IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 23, Issue 2, Ser. II (Mar. – Apr. 2021), PP 29-36
www.iosrjournals.org

HIV/AIDS Viral Load Prediction using Artificial Neural Network Approach

Magai, A.S.[1], Manga, I.,[2]Danladi,A.[3]
[1](Department of Computer Science, Adamawa State UniversityMubi, Nigeria)
[2](Department of Computer Science, Adamawa State UniversityMubi, Nigeria)
[3](Department of Physics, Adamawa State University Mubi, Nigeria)

Abstract:*Human Immunodeficiency Virus (HIV) is a virus which leads to acquired immune deficiency Syndrome (AIDS), which causes failure of the immune system thus, allowing other infections, ulcers, cancers and other diseases to affect the body and thrive. This research aimed at the use of multi-layer artificial neural networks with back propagation to predict the HIV/AIDS viral load levels over a given period of time. Data were collected from Adamawa State Specialist Hospital, Yola Adamawa State. R – Programming and STATA Version 15 were used to analyses the data. A total of 546 records were collected, in classification, 365 records were used in training set while 181 were used as a testng set. Performance accuracy of the model was checked using confusion matrix, ROC and other derived evaluation attributes such as sensitivity, specificity, and precision.*

Key Word*: Multi layer, Neural network, Viral load, Confusion Matrix, Sensitivity.*

Date of Submission: 09-03-2021 Date of Acceptance: 23-03-2021

I. Introduction

Since the discovery of Human Immunodeficiency Virus (HIV) as a virus in 1978 no cure has been found. This is majorly attributed to the dynamic nature of the virus that keeps morphing into new forms after very short periods of time. In Nigeria, the first case of Human Immunodeficiency Virus/Acquired Immune Deficiency Syndrome (HIV/AIDS) infection occurred in 1986. It was later declared a National disaster November 1999 by the then president. Then National Action committee on AIDS (NACA) was established to coordinate all AIDS programs in the country.HIV is a virus which leads to acquired immune deficiency Syndrome (AIDS), which causes failure of the immune system thus, allowing other infections and cancers to affect the body and thrive [1]. The virus is a retrovirus that attacks the human Cluster of Differential 4 (CD4+) cells, causing a decline in their natural defenses against pathogenic micro-organisms [2]. It belongs tothe Retroviridae family which is considered as a highly evolved virus type, and which can replicate in the host cells through Reverse Transcription process [3].

[4]States that there are two major phenotypes of the HIV virus, namely HIV-1 and HIV-2. HIV -1, which this study will focus on has three strains; labelled as M (Major), O (Outlier) and N (New i.e. not M or O). He also states that the strain that is almost entirely to blame for the global pandemic is the Group M, which has a lot of diversity. HIV-2 is relatively uncommon and is concentrated majorly in the West of Africa. This phenotype is less infectious and progresses slower as compared to HIV -1[5]. HIV in this study will refer to the more common HIV -1 phenotype.

The majority of the infected people within Nigeria are aged between 15-45 years which is a reflection of a population that has over 50% of its people being less than 16 years of age. The prevalence rate of HIV amongst this group is 5.9%. The major factor contributing to the high incidence of HIV/AIDS in Nigeria has been attributed to the high level of poverty among Nigerians where over 50 percent of the population lives with an average annual basic income of less than $1 per day [6].

Consequentially, a lot of research has gone into trying to come up with a solution to HIV/AIDS with the recent temporal solution being the invention of the Antiretroviral Therapy (ART) drugs, composed of a compound of medicines aimed at slowing down the rate at which the HIV virus replicates itself. However, a bigger quartile of the population of the third world countries is still suffering from logistical challenges such as lack of adequate medical equipment and medical supplies in the hospitals, and the high prices of undertaking the activities and tests. Worst case scenarios have included the introduction of ART in the late stages of HIV patients [7].

In the recent decades the use of data relating to HIV protein levels in the plasma, also known as clinical markers have been used to estimate prognosis in HIV-1 infection. In as much as it has been affirmed that the best predictor of AIDS onset characterized to date is the percentage or absolute number of circulating (Cluster of

Differentiation 4 positive) CD4+ T cells, a marker or combination of the same have recently been used to assess risk before substantial immune destruction kicks in [8]. The CD4+ count is a measure of the number of white blood cells per millilitre of blood that contain the CD4 glycoprotein. The CD4+ cells are usually developed in response to infections [1]. Viral load on the other end is a measureof the actual number of viral particles per millilitre of blood. This count is more accurate than the CD4+ count since CD4+ cells are usually detected after the drug resistance has been developed, and can also be affected by other factors other than HIV infection, such as other infection [9].

