



Strathmore
UNIVERSITY

Strathmore University
SU+ @ Strathmore
University Library

Electronic Theses and Dissertations

2017

A HIV/AIDS viral load prediction system using artificial neural networks

Titus Kipkosgei Tunduny
Faculty of Information Technology (FIT)
Strathmore University

Follow this and additional works at <https://su-plus.strathmore.edu/handle/11071/5684>

Recommended Citation

Tunduny, T. K. (2017). *A HIV/AIDS viral load prediction system using artificial neural networks*

(Thesis). Strathmore University. Retrieved from <http://su-plus.strathmore.edu/handle/11071/5684>

This Thesis - Open Access is brought to you for free and open access by DSpace @ Strathmore University. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of DSpace @ Strathmore University. For more information, please contact librarian@strathmore.edu

A HIV/AIDS Viral Load Prediction System Using Artificial Neural Networks

Tunduny Kipkosgei Titus

Master of Science in Information Technology

2017

A HIV/AIDS Viral Load Prediction System Using Artificial Neural Networks

Tunduny Kipkosgei Titus

063864

**A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of
Science in Information Technology (MSc.IT) at Strathmore University**

**Faculty of Information Technology
Strathmore University
Nairobi, Kenya**

June, 2017

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Declaration and Approval

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the research itself.

© No part of this thesis may be reproduced without the permission of the author and Strathmore University.

Tunduny Kipkosgei Titus[Name of Candidate]

..... [Signature]

7th June 2017[Date]

Approval

This thesis of Tunduny Kipkosgei Titus was reviewed and approved by:

Dr. Bernard Shibwabo,
Academic Director, Faculty of Information Technology,
Strathmore University

Dr. Joseph Orero,
Dean, Faculty of Information Technology,
Strathmore University

Prof. Ruth Kiraka,
Dean, School of Graduate Studies,
Strathmore University

Abstract

Human Immunodeficiency Virus (HIV) has been affecting people since it was first discovered in 1986. This is as a result of the HIV virus being present in the patient bloodstream for the remainder of their normal life, as there is no cure that exists as of now. HIV, if left unmanaged would end up developing into Acquired Immune Deficiency Syndrome (AIDS), a syndrome that weakens a patient's immune system and leaves them susceptible to other opportunistic infections. Antiretroviral therapy (ART) has been successfully used in managing the progression of the HIV virus in the human body. However, poor adherence attributable to ignorance, adverse drug effects, and age have derailed the attainment of viral load suppression amongst the HIV positive people. The progression of the virus is tracked by counting Cluster of Differentiation 4 positive cells, and the amount of virus in the blood (viral load) every 6 months. This research introduces the use of multi-layer artificial neural networks with backpropagation to predict the HIV/AIDS viral load levels over a given period of time (in weeks). The Data-driven Modelling methodology was used in the development of the model. This methodology was ideal since the model relied solely on pre-existing data, and supports artificial neural networks. The model developed performed at an accuracy level of 93.76% and a mean square error of 0.0323. The results showed that the neural network can be used as a suitable algorithm for HIV/AIDS viral load level prediction. The learning rate used in the study was 0.005 and the momentum was 0.9. The iterations for the training, testing and validation varied.

Keywords: Neural Networks, Viral Load, CD4+, Data-driven Modelling

Dedication

This research is dedicated to my family; my wonderful parents Ben and Susan Tunduny, and my siblings, Collins, Patrick and Isaack. Thank you for walking through the arduous journey with me. To my cousins; Robert, Nelson, Martin, Nathan and the entire family, thank you for the encouragement. To my friends; Bernard, Kevin, Sharon, Eunice, and Tiberius, thank you for walking with me. And to the Almighty God, for granting me the ability to be able to undertake the research. I appreciate it.

Acknowledgement

I would like to sincerely thank Dr. Bernard Shibwabo, whose guidance, insight and wisdom helped me complete the project on time. I would also like to thank Dr. Vincent Omwenga who also helped frame the ideas for the project. I would also like to appreciate Prof. Ismael Ateya who provided the knowledge on how to structure systems. I would also like to acknowledge my parents Ben and Susan Tunduny, my brothers Collins, Patrick and Isaack, my Aunts Esther, Rhoda, and my cousins Nelson, Robert, Martin, Terry, Nathan and Ivy for the direct support and words of encouragement and immense support throughout the project. I would like to acknowledge the MSc. IT class of 2017 for making light of the hard situations and making them manageable.

Table of Contents

Declaration and Approval.....	ii
Approval	ii
Abstract.....	iii
Dedication	iv
Acknowledgement	v
Table of Contents	vi
List of Figures.....	x
List of Tables	xi
List of Equations	xii
Abbreviations/Acronyms.....	xiii
Chapter 1 : Introduction	1
1.1. Background	1
1.2. Problem Statement	3
1.3. Aim.....	3
1.4. Research Objectives	4
1.5. Research Questions	4
1.6. Justification	4
1.7. Scope and Limitation	5
Chapter 2 : Literature Review.....	6
2.1. Introduction	6
2.2. Factors influencing HIV/AIDS Progression	6
2.2.1. How HIV Infects the Body	6
2.2.2. Antiretroviral Therapy Adherence	8
2.2.3. HIV Superinfection and Coinfection	10

2.3.	Methods for measuring HIV/AIDS Progression	11
2.3.1.	Viral Load	11
2.3.2.	Cluster of Differentiation 4	13
2.4.	Current HIV/AIDS Prediction Systems	13
2.4.1.	Prediction of HIV/AIDS Status using Artificial Neural Networks	13
2.4.2.	Random Forest Algorithm	16
2.4.3.	Support Vector Machines	18
2.4.4.	Linear Regression	18
2.4.5.	K-Nearest Neighbor	20
2.5.	Conceptual Model	21
Chapter 3 :	Research Methodology	23
3.1.	Introduction	23
3.2.	Research Design	23
3.3.	Model Development	23
3.3.1.	Obtaining Data	23
3.3.2.	Data Pre-processing	24
3.3.3.	Development of the Model	24
3.3.4.	Validation of the Model	24
3.4.	System Development Methodology	24
3.4.1.	Application of the Data-driven Modelling Methodology	25
3.4.2.	Techniques Used	25
3.4.3.	Model Testing and Deployment	25
3.5.	Research Quality	26
3.6.	Ethical Considerations	27
Chapter 4 :	System Design and Architecture	28
4.1.	Introduction	28
4.2.	Requirements Analysis	28
4.2.1.	Functional Requirements	28

