Journal of Research in Science, Technology, Engineering and Management (JoRSTEM)
Vol. 2, issue : 2, June 2016: ISSN(online): 2456-0197

Prediction of Pediatrics HIV/Aids Patient's Survival in Nigeria: A Data Mining Approach

Adebayo P. Idowu 1, Alademoko T.A[2] and Agbelusi Olutola[3]

[1] *Department of Computer Science & Engineering,*
Obafemi Awolowo University, Ile-Ife, Nigeria.
paidowu@oauife.edu.ng

[2] *Department of Peadiatric and Child Health,*
Obafemi Awolowo University
Ile-Ife, Nigeria,
Department of Computer Science, Rufus Giwa Polytechnic, Owo, Nigeria .
tola52001@yahoo.com

Abstract— ***Data mining methods which combine techniques from database research, statistics and artificial intelligence have been applied to many research disciplines including engineering, finance and medicine. This study identified survival variables for HIV/AIDS paediatric patients, developed predictive model for determining the survival of the patients who were receiving antiretroviral drug in the Southwestern Nigeria based on identified variables, compared and validate the developed model. Interviews were conducted with the virologists and Pediatricians at two health institutions from the study area in order to identify survival variables for HIV/AIDS Paediatric patients. Paediatric HIV/AIDS patients' data (216) were also collected from two health institutions, preprocessed and the 10-fold cross validation technique was used to partition the datasets into training and testing data. Predictive model was developed using supervised learning technique (The Multi-layer Perception (MLP)) and the Waikato Environment for Knowledge Analysis (WEKA) was used to simulate the models in which CD4 count, Viral Load, Opportunistic infections and Nutritional status were used as the independent variables for the prediction. The result showed that The Multi-layer Perception (MLP)) was suitable for carrying out the task of forecasting the survival of Paediatric HIV/AIDS patients with an accuracy of 99.07% (214 correct classifications out of 216), the mean absolute error rate was 0.022, 0.0962 for the root mean square error and 4.48% for the relative absolute error. The ROC area for the model was also 0.992 showing that the level of bias was also very low (0.008)99.07%.The validation was done by comparing the developed model with the historical data from the two selected health institutions in Nigeria..***

Keywords: - Virologist, Antiretroviral Therapy, Paediatric, Survival, Acquired Immunodeficiency Disease Syndrome, Human Immunodeficiency Virus, Antiretroviral drug, Virologist

I. INTRODUCTION

In the last three decades, children with HIV/AIDS suffer the usual childhood infections more frequently and more severely than uninfected children. HIV/AIDS is a type of virus called a retrovirus which infects human when it comes in contact with tissues such as those that line the vagina, anal area, mouth, eyes or through a break in the skin (Eric *et al.*, 2012), while Acquired Immunodeficiency Syndrome (AIDS) is the advanced stage of the retroviral infection that swept through sub-Saharan Africa with venom (Hoa, 2011; Joyce, 2014; Idowu, 2012). This infection can cause seizures, fever, pneumonia, recurrent colds, diarrhea, dehydration and other problems that often result in extended hospital stay (UNAID, 2008). HIV/AIDS continues to be a very serious health issue facing the world in which 34 million people were living with HIV at the end of 2011 and an estimated 0.8% of adults aged 15-49 years worldwide are living with the virus, although the burden of the epidemic continues to vary considerably between countries and regions (Ojunga *et al*., 2014; UNAIDS, 2012).

In Sub-Saharan Africa, roughly 25 million people were living with HIV in 2012, accounting for nearly 70 percent of the global total of infected patients. The epidemic has both social and economic consequences, not only in the health sector but also in education, industry and the wider economy (WHO, 2012; Shearer, 2000).

Moreover, Sub-Saharan Africa remains most severely affected with nearly 1 in every 20 adults (4.9%) living with HIV which accounts for 69% of the people living with HIV worldwide at present.