Other than just research, a lot of resources have gone into sponsoring activities to predict an improvement in a patient's viral load. This in most instances has normally entailed the participating competitors being provided with data on the nucleotide sequences of the Reverse Transcriptase (RT), the Protease (PR), the viral load and the CD4 count of different cases that suffer from HIV as at the beginning of therapy. The variables provided in these competitions have proven to be reliable and highly co-relational to the levels of HIV within a patient's body. The predictions output are normally then tested against a number of real cases. These exercises have over the time contributed to the use of data guided by the aforementioned parameters amongst others to better place the status of the HIV patients [10]. Both the Viral load and CD4+ count are instrumental in measure of HIV progression.

II. Material and Methods

Data were collected from Adamawa State Specialist Hospital, Yola of Adamawa State. In this study, Artificial Neural Networks were used to predict the HIV/AIDS viral load levels over a given period of time by study the existing models, and checked performance accuracy of the model. A total of 546 records were collected, in classification, 365 records were used in training set while 181 were used as a testing set. Performance accuracy of the model was checked using confusion matrix, ROC and other derived evaluation attributes such as sensitivity, specificity, and precision. R – Programming and STATA Version 15 were used to analyze the data.

Model Development

To develop the artificial neural network, the steps followed were as follows:

i) Obtaining data
ii) Pre-processing of data
iii) Development of the model
iv) Testing of the model
v) Validation of the model

Classifier Performance Evaluation Criteria

Confusion Matrix

In order to check the accuracy of our results, we will employ confusion matrix to evaluate the performance of the algorithm. This performs a more crucial analysis than accuracy alone. Each of the attribute in matrix represents the pattern in the anticipated data class where every false-positive (FP) metric". Confusion matrix is a table describing the various performance of a classifier; information about the predicted and actual classifications was done. To measure the performance, the data has to be represented in a table in the form of an m (n matrix i.e. number of rows equal's number of columns. The performance is carried out using the value found on each row and column intersection.

i. True-Positive indicates 'positives' pattern
ii. False-Positives indicates the amount of 'negative' patterns seen as positive patterns
iii. False-Negative implies amount of 'positive' seen as negative patterns
iv. True-Negative means negative 'patterns' seen as negatives

Table 1: Presents a Confusion Matrix Layout for two Class Classifiers

		Actual Class	
		Positive	**Negative**
Predicted Class	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Measures

Accuracy Measure–is the ability of a model to appropriately predict the class label of previously unseen data or new data. It is a measure of how well the classifier makes a prediction on average. A good classification

algorithm will try to minimize the number of times it makes the wrong prediction. Total number of correctly classified patterns divides by total number of patterns

Sensitivity: A true positive is when the outcome of a prediction is said 'P' and the classifier have actually predicted the value to be same 'P'. It is a measure of comprehensiveness or magnitude. It also implies how different values and independent variable affect a dependent variable.

Specificity: Is also called negative rate, measures the negative pattern correctly identified, it indicates the number of topples classified as false while they were actually false. It is also the ability of the classifier to classify those that were false correctly: or a negative rate, measures the negative pattern correctly identified

Positive Predictive Value – values of positive patterns

Negative Predictive Value - where negative occurs

Receiver Operator Characteristics test (ROC) - The plot of TP against FN

Multiple Linear Regression

Multiple linear regression models are concerned with predicting values of one response (Y) variable based set of predictor variables (X_i). Ordinary least square method was used to get the parameter (β_i) estimates of the model. The multiple linear regression models are given by:

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \mathrm{K} + \beta_k X_k \qquad \text{Where } i = 1,2,\Lambda\ ,k$$

Where
Y= Response or Dependent Variable
X=Predictor or Independent Variables
β_o =Intercept or Constant
β_1 =Slope or Regression Coefficient

III. Results

This framework of HIV classification using neural network classifier with a pre-processed; all the various stages of pre-processing, normalization, training and testing of data, classification, measures of accuracy are implemented using R programming.

