$

- **Precision:** is the proportion of the predicted positive/negative cases that were actually positive or negative. Equations (9) and (10) show the precision for positive and negative cases.

$$\text{Precision (positive class)} = \frac{TP}{TP + FP} \quad (9)$$

$$\text{Precision (negative class)} = \frac{TN}{FN + TN} \quad (10)$$

- **Area under the Receiver Operating Characteristics (ROC) curve:** 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 for HIV/AIDS survival.

- **Accuracy:** is the total number of correct classifications (positive and negative)

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP} \quad (11)$$

4. RESULTS AND DISCUSSION OF PREDICTIVE MODEL FOR HIV/AIDS SURVIVAL

For the purpose of this study, data was collected from 215 (136 from FMC Owo and 79 from OAUTHC) pediatric patients using the information elicited from their case files during the visit to the study locations selected for this study. This was done in order to use the dataset of the 215 pediatric patients to formulate (train and test) the predictive model for HIV/AIDS survival. Figure 5 shows a screenshot of the data collected from the 215 respondents selected for this study.

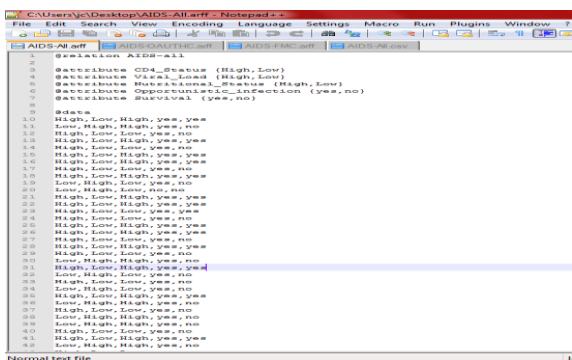


Figure 5: The .arff file for historical dataset

The data was stored in the attribute relation file format (.arff) which is the acceptable format for the data mining simulation environment selected for this study. The format required the identification of three (3) parts of the dataset, namely:

- The relation section:** was used to identify the name of the file identified which in this case is AIDS-all for the data containing all 215 patients, this data was selected for training and testing the model using the 10-fold cross validation for the SVM-SMO algorithm. The relations tag is identified using the name @relation before the relation name;
- The attribute section:** was used to identify the fields/attributes (risk factors) identified as the input variables for HIV/AIDS survival where the last attributes described HIV/AIDS survival status (Yes or No). There are 5 attributes identified in the file with the first 4 identifying the input variables (risk factors of HIV/AIDS survival) while the last variable is the HIV/AIDS survival class. Each attribute has its own respective label which shows the possible values that were stated for each attribute defined in the dataset. The attribute tag for each attribute is identified using the name @attribute before each attribute name; and
- The data section:** was used to identify the dataset values for each respondents collected in the same order as the attributes were listed. Each respondent's record of data is represented as the set of values on each line with the value of the HIV/AIDS survival class shown on the last portion of each line. The data containing the values of the attributes for each respondent is listed on the line following the name tag identified as data.

Table 2 shows the description of the data collected from the patients alongside the percentage distribution of the samples drawn for the values of each attribute.

Table 2: Distribution of Risk factors and HIV/AIDS survival

		FMC Owo	(%)	OAUTHC	(%)	Total	(%)
CD4 Count	High	93	43.26	59	27.44	152	70.70
	Low	43	20.00	20	9.30	63	29.30
Viral Load	High	43	20.00	20	9.30	63	29.30
	Low	93	43.26	59	27.44	152	70.70
Nutritional Status	High	62	28.84	31	14.42	93	43.26
	Low	74	34.42	48	22.33	122	56.74
Opportunistic Infection	Yes	101	46.98	54	25.12	155	72.09
	No	35	16.28	25	11.63	60	27.91
Survival of HIV/AIDS	Yes	59	27.44	40	18.60	98	45.58
	No	77	35.81	39	18.14	117	54.42

From the results of the data collection for the risk factors and their respective risks, it was discovered that the CD4 count was distributed in such a way that 70.7% were High while 29.3% were low; the viral load was distributed such that 29.3% were High while 70.7% were Low; the nutritional status had a distribution of 43.26% that were High while 56.74 were Low and the opportunistic infection showed distribution of 72.1% for Yes and 27.9% for No. Thus a

higher percentage of the pediatric patients had high CD4 count although with that same proportion having Low viral load. Also, 72.1% of the pediatric patients had opportunistic infection with 27.9% having no opportunistic infection and the nutritional status was distributed such that 43.3% was High while 56.7% were Low. Figure 6 gives a visual description of the dataset containing the risk factors collected from the pediatric patients considered for this study.