4.2.2.	Usability Requirements	28
4.2.3.	Reliability Requirements	29
4.2.4.	Supportability Requirements	29
4.3.	System Architecture	29
4.4.	Use Case Diagram	30
4.5.	System Sequence Diagram	34
4.6.	Flow Chart	35
4.7.	Database Schema	36
Chapter 5 :	Implementation and Testing.....	38
5.1.	Introduction	38
5.2.	Model Components	38
5.2.1.	System Components.....	38
5.2.2.	Neural Network Components	40
5.3.	Model Implementation	41
5.3.1.	Data Input.....	41
5.3.2.	Drug Resistance Computation	42
5.3.3.	Scaling Data	42
5.3.4.	Software Flow	43
5.4.	Model Architecture	44
5.5.	Model Validation and Testing.....	45
5.5.1.	Functional Requirements	46
5.5.2.	Usability Requirements.....	47
5.5.3.	Reliability Requirements	48
5.5.4.	Supportability Requirements	49
Chapter 6 :	Discussion	50
6.1.	Introduction	50
6.2.	Model Validation.....	50

6.3.	Model Implementation Outputs	51
6.3.1.	Training Outputs	51
6.3.2.	Testing Outputs	52
6.3.3.	Validation Outputs	53
6.4.	Contributions to Research	54
6.5.	Challenges	54
Chapter 7 :	Conclusions and Recommendations	55
7.1.	Conclusions	55
7.2.	Recommendations	56
7.3.	Suggestions for Future Research.....	56
References		58
Appendices		65

List of Figures

Figure 2.1: HIV virus infects the CD4+ cell	7
Figure 2.2: Graph showing relation between Transmission Probability of HIV per Sexual Act and the Viral Load	12
Figure 2.3: Structure of an Artificial Neural Network	14
Figure 2.4: Mathematical model of a Neuron	14
Figure 2.5: The Sigmoid Activation Function Graph	15
Figure 2.6: Treatment Change Episode	17
Figure 2.7: Support Vector Machines Example	18
Figure 2.8: Homoscedasticity	20
Figure 2.9: Sample of K-Nearest Neighbor Classification	21
Figure 2.10: Conceptual Framework for the system.....	22
Figure 3.1: Data-driven Modelling	25
Figure 4.1: System Architecture	30
Figure 4.2: System Use Case Diagram	31
Figure 4.3: System Sequence Diagram.....	35
Figure 4.4: Flow Chart for the System.....	36
Figure 4.5: Database Schema of the System.....	37
Figure 5.1: Public Input Section, Showing Nucleotide Input field, & Fasta file Upload Section	39
Figure 5.2: Add New Treatment Change Profile Interface.....	40
Figure 5.3: Model Architecture.....	45

List of Tables

Table 2.1: Table showing published cases of HIV-1 superinfection	11
Table 4.1: Data Pre-processing, Standardization and Feature Extraction	32
Table 4.2: Model Training and Testing	33
Table 4.3: Obtaining the Predicted Viral Load Level.....	34
Table 5.1: Table Showing Unstandardized Data	43
Table 5.2: Table Showing Standardized Data.....	43
Table 5.3: Functional Requirements	46
Table 5.4: Usability Requirements	47
Table 5.5: Reliability Requirements	48
Table 5.6: Supportability Requirements	49
Table 6.1: Sample Expected Output and Predictions on Training.....	51
Table 6.2: Sample Expected Outputs and Predicted Outputs on Testing	52
Table 6.3: Sample Expected Outputs and Predicted Outputs on Validation	53

List of Equations

Equation 2.1: Sigmoid Equation Used	15
Equation 2.2: Linear Regression Formula	19
Equation 2.3: Euclidean Distance	21
Equation 3.1: Formula for Computing Mean Squared Error	26
Equation 3.2: Coefficient of Determination Formula	26
Equation 5.1: Logarithmic Formulae for Finding the logload value	41
Equation 5.2: Determining Drug Resistance for a Patient	42
Equation 5.3: Standard Scaler Formula	42

Abbreviations/Acronyms

AIDS	- Acquired Immune Deficiency Syndrome
ANN	- Artificial Neural Networks
ART	- Antiretroviral Therapy
CDC	- Centers for Disease Control and Prevention
DNA	- Deoxyribonucleic Acid
DDM	- Data-driven Modelling
HIV	- Human Immunodeficiency Virus
PLWHA	- People Living with Human Immunodeficiency Virus and Acquired Immune Deficiency Syndrome
RNA	- Ribonucleic Acid
STBBI	- Sexually transmissible diseases and blood-borne infections
TCE	- Treatment Change Episode

Chapter 1 : Introduction

1.1. Background

Since the discovery of Human Immunodeficiency Virus (HIV) as a virus in 1986 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 Kenya, the first case of Human Immunodeficiency Virus/Acquired Immune Deficiency Syndrome (HIV/AIDS) infection occurred in 1978, in communities living around the shores of Lake Victoria. It was later declared a National disaster November 1999 by the then president. Then National Aids Control Council (NACC) was established to coordinate all Aids programs in the country (National AIDS and STI Control Programme, 2014).

HIV is a virus which leads to acquired immunodeficiency virus (AIDS), which causes failure of the immune system thus allowing other infections and cancers to affect the body and thrive (Chou, Iu, Krishna, & Liang, 2012). The virus is a retrovirus that attacks the human Cluster of Differential 4 (CD4+) cells, causing a decline in their natural defenses against pathogenic microorganisms (Rosa, Santos, Brito, & Guimaraes, 2014). It belongs to the Retroviridae family which is considered as a highly evolved virus type, and which can replicate in the host cells through Reverse Transcription process (Levy, 2007).

Archer (2008) 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 (Averting HIV and AIDS, 2016). HIV in this study, will refer to the more common HIV -1 phenotype.

The majority of the infected people within Kenya are aged between 15-39 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 Kenya has been attributed to the high level of poverty among

Kenyans where over 50 percent of the population lives with an average annual basic income of less than \$1 per day (National AIDS and STI Control Programme, 2014;UNAIDS, 2017).

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 a HIV patients (Lopez, 2011).

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 (United Nation AIDS, 2016). The CD4+ count is a measure of the number of white blood cells per milliliter of blood that contain the CD4 glycoproteins. The CD4+ cells are usually developed in response to infections (Chou, et al., 2012). Viral load on the other end, is a measure of the actual number of viral particles per milliliter 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 (HIV Viral Load Blog, 2016; Chou, et al., 2012).

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 (Shafer, Dupnik, Winters, & Eshleman, 2001).