Table 2: Case Processing Summary

Sample	Frequency	Percentage (%)
Training	365	66.8%
Testing	181	33.2%
Valid	546	100.0%
Excluded	0	0
Total	546	

Table 2 above shows the case processing summary of the result, the training process used 365 representing 66.8% of the data while the testing process used 181 representing 33.2% of the data.

Table 3: Classification of Training and Testing Result

Sample	Observed	Absence Viral Load	Present Viral Load	Percent correct
Training	Absence Viral Load	129	62	67.5%
	Present Viral Load	11	163	36.2%
	Overall Percent	65.8%	34.2%	52.6%

Testing	Absence Viral Load	66	25	72.5%
	Present Viral Load	59	13	34.4%
	Overall Percent	69.1%	30.9%	53.6%

Dependent Variable: Categorized Test result

Table 3 shows the classification performance of the training and testing process. Of the 365 cases considered in the Training set, 67.5% (129 cases) were correctly classified as showing the absence of the viral load while 36.2% (63 cases) show the presence of viral load. The model correctly classified 65.8% of the Training set with absence of viral load after the medical test and 34.2% with presence of viral load after the medical test. An overall correct classification rate was obtained after analysis as 0.526 (52.6%). From the Testing set of 181 cases, 72.5% (66 cases) are correctly classified as showing absence of the viral load while 34.4% (31 cases) show correctly the presence of the viral load. The model correctly classified 69.1% and 30.9% as having the absence and presence of viral load respectively. The Testing set thus show an overall correct classification of 53.6%.