There is no cure for HIV but it is being managed with antiretroviral drugs (ARV) and Highly Active Antiretroviral drugs (HAART) which is the optimal combination of ARV (Rosma *et al*., 2012; Kama and Prem, 2013). ARV does not kill the virus but slow down the growth of the virus (Ojunga *et al*., 2014; WHO, 2012a). Antiretroviral therapy (ART) and highly antiretroviral therapy (HAART) are the mechanisms for treating retroviral infections with drugs.

The epidemic of HIV/AIDS affects two classes of people: the paediatric and the non-paediatric individuals. The non-Paediatric patients are patients above 15 years of age while the paediatric patients who form the main target of this research are patients whose age is less than 15 years (Sanjel, 2009).

There are four distinct stages of HIV infection which includes: the primary HIV infection stage or clinical stage 1 which involves asymptomatic and acute retroviral syndrome ; clinically asymptomatic stage or clinical stage 2 which involves moderate and unexplained weight loss (<10% of presumed or expected body weight; symptomatic stage or clinical stage 3 is also a stage where a presumptive diagnosis could be made on the basis of clinical signs or simple investigations like unexplained chronic diarrhea for longer than one month, unexplained persistent fever (intermittent or constant for longer than one month) and progression or Clinical stage 4 which is a condition where a presumptive diagnosis can be made on the basis of clinical signs or simple investigations like HIV wasting syndrome, pneumocystis pneumonia, recurrent severe or radiological bacterial pneumonia etc. Patient at this stage is in a condition where confirmatory diagnostic testing is necessary (Avert, 2014a; WHO, 2012b; Centre for Disease Control, 2011).

There are different modes of transmission of this virus, one of which is mother-to-child transmission. About nine out of ten children exposed are infected with HIV during pregnancy, labour, delivery or while breastfeeding (UNAIDS, 2010). Without treatment, 15-30 percent of babies born to HIV positive women are infected with the virus during pregnancy and delivery and a further 5-20 percent are also infected through breastfeeding (WHO, 2006a). In high-income countries, preventive measures are undertaken to ensure that the transmission of HIV from Mother to Child is relatively rare and in cases where it occurs, a range of treatment options are undertaken so that the child can survive into adulthood. Blood transfusion is another route in which HIV infection can occur in medical setting (Mira and Dennis, 2010)**.**

In order to predict the survival of paediatric HIV/AIDS patients that are receiving ARV in South western Nigeria, data mining technique was introduced to this research. Data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends.

II. RELATED WORK

Data mining is an important tool in disease prediction. The literature selected and discussed in this section are those that are more related and relevant to the survival of Paediatric HIV/AIDS patients. The various mathematical, statistical and data mining approaches have been used in different research for the prediction of survival in HIV/AIDS infected patients. Basically, Cox regression and data mining approach are mostly used for survival prediction of HIV/AIDS.

Lucia (1996) developed a model to predict survival of HIV/AIDS using sequential and standard neural networks. The aim of the study was to produce a model for disease progression in AIDS using sequential neural network and compare the model's accuracy with that of a model constructed using only standard neural networks based on demographic and socioeconomic variables. The strength of the study was that sequential neural networks could discriminate patients who die and patients who survive more accurately than the standard neural networks. The weakness of the study is that only demographic and socioeconomic variables were used, which is not sufficient to predict survival. CD4 count, viral load, opportunistic infections and nutritional status would have been enough to predict survival accurately in this study.

Brain *et al* (2006) applied neural network in the prediction of HIV status of an individual based on demographic and socioeconomic characteristics. The aim of the study was to classify the HIV status of an individual given certain demographic factors. The strength of the study is that the neural networks used for prediction has high predictive capability. The weakness is that demographic and socioeconomic characteristics are not enough to accurately predict survival status.