1.2. Problem Statement

ART medicine dispensation has for some time been used as a combative mechanism to slow the quick spread of HIV virus within the infected persons. In addition to this, viral load tests are normally carried out as an important exercise to aid the medical practitioners know how much the virus has spread in the body for purposes of monitoring the levels or kinds of ART to administer to the infected persons. This is because, as the viral load increases, the response to the medicines administered also gets altered and in most cases thus needs to be changed; with fewer cases of adherence to the original prescribed drug. Failure to this leads to unresponsiveness to the ART and thus puts the patients at a graver risk (Institut National de Sante' Publique du Quebec, 2014).

At the moment, especially within Kenya, patients are advised to visit the hospital at least twice a year for viral load tests which are conducted clinically through the use of various clinical markers within the hospitals' laboratories. This approach is so far not effective due two major reasons. First, most patients, as dictated by their lifestyle are not committed to keeping these yearly appointments and thus with time end not responding to the prescribed ART. Such lifestyle range from poor eating habits to alcoholism (Nyaga, et al., 2004). The other major factor that makes this approach ineffective is the limited resources required to conduct these activities within the hospitals in relation to the large population of HIV infected patients. There are CD4 prediction systems that are currently existing. However, the CD4 prediction comes later after the viral load of a given patient is far much overwhelming and thus no longer useful in preventing the quick spread of the HIV virus.

1.3. Aim

The aim of this research is to develop a prediction system for determining viral load levels at desired times for HIV-positive patients.

1.4. Research Objectives

- i. To investigate the factors influencing HIV/AIDS progression in human beings
- ii. To analyze the methods used for the measurement of HIV/AIDS progression
- iii. To review the existing prediction algorithms and models used in prediction of HIV/AIDS progression
- iv. To develop an HIV/AIDS Viral load prediction system
- v. To test the functionality of the developed HIV/AIDS Viral load prediction system

1.5. Research Questions

- i. What are the factors influencing the progression of HIV/AIDS amongst HIV positive patients?
- ii. How is the HIV/AIDS progression measured?
- iii. What are the prediction algorithms and models currently being used to predict HIV/AIDS?
- iv. How can the HIV/AIDS viral load prediction system be developed, and what methodology or framework will be used in the implementation?
- v. How can the functionality of the developed system be tested?

1.6. Justification

Albu and Stanciu (2015) highlight the benefits of using artificial intelligence in the medical field. Simplifying the physician's work, saving their time and energy, detecting imperceptible things that would have otherwise been missed are some of the benefits that they state. They conjecture that through the use of artificial intelligence, we can be able to have more insight into hidden attributes, especially when using a prediction system.

Winters-Miner (2014) also notes that the use of predictive analytics in medicine can lead to improved diagnosis. The doctors will be harnessing the power of models created over a period of time to augment their skill in patient-care. This would ultimately lead to better outcomes for the patient. The predictions can lead to the use of preventive medicine where the patient's lifestyle is changed to prevent health risks in future. Stayerberg (2009) agrees with the assertion on better outcomes for the patient and conjectures that this prediction ability can help weigh the benefits of risk of medicinal side-effects with the rewards, and therefore better medical decisions made.

Therefore prediction systems can be used in medicine to improve patient-care together with doctors' skills and input.

Lopez (2011) states that the People Living with Human Immunodeficiency Virus and Acquired Immune Deficiency Syndrome (PLWHA) can lead normal lives, even after the initial diagnosis with the HIV virus. The author stipulates that the management of HIV/AIDS has been successful through the use of ART, which prolongs the lives of the PLWHA. These ART regimens are provided based on the viral load of the HIV virus in the blood of the patient (Medecins sans Frontieres, 2010).

1.7. Scope and Limitation

The scope of this research was limited to the use of online data to develop the HIV/AIDS viral load prediction system. Due to the sensitivity of the data under consideration, getting the data from local hospitals in Kenya is very challenging as most are very protective. For this reason, the study would utilize anonymized patient data that is publicly available online. The data obtained does not contain any adherence information, so this research is limited to the attributes available, but will also focus on the impact of adherence.

Chapter 2 : Literature Review

2.1. Introduction

In this chapter, we look at the working of the HIV virus in the human body, and how the antiretroviral medication has been used to manage the virus. We also look at the ways in which the HIV virus has been observed to have formed resistance against the available antiretroviral medication. Finally, we evaluate the existing systems that have been used to try and predict the viral load levels in an HIV positive patient and highlight the shortcomings of the research.

2.2. Factors influencing HIV/AIDS Progression

2.2.1. How HIV Infects the Body

Infection of HIV occurs when the HIV pathogen comes in direct contact with the blood of a HIV – negative person. This can occur during unprotected sexual contact with an HIV positive person, sharing of sharp objects such as syringes with an HIV positive person, or an HIV-negative person with an open wound coming into contact with contaminated blood such that their blood mixes with the contaminated blood.

The major means of transmission of HIV is contaminated blood products, sexual contact-homosexual contact being higher than heterosexual contact, and mother to child transmission. It is impossible to detect HIV through the symptoms that are displayed as they largely mimic other infections and diseases. Thus the only definitive way of knowing if one is infected with HIV is through testing of blood or saliva (Cohen, 2007; Levy, 2007).

The HIV tests that are submitted for adults look for the antibodies that are present in the blood. These antibodies are produced in response to the presence of the HIV virus in the blood. In adults, a rapid test can be used to test. The efficiency rate of the tests varies depending on the time since the exposure of the person to the HIV virus. During the first three months, there is a very high level of viral replication, and thus the antibodies may not be detectable. Modern rapid tests, however, have a higher efficiency rate and are able to detect the presence of the virus within two weeks of possible exposure (Miller, et al., 2002).

Once the virus has entered the bloodstream, they begin the process of replication. The virus attacks the white blood cells of a body, also called T-helper cells or CD4+ cells. These CD4+ are important since they help in fighting off diseases and infection in the human body. Given that the

HIV virus does not have the ability to multiply, they replicate within the CD4+ cells through reverse transcription process. This results in the damage of the immune system and thus weakens the body's natural defense system (Averting HIV and AIDS, 2016). Figure 2.1 shows how the HIV virus infects the CD4+ cells.