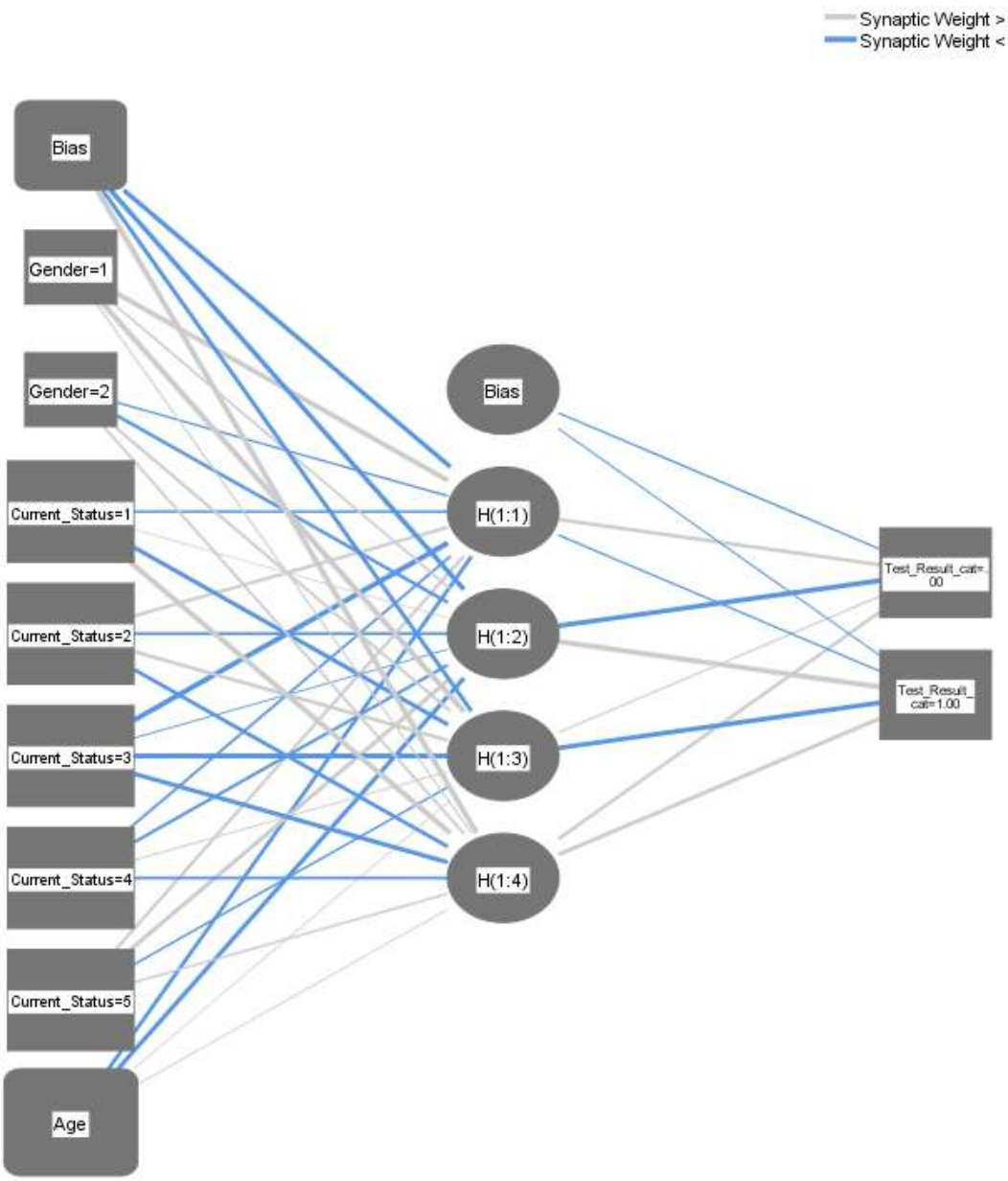

Figure 1: *Hidden layer of neural network*

Performance Evaluation

To demonstrate the accuracy of the system, we divide the data into 365 records for training and 181 records for the validation testing set in phases. The system has achieved 53.6% accuracy of a correctly classified instance during prediction in *classification for testingwhile 46.4% is i*ncorrectly classified.

Confusion Matrix

Classification for Testing Model $Confusion\ Matrix = \begin{pmatrix} 66 & 25 \\ 59 & 31 \end{pmatrix}$

Where,

66 is the *N0* of accurate predictions that a given instance is "-",
59 is the *N0* of incorrect predictions that a given instance is "+",
25 is the *N0* of incorrect of predictions that a given instance "-",
31 is the *N0* of accurate predictions that a given instances "+".

Accuracy= $\frac{TP+TN}{N} = \frac{66+31}{181} = 0.5359 = 53.6\%$

Sensitivity= $\frac{TP}{FN+TP} = \frac{66}{25+66} = 0.7253 = 72.5\%$

Specificity= $\frac{TN}{TN+FP} = \frac{31}{31+59} = 0.3444 = 34.4\%$

False positive rate (FPR) = $\frac{FP}{TN+FP} = \frac{59}{31+59} = 0.6555 = 65.6\%$

False negative rate (FNR) = $\frac{FN}{FN+TP} = \frac{25}{25+66} = 0.2747 = 27.5\%$

Precision= $\frac{TP}{TP+FP} = \frac{66}{66+59} = 0.528 = 52.8\%$

ROC= $\frac{Sensitivity + Specificity}{2} = \frac{72.5+34.4}{2} = 53.45\%$

Plot of Learning Convergence

We can as well plot the neural network's convergence after the learning phase for a pictorial representation and clarity showing the inputs vectors, weights, hidden layers, the bias and the output. There are two coloured lines visible in the plot; they blue lines which represent the bias factors with their weight and the black lines for the forward neural connection and with their weights too. Fig. 2 is the graphical presentation of the neural network structure of the trained data.

Figure 2 is the graphical presentation of the neural network structure of the trained data.