Rosma *et al* (2012) developed a predictive model for AIDS Survival using Data Mining Approach. The aim of the research was to describe the feasibility of applying data mining technique to predict the survival of HIV/AIDS. An adaptive fuzzy regression technique, FuReA, was used to predict the length of survival of AIDS patients based on their CD4, CD8 and viral load counts. The strength of the research is that CD4, CD8 and viral load counts were used because the authors believed that predictors / markers are appropriate for predicting AIDS survival due to the high accuracies demonstrated by Fuzzy regression analysis (FUREA). The weakness is that fuzzy neural network

prediction results on AIDS survival could not be made possible because of data limitation. This is because Fuzzy neural network requires the use of large volume of data for prediction. Opportunistic infections and Nutritional status are important predictor variables together with CD4, CD8 and viral load counts that can be used to predict the survival of HIV/AIDS patients.

Moore *et al* (2006) carried out a study on CD4 % as an independent predictor of survival in patients starting antiretroviral therapy with absolute CD4 cell counts between 200 and 350cells/Micro. The aim of this study was to carry out a cumulative mortality rate of patients starting antiretroviral therapy with absolute CD4 cell counts between 200 and 350 cells/mm^3 using Kaplan-Meier methods. Cox proportional hazards regression was also used to model the effect of baseline CD4 strata, CD4 percentage strata and other prognostic variables on survival. The strength of this study shows that Low CD4 percentage, poor adherence to therapy and high viral load were associated with increased risk of death in HIV/AIDS infected patient. The limitation is that only CD4 count was used as predictor variable and there was no availability of enough data in the study area. In this research, CD4, viral load, opportunistic infections and nutritional status were used to predict survival of paediatric HIV/AIDS patient.

Sameem *et al* (2010) used Classification and Regression Tree (CART) for the Prediction of Survival of Aids Patients receiving antiretroviral therapy in Malaysia. The aim is to investigate the use of CART as a tool for prediction of AIDS survival using CD4, CD8, Viral Load and Weight as predictor variables. The strength of the research is that the potential treatment methods and monitoring the progress of treatment of AIDS patients could be determined with the approach experimented and the results obtained. Fewer variables were considered for the prediction of survival of AIDS patients and data limitation is also a constraint in the study. Also, opportunistic infections and nutritional status that are very important for HIV/AIDS prediction were not used in the research. CD4, CD8, viral load and weight are not enough factors to predict survival.

Petro *et al* (2010) worked on high survival and treatment success sustained after two and three years of first-line ART for children in Cambodia. The aim of this study was to assess two-and three-year survival, CD4 evolution and virological response among children on ART in a programmatic setting in Cambodia. Kaplan-Meier analysis was used to estimate survival and Cox regression was used to identify risk factors associated with treatment failure. The strength shows that good survival, immunological restoration and viral suppression could be sustained after two to three years of ART among children in resource-constrained settings. The weakness is that only CD4 count was used as predictor variable in this study. CD4, viral load, opportunistic infection and nutritional status were variables used in this study as predictive factors.

Ojunga *et al* (2014) applied logistic regression in modelling of survival chances of HIV-positive patients on 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 on highly active antiretroviral therapy (HAART). The study was expected to provide suitable model for predicting the chances of survival among the HIV positive patients attending ART clinic in Nyakachi District and also provide information for policy makers on the factors affecting survival of HIV positive individual on ARV drugs. 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, viral load, opportunistic infections and nutritional status were added to the existing study in this thesis as predictive factors.

Dalton *et al* (2010) carried out a prospective cohort on the Predictors of mortality in HIV-1 infected children on antiretroviral therapy in Kenya. The aim of this work was to carry out a study on early mortality following highly active antiretroviral therapy (HAART) using Cox proportional hazard model to determine the baseline characteristics associated with mortality and Kaplan-Meier method to estimate the probability of survival. The study shows that low baseline hemoglobin was an independent risk factor for death. The weakness is that only hemoglobin was used as predictor variable in this research which is not enough to determine survival rate.

Heamoglobin is not enough to predict the survival of paediatric infected HIV patients but in this study, CD4, and viral load, opportunistic infection and nutritional status were used as predictive factors.