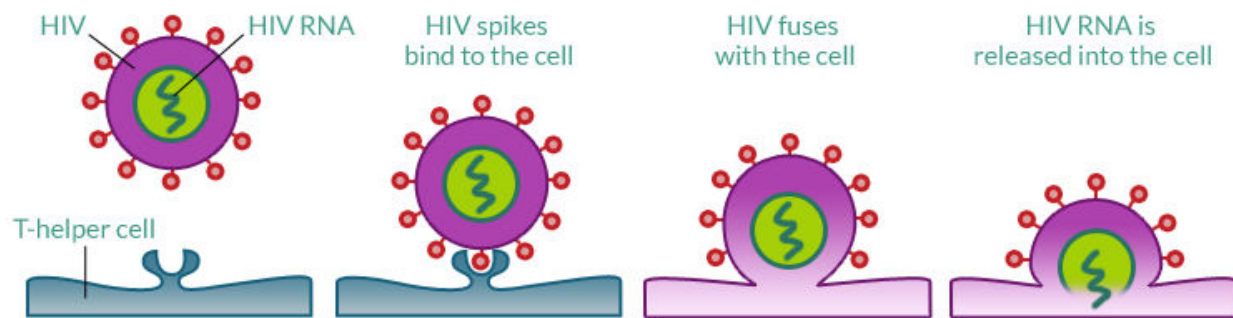


Figure 2.1: HIV virus infects the CD4+ cell (Adapted from Averting HIV and AIDS (2016))

The HIV virus infects the CD4+ cell and the two Ribonucleic Acid (RNA) strands which are synthesized to HIV Deoxyribonucleic Acid (DNA) through the enzyme reverse transcriptase. This process is referred to as the reverse transcription (Averting HIV and AIDS, 2016).

During the reverse transcription process, the reverse transcriptase enzyme converts the RNA template into DNA. This is done and the newly synthesized DNA is integrated into the cell DNA through the use of the integrase protein. This is what determines the lifelong nature of HIV infection. A new HIV virus is then assembled within the cell and exits the cell while not yet fully mature. It takes some essential proteins as it leaves the cell, and matures outside the infected cell. The mature cell is then able to infect other CD4+ cells (Fevrier, et al., 2011; AIDS Gov, 2016).

The available antiretroviral regimens in use target the HIV virus during the initial fusion of the HIV cell, the integration of the HIV DNA into the cell DNA and the conversion of the HIV RNA to DNA. This is done by introducing blockers that prevent the execution of the same processes. In the initial fusion, a blocker to prevent the HIV virus from fusing with the CD4+ cells prevents the introduction of the HIV RNA into the cell. Some antiretroviral medication introduce blockers to prevent the integration of the HIV DNA into the cell DNA. This thus prevents the lifelong infection of HIV (myMVC, 2016).

2.2.2. Antiretroviral Therapy Adherence

Antiretroviral adherence is the strict following of the prescribed antiretroviral medication regimen, by an HIV positive patient. AIDS INFO (2016) states that strict adherence to the antiretroviral therapy is key to sustained HIV suppression, reducing the drug resistance and improving the quality of life, overall health and survival. This is important because lack of adherence to the medication could lead to the patient losing all future treatment options as a result of the drug resistance. However, adherence to therapy is difficult to measure. Chesney (2000) identifies four basic techniques that have been developed to quantify adherence. These techniques are having the patient self-report, patients' reports of missing pills, assays of drug levels, and electronic monitoring systems. These methods have their challenges. In the self-reporting method, the challenge is that the patient may inflate their adherence. This method will also only work in the short term.

Chesney (2000) points out that the challenge to the patients' reporting the missing pills is the discarding of the medication packaging and pill-dumping. This method requires that the patient brings back the remaining pills and the original medication package, which they may have discarded. Some patients will also discard some pills so that they appear to be adhering to the regimen. Nonetheless, this method is said to be almost always reliable. Drug assays are expensive, and can be misleading if a patient who is aware takes medication before visiting the clinic. Electronic monitoring systems on the other hand, assume that single doses are taken whenever the pill bottle is opened. This is misleading especially if multiple doses are removed at once.

Several studies, Wasti, et al., (2012); Beer, et al., (2012), report that a high level of adherence to ART is necessary to maintain viral suppression and maintain optimal clinical outcomes for HIV-positive persons. One study, Wasti, et al., (2012) , goes further and identifies that an adherence level of 95% and above is required to achieve virological success.

2.2.2.1. Factors Affecting ART Adherence

Studies have revealed various reasons for nonadherence, ranging from social to side effects of medication. Perception about ART, religion and rituals, alcohol intake, lack of family support, economic issues, distance, stigma and discrimination, short period of medicine prescription, gender and insufficient pills in the packaged bottles were some of the reasons identified. It is important to note that in that particular study, gender was identified given the socio-cultural and economic limitations that Nepal has placed on the people of the female gender. The short period in which medicine is prescribed, in this case, either one or two months is influenced by the financial aspect of medication.

This therefore means that should a person be located far away from a dispensing area, then they might miss getting refills. Age (younger age) and binge drinking (related to alcohol abuse) were consistently found to have negatively affected adherence to ART regimen. Chesney (2000) however cautions that some, and not all studies have shown an association between nonadherence and youth, female gender, low educational level, or a current or past history of substance abuse.

2.2.2.2. Impact of ART resistance in HIV Positive Patients

As seen in the studies above, there is a clear correlation between the social issues and the adherence levels among HIV-positive patients. This non-adherence has also been established to lead to higher viral loads amongst the patients as shown by (Beer, et al., 2012). This therefore, could lead to ART treatment resistance.

ART treatment resistance can result in reduced survivability and increased mortality, cross-resistance, and transmission of ART-resistant HIV strains. Bertagnolio et al., (2013) states that the survivability of the HIV-positive patient in the future would be reduced by 18% where the ART resistant strains are transmitted. This indicates a high mortality rate that can be purely attributed to the development of ART-resistance.

AIDS INFO (2016) defines cross-resistance as the situation in which resistance to a HIV medicine causes resistance to drugs within the same HIV medicine's class. This class refers to the group within which they fight HIV. This could lead to a person having resistance to a drug which they have never even interacted with before.

In mother-to-child transmission, the strain of the HIV virus that is resistant to the ART regimen can be passed to the unborn child. This means that they will in effect lose all future options in the use of the same line of ART regimen. In cases where there is cross-resistance, the unborn child could also lose the second line regimen if the drug is in the same class (Bertagnolio et al., 2013).

2.2.3. HIV Superinfection and Coinfection

Grant and McConnell (2017) define superinfection as the acquisition of multiple HIV strains from multiple partners, where one of the strains of the virus is obtained after seroconversion. Smith, Richman and Little (2005) expand this definition and define it as the infection by a second strain after the initial infection and immune response to it has already been established.