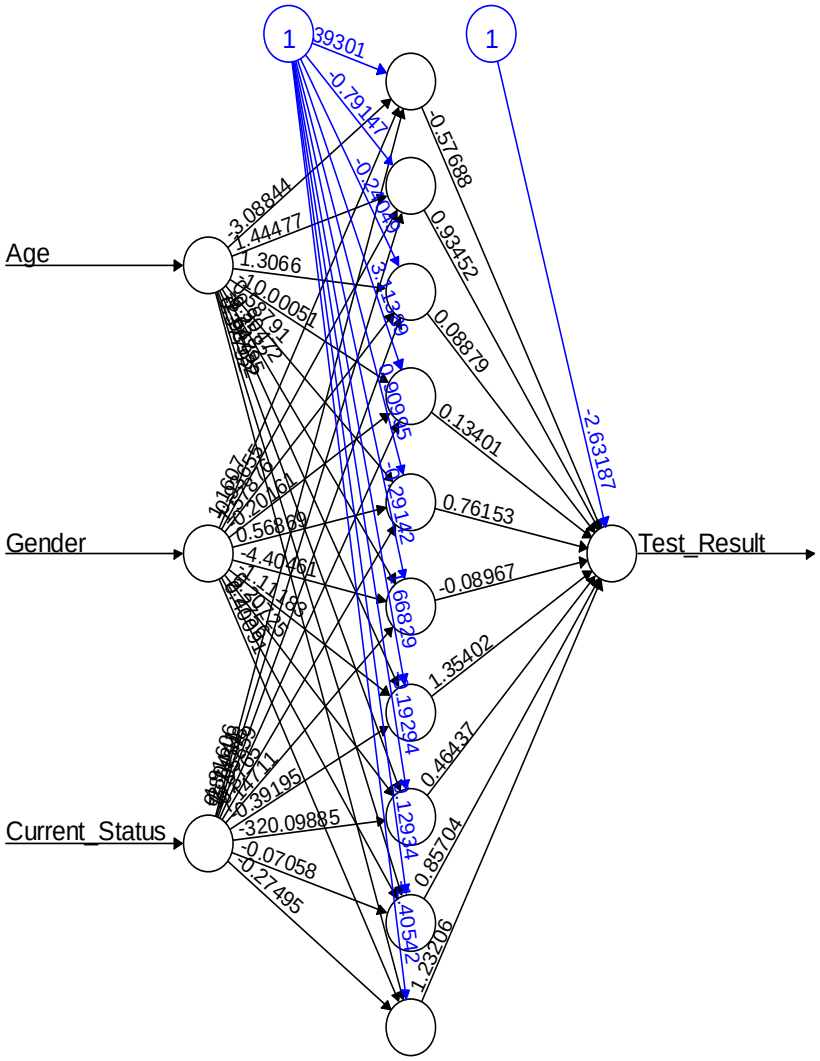

Table 4: Multiple Regression of test result on age, gender and current status

Variables	Coefficient	Standard Error	t-statistic	Prob
Constant	39225.6	17612.4	2.227	0.0263*
Age	-5.18	0.36	-1.438	0.1511
ART Transfer in (Ref)				
ART Start	16312.3	15875.1	1.028	0.3046
HIV + non Start	127531.3	71969.4	1.772	0.0770**
Lost to follow-up	-25547.3	58793.3	-0.435	0.6641
ART Transfer Out	-17517.1	58862.8	-0.298	0.7661
Female (Ref)				
Male	10313.5	10331.7	0.998	0.3186

Residual standard error: 101500 on 539 degrees of freedom

F-statistic: 1.386 on 6 and 539 df, p-value: 0.2181

** Significant at 5%* *** Significant at 10%*

Table 4 is obtained from the R- Programming output for the analysis of the multiple linear regressions relating the test result as a function of age, gender and current status. T From the result above HIV + non start has significant effect on HIV patient at 10% level.

Table 5: Independent Variable Importance

Variable	Importance	Normalized Importance
Gender	0.0336	3 %
Current Status	0.4408	3.4 %
Age	0.5271	100.0 %

Importance of Independent Variables:

Table 5 shows the importance and normalized importance of each predictor in determining the neural network. The analysis is based on the training and testing samples. The importance of an independent variable is a measure of how much the network's model-predicted value changes for different values of the independent variable. Moreover, the normalized importance is simply the importance values divided by the largest importance values and expressed as percentages. From table 5, it is evident that "Age" contributes most in the neural network model construction, followed by "Current Status" of the HIV/AIDS patients as shown in the diagram below.