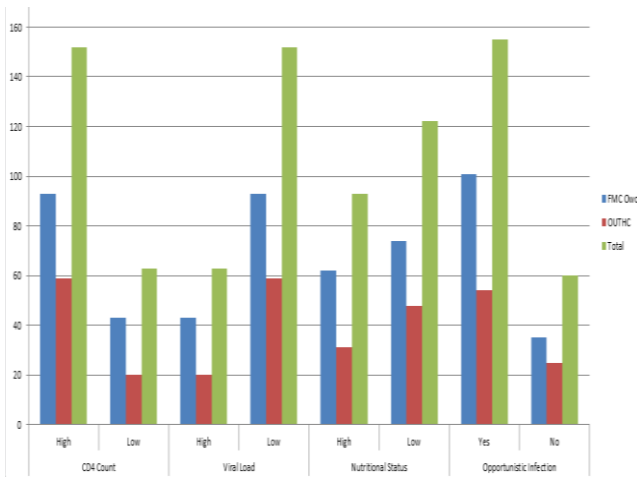


Figure 6: Visualization of risk factors for HIV/AIDS survival

Furthermore, table 2 showed the distribution of the survival of HIV/AIDS among the 215 pediatric patients considered for this study from the two selected locations. From the dataset collected from observing the records in the patients' case files, it was discovered that in the overall dataset with 215 records there were more patients who didn't survive than there were those who survived. 54.4% of the patients did not survive while 45.6% survived HIV/AIDS although there were more patients that did not survive in the dataset collected from FMC Owo but an equal distribution between those that survived and didn't survive in the dataset were collected from OAUTHC. Figure 7 shows a graphical plot of the distribution of the HIV/AIDS survival class among the pediatric patients selected for this study.

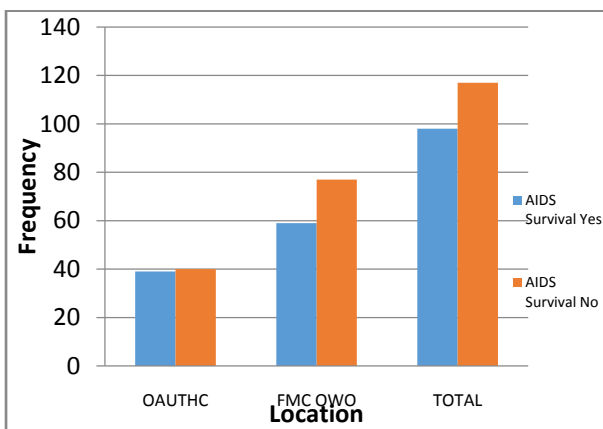


Figure 7: Distribution of HIV/AIDS survival

Following the conversion of the dataset into the preferred .arff file format of the simulation environment and the visualization and summarization of the dataset using

frequency tables and graphs, the dataset was loaded into the WEKA environment for the purpose of the classification tasks that was required for the development of the predictive model for HIV/AIDS survival in pediatric patients. The 10-fold cross validation method was used as the training and testing method during the simulation of the predictive model using the historical dataset selected for this study.

Figure 7 shows the WEKA simulation interface where the classification task of the development of the predictive model for HIV/AIDS survival was performed using the sequential minimal optimization (SMO) algorithm implemented by the support vector machine (SVM). The WEKA class that was used to implement the SVM-SMO algorithm was the class `weka.classifiers.functions.SMO` – an implementation of John Platt's SMO algorithm for training a SVM classifier. The polynomial kernel was the kernel used for estimating the Lagrangian of the problem which is usually expressed as in equation (12) as follows:

$$K(x, y) = \langle x, y \rangle^p \quad \text{of } k(x, y) = (\langle x, y \rangle + 1)^p \quad (12)$$