III. METHODLOGY

Extensive review of literature on related areas in HIV/AIDS survival prediction was carried out and interview was conducted with virologists and Paediatrician in order to identify required survival variables for HIV/AIDS. Predictive model was developed using supervised learning technique (Multilayer perception classifier) and the Waikato Environment for Knowledge Analysis (WEKA) was used to simulate the model. The result of this prediction shows that The Multi-layer Perception (MLP)) technique was suitable in carrying out the task of forecasting the survival of Paediatric HIV/AIDS patients with an accuracy of 99.07% (214 correct classifications

out of 216) based on selected dependent variables. The validation was done by comparing the developed model with the historical data collected from the two selected health institutions in Nigeria.

IV. RESULTS AND DISCUSSIONS

The mathematical model for the artificial neural network in figure 1 is as follows

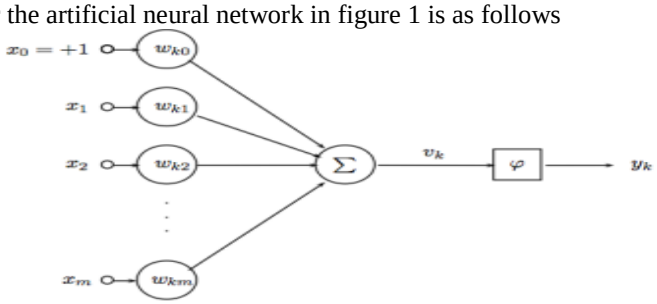

Figure 1: Signal-flow of a Multilayer Perception Model

- **THE INPUT LAYER**– at this part of the multi-layer perception (MLP) the input values are entered into the MLP system and the weights, w_i of each input, x_i are applied after which the summation, U_k is sent to the hidden layer for the activation function, to take effect.
If X = (x_1 = "CD4 status = value", x_2 = "viral load = value", x_3 = "nutritional status = value", x_4 = "opportunistic infection = value")
- **THE HIDDEN LAYERS**– at this part of the MLP the summation of the input variables are all sent to the activation function which is fired through all the hidden layers (for the purpose of this research 20 layers were used) and finally producing the required output variable, y_k which determines whether the survival is YES/NO.

$$y_k = \varphi(v_k) = \varphi(u_k + b_k), \quad y_k = \pi r^2 \tag{1.1}$$

$$\varphi(w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + b_k \tag{1.2}$$

Where the activation function is:

$$\varphi(v) = \frac{1}{1+e^{-av}} \tag{1.3}$$

- **THE OUTPUT LAYER**– the value of the output (Survival) is determined with the error rate as low as possible. Also, the back-propagation algorithm is applied which tries to reduce the error rate, of the model via gradient descent. At iteration n (the nth row in the training set), the error for neurons in the output layer is calculated as:

$$e_j(n) = d_j(n) - y_j(n) \tag{1.4}$$

$$v_j = \sum_{i=0}^{m} w_{ji}(n) y_i(n) \tag{1.5}$$

$$y_j = \varphi_j(v_j(n)) \tag{1.6}$$

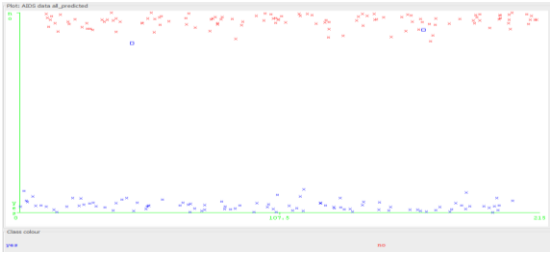

Figure 2: Multi-layer Perception Predictions of all 216 Data

Classified as	YES	NO
YES	95	2
NO	0	119

Figure 3: Confusion Matrix of M P Classification for all 216 Data

Table 1: Summary of Results for M P Prediction Model

Item	FMC, Owo (137)	OAUTHC/ Wesley Guild (79)	Both Data Sources (216)
Accuracy	99.2704%	97.4684%	99.0741%
Correct Classification	136	77	214
Incorrect Classification	1	2	2
TP rate	0.983	0.976	0.979
FP rate	0.00	0.026	0.00
Precision	1.00	0.976	1.00
Area under ROC	0.993	0.981	0.992
MAE	0.0193	0.0428	0.022
RMSE	0.0862	0.1593	0.0962
RAE	3.9325%	8.5184%	4.477%