Coinfection is where an HIV-positive patient acquires another/other strains of the HIV virus from multiple partners during the initial stages of infection i.e. before an immune response to the first strain has been established (Smith et al., 2005). The high amounts of viral load in the blood could possibly lead to superinfection with the HIV virus, although Grant & McConnell (2017) agree that more research as to this relation should be done.

The assumption that once a person who has already been infected with the HIV virus cannot be reinfected leads to risky sexual behaviour amongst the HIV positive patients, which might lead to superinfection. In the reported cases of superinfection, there was reported increase in the viral load levels and decrease in the CD4+ cells, which is similar to that during the primary (initial) infection stage. Table 2.1 shows the published cases of HIV superinfection (Smith, Richman, & Little, 2005).

Table 2.1: Table showing published cases of HIV-1 superinfection (Adapted from (Smith, Richman, & Little, 2005))

Citation	Time to superinfection	HIV subtypes	Risk factor	Accelerated disease progression after superinfection ^a	Drug resistance
Ramos et al. (2002) [2]	<3 months	B after AE	IDU	Yes	Not reported
Ramos et al. (2002) [2]	<11 months	AE after B	IDU	Yes	Not reported
Jost et al. (2002) [3]	<28 months	B after AE	MSM	Yes ^b	Not reported
Yerly et al. (2004) [23]	18–24 months	B after CRF-11	IDU	No ^c	Not reported
Yerly et al. (2004) [23]	~36 months	CRF-11 after B	IDU	Yes	Not reported
Yerly et al. (2004) [23]	45–55 months	CRF-11 after B	IDU	Yes	Not reported
Fang et al. (2004) [46]	Unable to be determined	C after A	WSM	? ^d	Not reported
Manigart et al. (2004) [52]	Unable to be determined	G after AG	WSM	Yes	Not reported
Manigart et al. (2004) [52]	Unable to be determined	CRF-06 after AG	WSM	Yes	Not reported
Altfield et al. (2002) [4]	<32 months	B after B	MSM	Yes ^b	Not reported
Koelsch et al. (2003) [5]	<4 months	B after B	MSM	Yes	DS after DR
Gottlieb et al. (2004) [50]	<15 months	B after B	MSM	Yes	Not reported
Chakraborty et al. (2004) [58]	Not reported	B after B	Not reported	Yes ^b	DS after DR
Brenner et al. (2004) [59]	10 months	B after B	MSM	Yes	DR after DR
Yang et al. (2004) [49]	<5 months	B after B	MSM	Yes	DS after DR
Smith et al. (2004) [57]	<14 months	B after B	MSM	Yes	DR after DS

NOTE. CRF-06, mosaic of clades A, G, K, and J; CRF-11, mosaic of clades A, G, E, and J; DR, drug-resistant strain; DS, drug-sensitive strain; IDU, injection drug use; MSM, men who have sex with men; WSM, women who have sex with men.

^a Those individuals referenced as having disease progression were reported to have had increases in HIV load around the time of suspected superinfection.

^b Individual reported recent treatment interruption, which could explain the change in disease progression.

^c Superinfection was noted to be transient and was detected in only 1 sample.

^d Reported disease progression could have been the result of the natural history of advanced HIV disease.

From the table above, it is clear that there was accelerated/increased disease progression after the superinfection. This disease progression can be measured by taking the CD4+ counts and/or the viral load. It was also clear that there was some drug resistance in some of the studies where there was superinfection.

2.3. Methods for measuring HIV/AIDS Progression

HIV/AIDS progression can be measured in two ways, by measuring the CD4+ count and the viral load levels in the human body.

2.3.1. Viral Load

(Government of Quebec, 2015; Miller, et al., 2002) define viral load as the amount of virus in an organism, given in terms of virus particles per milliliter. This viral load refers to the number of viral copies of the HIV RNA strands per milliliter (mL) of blood. In HIV/AIDS management, this is usually a good indicator as to the immunity status of the body, and the efficacy of the ART regimen being used. Chou et al., (2012) affirms this by stating that this variable (viral load) is useful in the determination of whether a treatment is working on a patient for a given disease.

Wilson, et al. (2008) assert that the people with HIV infection receiving effective antiretroviral therapy i.e. those with undetectable plasma HIV viraemia (that is <40 copies of HIV RNA per mL), who do not have other genital infections cannot transmit HIV through sexual contact. Figure 2.2 shows the relation between the probability of HIV transmission per sexual act, and the amount of viral copies in the blood. It shows an gradual increase in the transmission risk with an increase in the viral load. Should the assertion by Wilson et.al be proved as true, since at the time of the study, medical and biological data did not permit proof of that assertion, predicting the viral load could be instrumental in the attainment of the 90-90-90 treatment objective.

The 90-90-90 treatment target sets the goal of, having 90% of PLWHA knowing their HIV status, 90% of people who know their HIV-positive status accessing treatment and 90% of people on treatment having suppressed viral loads by the year 2020 (United Nation AIDS, 2016; Levi, et al., 2016). The suppressed viral load in this case is having the undetectable viral load levels. This would also be helpful in preventing the spread of HIV amongst serodiscordant couples; where one partner is HIV positive and the other partner is HIV negative. It is however important to note that the absence of HIV transmission by HIV positive patients was observed in the studies where treatment adherence, absence of STBBIs, regular medical follow-up and counselling were met (Government of Quebec, 2015).

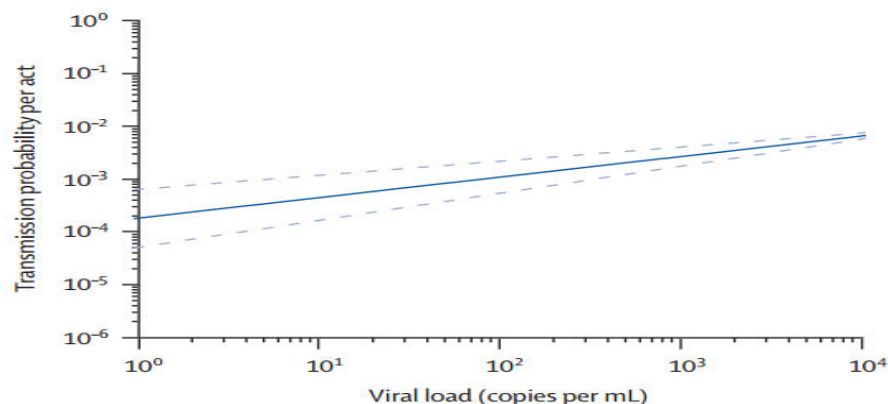


Figure 2.2: Graph showing relation between Transmission Probability of HIV per Sexual Act and the Viral Load (Adapted from Wilson, et al., (2008))

From the figure above, it is clear that the probability of HIV transmission is higher where there is a greater viral load. This highlights the need to lower the viral load levels in order to attain one of the tenets of the 90-90-90 goal and prevent future occurrence of the virus.