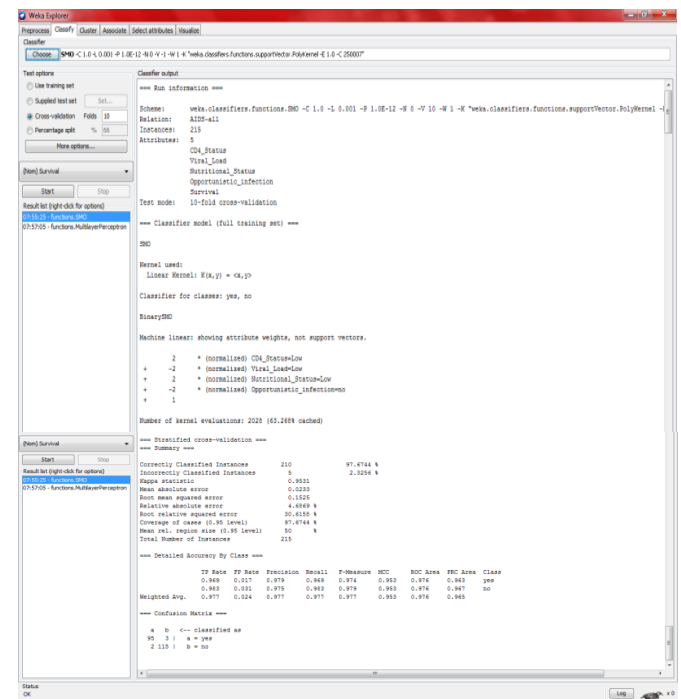


Figure 7: WEKA interface for the simulation of the SMO-SVM algorithm

Also, the batch size which defined the number of preferred number of instances during the batch prediction performed used by the SVM-SMO was 100, and the filter type used in transforming the data was by normalization the training data. The random seed used was 1 while the tolerance parameter used was 0.001. Figure 8 shows a description of the settings for the SVM-SMO algorithm used in implementing the algorithm on the WEKA environment.

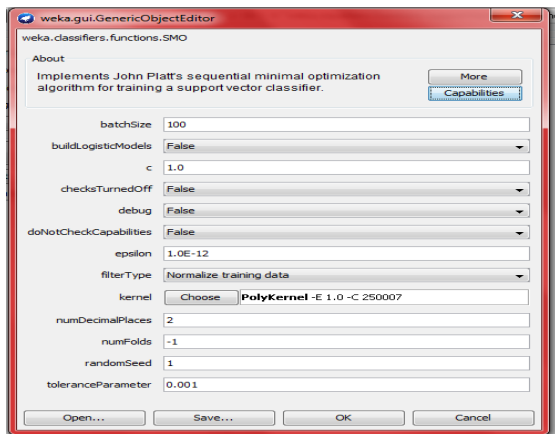


Figure 8: WEKA interface for SVM-SMO implementation

Following the implementation of the SVM-SMO algorithm on the training dataset sample used for this study using the 10-fold cross validation, the results of the predictions made by the model on the 215 datasets used were plotted onto a confusion matrix in order to plot the true and false positives/negatives. Figure 9 shows the confusion matrix of the results of the implementation of the SVM-SMO algorithm on the dataset used in this study. The predicted values are summed up along the vertical for both Yes and No cases while the actual values are summed up along the horizontal for both Yes and No cases in the dataset.

	A	B	< - - Predicted as
Yes = A	95	3	
No = B	2	115	

Figure 9: Confusion matrix of SVM-SMO algorithm

The results of the implementation of the SVM-SMO algorithm for the development of the predictive model for HIV/AIDS survival in Nigerian pediatric patients based on the historical dataset collected from the study locations selected for this study. Out of the 98 patients that survived HIV/AIDS disease, 95 were correctly predicted while 3 were misclassified as No and out of the actual 117 patients that did not survive HIV/AIDS the SVM-SMO predicted 115 correctly but misclassified 2 as Yes. Hence, the total correct classifications made by the SVM-SMO algorithm was 210 out of the 215 records in the historical dataset used to train and test the predictive model owing for a predictive accuracy of 97.7%.

The result of the true and false positives/negatives recorded in the confusion matrix was used to evaluate the performance evaluation metrics of the predictive model developed for HIV/AIDS survival (see Table3). The result presented the true positive rate (TP) (recall/sensitivity) values of the Yes and No cases as 0.969 and 0.983 respectively which implies that the model was able to predict 96.9% and 98.3% of the Yes and No cases in the actual dataset respectively owing for an average of 0.977. The false positive (FP) rate (false alarm)

values for the Yes and No cases were recorded as 0.017 and 0.031 respectively owing for an average value of 0.024 – about 2.4% of a Yes/No cases was misclassified a No/Yes respectively. The precision which gave a description of the proportion of predicted cases that was correct showed values of 0.979 and 0.975 for the Yes and No cases respectively with an average of 0.977 owing for a value of 97.7% of the predicted Yes/No cases to be correct classifications. Finally, the area under the receiver operating characteristics (ROC) curve which is a plot of the TP rate against the FP rate in order to identify how well the classifier was able to discriminate between the Yes and No cases showed values of 0.976 for the Yes/No cases.