A number of observations were made on the dataset using Multilayer perception classification in the developing the predictive model for peadiatric HIVAIDS patients. Figures 2, 3 and 4 show the results of the multi-layer perception prediction model for the data collected from FMC, Owo in Ondo State and OAUTHC in Osun State respectively. In figure 2, the blue crosses identify a Survival (YES) and red crosses identify a no survival (NO) while the boxes show misclassifications (i.e. YES classified as NO and vice versa).

Out of 97 dataset which had survival of YES, 95 dataset that were actually yes were classified as YES as shown by red crosses and 2 that were yes were mis- classified as NO and this is represented with two blue boxes. Furthermore, out of 119 dataset which had a survival of NO, 0 were wrongly classified as YES as represented with blue crosses and 119 which were actually no were classified as NO and represented with blue boxes.

Figure 3 is the confusion matrix which was used to evaluate the performance of Multi layer Perception classification for 216 data. From the confusion matrix table, out of 97 dataset which had survival of YES, 95 which was actually yes were also classified as YES and this shown in the first box on the table and 2 which was actually yes were mis-classified as NO in the second box. Secondly, out of 119 dataset which had a survival of NO, 0 were actually no but mis-classified as YES and 119 dataset that were actually no were truly classified as NO.

Table (1) shows that Multilayer perception Prediction model has accuracy of 99.0741%, 214 data were correctly classified and 2 incorrectly classified. The true positive rate (TP) , false positive rate (FP), precision, area under ROC, mean absolute error (MAE), root mean standard error (RMSE) and root absolute error (RAE) values were 0.979, 0.00,1.00, 0.992, 0.022, 0.0962 and 4.477% respectively.

Validation of the Model : Validation is the task of demonstrating that the model is a reasonable representation of the actual system. It reproduces system behaviour with enough fidelity to satisfy analysis objectives. In Table 1 above, the validation was done by comparing the values of the developed survival predictive model with the actual values collected from the two health institutions (historical data) and it was discovered that the values in the Multilayer perception model column were almost the same as the actual values (Historical data) using all the input variables (CD4 count, Viral load, Opportunistic infections and Nutritional status).

Table 2: Validation of the Result of the three Models

CD4	Viral Load	Nutritional Status	Opportunistic Infection	Survival	MLP
High	Low	High	Yes	Yes	Yes
Low	High	High	Yes	No	No
High	Low	Low	Yes	No	No
High	Low	High	Yes	Yes	Yes
High	Low	Low	Yes	No	No
High	Low	High	Yes	Yes	Yes
High	Low	High	Yes	Yes	Yes
High	Low	Low	Yes	No	No
High	Low	High	Yes	Yes	Yes
Low	High	Low	Yes	No	No
Low	High	Low	No	No	No
High	Low	High	Yes	Yes	Yes
High	Low	High	Yes	Yes	Yes
High	Low	Low	Yes	Yes	No
High	Low	Low	Yes	No	No
High	Low	High	Yes	Yes	Yes
High	Low	High	Yes	Yes	Yes
High	Low	Low	Yes	No	No
High	Low	High	Yes	Yes	Yes

V. CONCLUSION

In order to reduce child mortality which is one of the important millennium development goals and one of the development indicators of any country, there is the need to have effective and efficient model that can be used to predict the survival status of pediatric HIV/AIDS patients in Southwestern Nigeria. Hence, this particular study. The Multilayer Perception predictive model would be recommended for the prediction of peadiatric HIV/AIDS patient's survival from the above analysis.