Figure 3 *Variable importance*

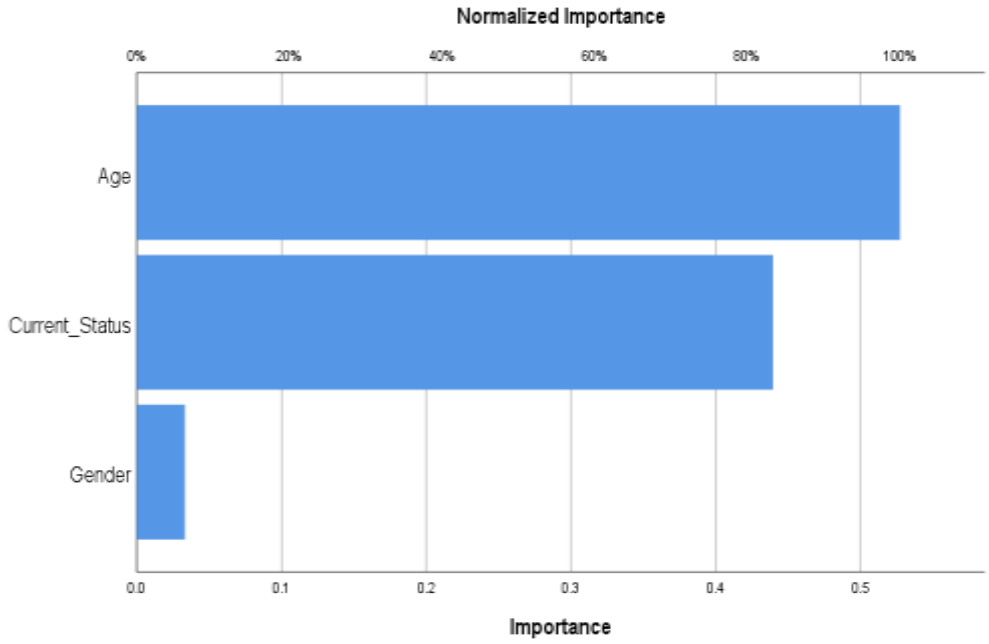

Table 4.6: Summary of the Measures and Performance Evaluation

Measure	**Performance**
Accuracy	53.6 %
Sensitivity	72.5 %

Specificity	3	4	.	4 %
False Positive Rate	6	5	.	6 %
False Negative Rate	2	7	.	5 %
Precision	5	2	.	8 %
ROC	5	3	.	45 %

V. Conclusion

The techniques used in this research demonstrated a better performance. The parameters used in this study are distinct, non-influencing by one another. The system is designed to accept input of parameters from the user and automatically predict the class label based on the training data. During prediction, the system will also automatically display the performance accuracy of the model. The model correctly classified 69.1% and 30.9% as having the absence and presence of viral load respectively. Also, ROC demonstrated a more optimal result than confusion matrix. In general, the entire result demonstrated that the system can be efficient and reliable in predicting HIV/AIDS after treatment. We therefore recommended that the system would assist the clinician in understanding the various complications that may result from HIV/AIDS cases.

References

[1]. Chou, D., Iu, R., Krishna, R., & Liang, A. (2012). An Analysis on the Prediction of HIV Progression.
[2]. HIV Viral Load Blog. (2016, 11 28). The difference between HIV viral load and CD4 tests. Retrieved from HIV Viral Load Blog: http://www.hivviralload.com/ blog/2008/7/10/thedifference-between-hiv-viral-load-and-cd4-tests.html
[3]. Levy, J. A. (2007). HIV and the Pathogenesis of AIDS. Washington: ASM Press.
[4]. Archer, J. P. (2008). The Diversity of HIV -1 Manchester: University of Manchester.
[5]. Averting HIV and AIDS. (2016, 11 25). HIV Strains and Types. Retrieved from Averting HIV and AIDS: http://www.avert.org/professionals/hiv-science/types-strains
[6]. Lopez, W. (2011). HIV/AIDS: A New Era of Treatment. The York Scholar, 11-17. Medecins sans Frontieres. (2010). The Ten Consequences of AIDS Treatment. Medecins sans Frontieres.
[7]. National AIDS and STI Control Programme. (2014). KENYA HIV ESTIMATES. Nairobi: National AIDS and STI Control Programme.
[8]. Rosa, R. S., Santos, R. H., Brito, A. Y., &Guimaraes, K. S. (2014). Insights of prediction of patients' response to anti-HIV therapies through machine learning. Recife: Federal University of Pernambuco.
[9]. Shafer, R. W., Dupnik, K., Winters, M. A., &Eshleman, S. H. (2001). A Guide to HIV-1 Reverse Transcriptase and Protease Sequencing for Drug Resistance Studies. Stanford: Stanford University.
[10]. United Nation AIDS. (2016). The need for routine viral load testing. United Nation AIDS. University of Edinburgh. (2017, March 26). *Note 14.* Retrieved from University of Edinburgh Informatics: http://www.inf.ed.ac.uk/teaching/courses/ms/notes/note14.pdf