Viral load testing is also instrumental in determination of the necessity of a change in the ART regimen for a HIV positive patient. This aid in preventing unnecessary ART regimen changes and can encourage the patient to adhere to the ART regimen given (United Nation AIDS, 2016). Viral load is a more sensitive measure of HIV progression.

2.3.2. Cluster of Differentiation 4

The cluster of differentiation 4 (CD4+) is a glycoprotein found on the surface of the T-cells. The T-cells are a type of white blood cells. The CD4+ count is the count of the number of CD4+ cells per milliliter (mL) of blood. This count is used to estimate the number of white blood cells that are existing per milliliter of blood. The CD4+ cells are used as a conduit by HIV to bind itself to the T-cells, therefore can be used as an indicator of the HIV progression in a HIV positive patient. A higher CD4+ count denotes a healthier patient, but also denotes a high amount of HIV reproduction (Chou et al., 2012).

Immunological markers such as CD4+ counts has been traditionally used in the diagnosis of treatment failure in HIV positive patients. The drawback in using CD4+ counts as a measure of treatment failure in HIV positive patients is that their detection comes after resistance to the drug has already been developed. This therefore necessitates switching of the regimen to a second-line or third-line regimen, which are more expensive (United Nation AIDS, 2016).

2.4. Current HIV/AIDS Prediction Systems

Here, we look at the various existing prediction systems and frameworks, analyzing their strengths and weaknesses.

2.4.1. Prediction of HIV/AIDS Status using Artificial Neural Networks

Artificial Neural Networks (ANNs) are processing devices, that are loosely modeled after the structure of the human brain. The structure is more like a miniature brain in that whereas the brain contains billions of neurons, a large ANN might have hundreds or thousands of processor units. ANNs are organized in layers that are made up of interconnected nodes containing an activation function. Information is provided to the network via the input layer, which

communicate to the hidden layer(s) where the actual processing is carried out. The processing is undertaken via a system of weighted connections which are linked to the output layer. The output layer is responsible for showing the answer (Chebet, et al., 2014) . Figure 2.3 shows the ANN.

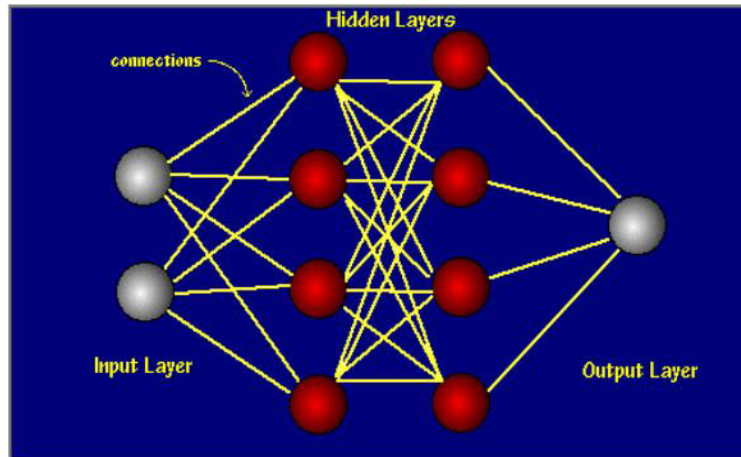


Figure 2.3: Structure of an Artificial Neural Network (Adapted from Chebet, et al., (2014))

2.4.1.1 Neuron

Figure 2.4 shows the mathematical model for a neuron. A typical neuron consists of inputs and weights w_0 to w_i that will be fed into it. A bias weight is also included into the neuron. These inputs with the corresponding weights are then passed through the activation function to produce an output. This output can be used as an input to another neuron.

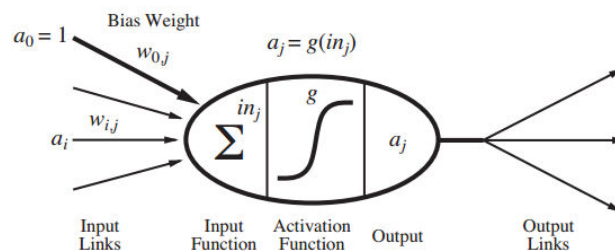


Figure 2.4: Mathematical model of a Neuron (Adapted from (Russell & Norvig, 2010))

2.4.1.2 Activation Function

The activation function's task in a neural network is to transform the input value to the output value in a neuron (Doorn, 2014). Every neuron in the network will require that the input is converted into a value between 0 and 1, and this is done by the use of activation functions. The sigmoid activation function is the most widely used activation function. Equation 2.1 shows the sigmoid function.

Equation 2.1: Sigmoid Equation Used (Adapted from (Doorn, 2014))

$$sig(x) = \frac{1}{1 + e^{-x}}$$

The output of the sigmoid function is shown in Figure 2.5 below. The graph shows the values from zero to a maximum of one. The output, therefore, will fall in between the range 0-1.

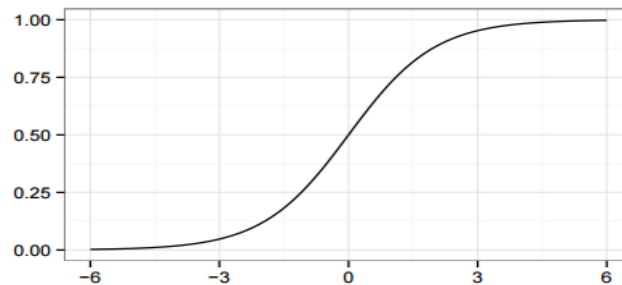


Figure 2.5: The Sigmoid Activation Function Graph (Adapted from (Doorn, 2014))

Emuoyibofarhe, et al., (2016) introduce a confirmatory test for determining the accurate diagnosis of HIV, by use of Artificial Neural Networks. They use the result of the Enzyme Linked Immuno Sorbent Assay (ELISA) test as inputs into the neural network. The ELISA test represents the standard for screening blood samples infected with HIV from none infected samples in Nigeria.

However, the diagnosis from this test is still far from being standard in the Nigerian Health sector. Since the diagnosis is subject to human error. (Emuoyibofarhe, et al., 2016) successfully applied multi-layer Neural Networks to the prediction and hence diagnosis of HIV using the results of the ELISA tests. They used a feed forward back propagation network that was successfully trained with data of various HIV positive sample and HIV Negative samples and

upon completion of the training, new data were sought, interpreted, and supplied as input into the Network.