Table 3: Performance Evaluation of SVM-SMO algorithm

	Yes	No	Average
TP rate/recall	0.969	0.983	0.977
FP rate/false alarm	0.017	0.031	0.024
Precision	0.979	0.975	0.977
Area under ROC	0.976	0.976	0.976
Accuracy (%)	97.6744		
Correct Classification	210		
Incorrect Classification	5		

5. CONCLUSION

This study focused on the development of a predictive model for HIV/AIDS survival in pediatric patients using the risk factors in order to classify the survival of each patient's individual record. Following the collection of the historical dataset on the distribution of the HIV/AIDS survival among the pediatric patients selected for this study by observing the information stored in their respective case-files the model was formulated using the sequential minimal optimization (SMO) algorithm of the support vector machines (SVM) – a binary classification algorithm. The simulation of the implementation of the SVM-SMO algorithm was done using the WEKA simulation environment via the 10-fold cross validation training method.

The predictive model developed for the prediction of the survival of pediatric HIV/AIDS patients in south-western Nigeria showed good results with the accuracy of the algorithm as high as 97.7%. The predictive model is expected to give very promising results when used on other pediatric patients in order to determine their respective survival of HIV/AIDS. Also, it was discovered that the SVM-SMO algorithm was able to infer as much information from the historical dataset used for this study about the relationship between the risk factors and the survival of HIV/AIDS in Nigerian pediatric patients.

The model can also be integrated into existing Health Information System (HIS) which captures and manages clinical information which can be fed to the HIV/AIDS survival classification model thus improving the clinical decisions affecting HIV/AIDS survival and the real-time assessment of clinical information. It is advised that a continual assessment of other variables that have a relationship with HIV/AIDS survival be made in order to increase the number of information relevant to creating an



improved prediction model for HIV/AIDS survival classification using the proposed model in this study.

6. REFERENCES

- [1] Eric, J. and Daria, J. , 2012. HIV-1 Antiretroviral Drug Therapy. *Journal of Cold Spring Harb Perspect Medical*. Vol. 2, No 4. pp 23-45
- [2] Henry, K., 2013. The Global HIV/AIDS Epidemic. Available from www.kff.org/global-health-policy/fact.../the-global-hiv-aids-epidemic/. [Accessed 12 January, 2014].
- [3] Hoa, M., 2011. ART Adherence among People Living With HIV/Aids in North Vietnam, Queensland University of Technology, Brisbane Australia
- [4] Idowu, P., et al., 2009. Spatial Predictive Model for Malaria in Nigeria, *Journal of Health Informatics in Developing Countries*, Vol. 3, No 2. pp 31-36.
- [5] Idowu, P., 2012. Development of A Web-Based Geo-Spatial Environmental HealthTracking System for South Western Nigeria. Unpublished PhD Thesis Submitted To Department Of Computer Science and Engineering, Obafemi Awolowo University, Ile- Ife, Nigeria.
- [6] 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
- [7] 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].
- [8] Ojunga, N., et al., 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.
- [9] Picat, M., et al ., 2013. Predicting Patterns of Long-Term CD4 Reconstitution in HIV-Infected Children Starting Antiretroviral Therapy in Sub-Saharan Africa: A Cohort-Based Modelling Study. *Journal of Pediatric Medicine*.Vol. 10, No 10. pp 45-49
- [10] Shearer, W. , 2000. Evaluation of Immune Survival Factors in Pediatric HIV-1 Infection. *Annual National Academic Journal*.Vol. 91, No. 8. pp 298-312.
- [11] UNAIDS, 2012. Together We Will End AIDS. Available from www.unaids.org. [Accessed 12 June, 2013] UNAIDS, 2013. Global Report on HIV. Available from http://www.unaids.org/en/resources/documents/2013/name_85053_en.asp. [Accessed 2nd August, 2014]
- [12] WHO, 2012. Towards Universal access: Scaling up Priority HIV/AIDS interventions in the HealthSector. Progress Report 2010. Available from <http://whqlibdoc.who.int> [Accessed January 15, 2014]