REFERENCES

Avert (2014b). HIV Treatment for Children. Available from www.avert.org/hiv-treatment-children.htm. [Accessed 13, August 2014]

Brain, L., Tshilidzi, M., Taryn, T.and Monica, L (2006). Prediction of HIV Status from Demographic Data using Neural Networks. IEEE conference on Systems Man and Cybernetics. Taipei, Taiwan.

Centre for Disease Control (2011). HIV Classification: CDC and WHO Staging System. Available from www.hab.hrsa.gov.

Dalton, C., Elizabeth, M., Carey, F., Barbra, A., Dorothy, A., Irene, I., Sara, B. and **Grace, J. (2010).** Predictors of Mortality in HIV-1Infected Children on Antiretroviral Therapy in Kenya: A Prospective Cohort. *Journal of Pediatrics.* 10(33) : 1471-2431.

Eric, J. and Daria, J. (2012). HIV-1 Antiretroviral Drug Therapy. *Journal of Cold Spring Harb Perspect Medical,* 2(4) : 10-12.

Hoa, M. (2011). ART Adherence among People Living With HIV/Aids in North Vietnam, Queensland University of Technology, Brisbane Australia.

Idowu, P. (2012). *Development of A Web-Based Geo-Spatial Environmental Health Tracking System for South Western Nigeria*. Un Published Phd Thesis Submitted to Department of Computer Science and Engineering, Obafemi Awolowo University Ile-Ife, Nigeria.

Joyce, B. (2014) **Towards a New Global Development Agenda. Available from** www.unaids.org.

Kama, K. and Prem, S. (2013). Utilization of Data Mining Techniques for Prediction and Diagnosis of Major Life Threatening Diseases Survivability-Review. International *Journal of Scientific & Engineering Research*, 4(6) : 923-932.

Lucia, O. (1996). Sequential Use of Neural Network for Survival Prediction in AIDS. Available from www.ncbi.nlm.nih.gov/pmc/articles/PMC2233186/. [Accessed 4th August, 2014]

Mira, J. and Denis, A. (2010).The Cost-Effectiveness of Preventing Mother-to-Child Transmission of HIV in Low- and Middle-Income Countries: Systematic Review. *Journal of Bio Med Central*, 9(3) : 45-52.

Moore, D., Hogg, R., Yip, B., Craib, K. Wood, E. and Montaner, J. (2006). CD4 Percentage is An Independent Predictor of Survival in Patients Starting Antiretroviral Therapy with Absolute CD4 Cell Counts between 200 And 350 Cells/Microl. *Journal of HIV Medical*, 7: 383-8.

Ojunga, N., Peter M., Otulo W., Omollo, O. and Edgar O. (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*, 3(3) : 14-20.

Petros, I.,Marie-Eve, R., Vantha, T., Chhraing, S., Kazumi, A., Varun ,K., Sopheak, N., Eric, N., Rony, Z. (2010). High Survival and Treatment Success Sustained After Two and Three Years of First- Line ART for Children in Cambodia. *Journal of International AIDS Society*, 13(11) : 13-19

Sameem, A., Raviraja, S., Namir, A., Adeeba K., and Annapurni, K. (2010). Classification and Regression Tree in Prediction of Survival of AIDS Patients, *Malaysian Journal of Computer Science*, 23(3) : 153-165.

Sanjel, S., (2009). An Application of Cox Model for Lifetimes of HIV Patients. Unpublished M.Sc. thesis submitted to the department of Power Analysis. USA. MC Master University.

Shearer W. (2000). Evaluation of Immune Survival Factors in Pediatric HIV-1 Infection. *Annual National Academic Journal*, 91(8) : 298-312.

UNAIDS (2008). Report on The Global AIDS Epidemics Available from http://www.unaids.org. UNAIDS, (2012). Report on the Global AIDS Epidemic**.** Available from http://www.unaids.org. [Accessed February 10, 2013].

World Health Organization (2012). Towards Universal access: Scaling up Priority HIV/AIDS interventions in the Health Sector. Progress Report 2010.