The diagnosis of the network showed 94% prediction accuracy based on 9 epochs. (Emuoyibofarhe, et al., 2016)'s research successfully established the applicability of Artificial Neural network to the diagnosis of HIV based on the result of the ELISA test and it is expected that it will assist medical practitioners to give a more accurate diagnosis of HIV.

2.4.2. Random Forest Algorithm

Ali, Khan, Ahmed and Maqsood (2012) define random forests as a classifier combination utilizing the L tree-structured base classifiers $\{h(X, \Theta_n), N=1,2,3,\dots,L\}$, where X denotes the input data and $\{\Theta_n\}$ is a family of identical and dependent distributed random vectors. Each decision tree here, is made through the random selection of data, from the entire set of all available data. A random forest can be built through the random sampling of a feature subset, or the random sampling of the data subset used in training of each decision tree. This is also called bagging. (Liaw & Wiener, 2017) explains the concept of bagging as being the case where the succeeding decision trees do not rely on the earlier trees, but each of them is developed independently from a striped down sample of the data set.

Random forest algorithm has been used in the prediction of the probability of the HIV positive patient's viral load falling below 50 copies per ml following the change in therapy. In this study, (Revell, et al., 2012), the input variables used were baseline Viral Load, CD4 count, treatment history and the time to follow-up. This basically forms the Treatment Change Episode (TCE). Figure 2.6 show a sample TCE.

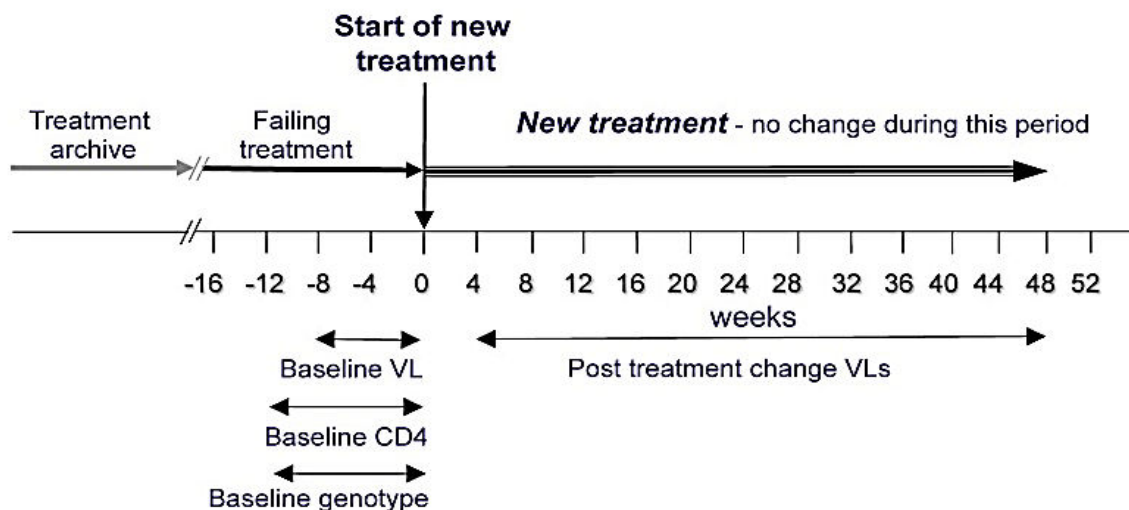


Figure 2.6: Treatment Change Episode (Adapted from (Revell, et al., 2012))

The TCE captures the initial viral load level, CD4+ count and the genotype before the start of the new treatment. During the course of the new treatment, there will be periodic measurement of the viral load levels within a patient to see if the treatment is working. Revell, et al., (2012) used 3188 TCEs and tested with 100 TCEs, and developed two models; one considering the genotype, and the other ignoring the genotype. In their study, they conjecture that the use of the genotype affected the performance of the prediction by a small factor, and that it was insignificant. Their model without genotype resulted in the area under the curve (AUC) for the 0.88 while that with genotype had an AUC result of 0.86. This result was lowered when the TCEs originating from Romania to 0.60. This shows that their model can be used in limited resource setting where the genotype information may not be readily available.

Zoya and Sezerman, (2016), used random forests to test which structural and sequence features can be used to predict drug resistance amongst HIV positive patients with remarkable results. Their random forest classifier showed accuracy measures ranging between 98-99.2%. The study focusses more on the amino acid sequences of the HIV-1 Protease and Reverse transcriptase, and looked at the single and multiple mutations of HIV resistance, while inferring the interactions between them. The feature set considered included hydrophobicity measure, evolutionary conservation, flexibility measure, disordered proteins, and amino acid volume information, for the sequence features, whereas the structural features considered the 2D and 3D representation together with the contact energies.

2.4.3. Support Vector Machines

Cai, Liu, Xu and Zhou (2001) define support vector machines as a kind of learning machine that is based on statistical learning theory. It works by first mapping the input vectors into one feature space, linearly or non-linearly, relevant to the selection of the kernel function. Then withing the feature space from the first step, a hyperplane separating the two classes (or multiclass) is constructed. Figure 2.7 shows an example of the classification using support vector machines.

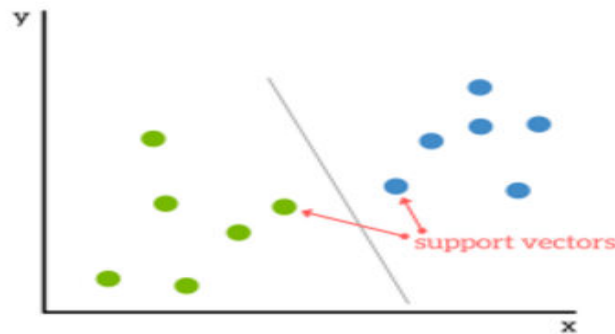


Figure 2.7: Support Vector Machines Example (Adapted from (Aylien, 2016))

Zoya and Sezerman, (2016) also developed a support vector machine classifier in their study. They used the same structural and sequence features in this case and obtained accuracy measures of 95-96%. The parameters used had to be tweaked in order for the study to achieve the highest possible accuracy. Due to the sensitivity of the classifier with regard to class imbalance, they applied sampling techniques to the dataset in order to prevent misclassification.

2.4.4. Linear Regression

Craenenbroek, et al., (2007) applied linear regression on an HIV-1 genotype and phenotype database to predict the resistance of HIV-1 phenotype from the viral genotype. In this study, the phenotypic measurement was estimated as the weighted sum of the effects of the individual mutations. The mutation pairs were included in order to account for the synergistic and antagonistic effects of the various mutations.

Linear regression is used to study the linear relationships that exist between a dependent variable and one or more independent variables. The dependent variable is expected to be continuous while the independent variable could be continuous, binary or categorical in nature (Schneider, Hommel, & Blettner, 2010). In this method, the relationship between the dependent and independent variables under consideration must be linear.

In cases where more than one independent variables linearly relate to giving the output of the dependent variable, then multivariable linear regression is used. The independent variable can then be defined as the linear function of the independent variables X_i . Equation 2.2 shows the liner regression formula. The dependent variable \hat{Y}_t is a function of the relation between the independent variables X_{1t} and X_{kt} where k is the number of variables. β_0 represents the intercept for the regression equation and β_1 to β_k represent the gradients for each of the independent variables.

Equation 2.2: Linear Regression Formula (Adapted from (Freyder, 2014))

$$\hat{Y}_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + \epsilon_t$$

Freyder (2014), states that some assumption have to be made in order to use linear regression. First, the effects of each of the independent variables on the dependent variable should be linear and additive. This means that they should each contribute to the dependent variable in an even manner. Secondly, the independent variables should be independent of each other. The independent variables should not rely on each other nor should they contribute to each other. They should all be contributing to the dependent variable in an equal manner. Thirdly, the errors contained in the model are normally distributed and the variance (homoscedasticity) for the errors are the same for all the independent variables. Figure 2.8: Homoscedasticity (Adapted from (Kumar, 2014)) shows the concept of homoscedasity. The points are aligned along the regression line indicating the constant variability of the values.

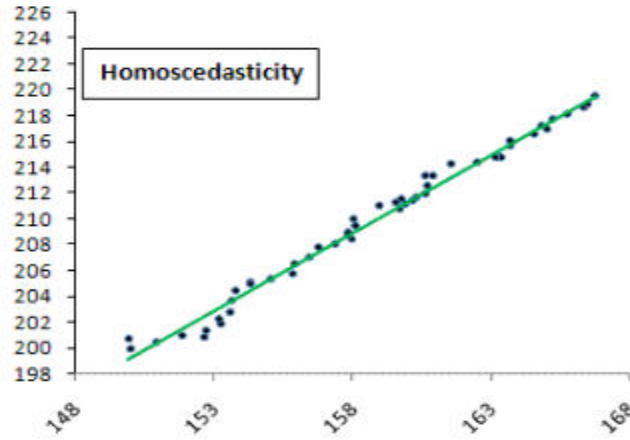


Figure 2.8: Homoscedasticity (Adapted from (Kumar, 2014))

For the viral load prediction, the attributes in question will not follow the assumptions listed above, and thus linear regression cannot be used for this research.

2.4.5. K-Nearest Neighbor

K-Nearest neighbor (KNN) algorithm is a non-parametric method that uses the full training set. It finds the k nearest neighbors to a query point and reports either their class by majority vote or the average of their resistance value (Shen, Yu, Harrison, & Weber, 2016).

The neighbours are found by use of the distance measure, which can either be euclidean, manhattan or minkowski. The value of k is used to determine the number of neighbours which we aim to consider for the case, based on their distance. The algorithm stores all the cases that will be used for the testing, and will be accessed during runtime. Once the test case has been selected, the test case is compared with all the existing cases, calculating the distances between the test case and the cases. The distances obtained are then ranked, and the k smallest distances are selected as the neighbours (Sayad, 2017). To classify, the majority of the classes in the selected neighbors are chosen as the class of the test case.

Figure 2.9 shows how the k-nearest neighbor algorithm. The point of interest is the red dot. The nearest neighbors to the dot are the two plus and the three minus. The 5 nearest neighbors in this case would be taken as a minus, since amongst the 5 neighbors, the majority is a minus.

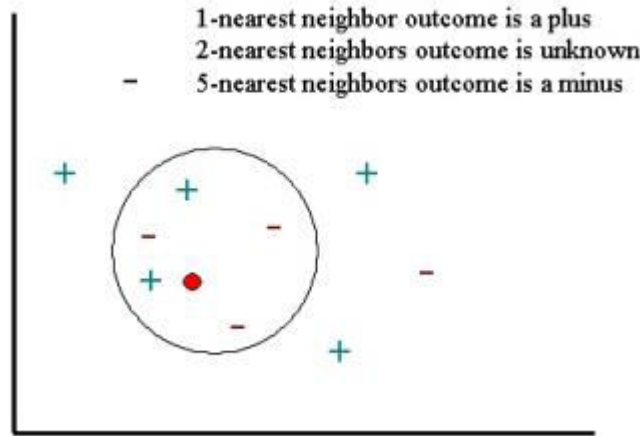


Figure 2.9: Sample of K-Nearest Neighbor Classification (Adapted from (Statistica, 2017))

The distance could be determined by the Euclidean distance whose formula is shown in Equation 2.3 where x is the instances, k is the number of features. We, therefore, sum the distance for each of the features, for the first instance x_i and the test instance x_j . This absolute value is obtained by squaring the difference.

Equation 2.3: Euclidean Distance (Adapted from (Kataria & Singh, 2013))

$$D(x_i, x_j) = \sqrt{\sum_{k=1}^n ((x_{ik} - x_{jk}))^2}$$

In the study by Shen, et al., (2016), the k nearest neighbor algorithm was used. The k value was set to 6, with 210 dimension vectors for the training sample. The use of KNN algorithm makes the training of the model faster, but will utilize the complete training data in the prediction stage since the reporting of the result is done based on the training data. This makes the validation of the model very slow especially when working with very large number of cases. In this study, they were able to obtain R^2 values ranging from 0.719 to 0.928 across 5 –folds. The closer the R^2 value is to 1, the better the prediction accuracy of the model.

2.5. Conceptual Model

The data to used in the system was obtained from the publicly available HIV patient databases. This data was cleaned for use in the training of the ANN system. This means that only specific data attributes were extracted, including the phenotype, the viral load (current), past viral loads, past and present CD4+ counts for the patient.

This data was then combined together based on the patient serial number and the specific dates in which the data for CD4+ counts and the viral load levels were measured. The ART history of the patient was also considered.

The cleaned data was then categorized into two; one for training and one for testing. The drug resistance information for the specific ART regimen phenotypes was used together with the cleaned data to train the Artificial Neural Network, that first begins with arbitrary weights. Whenever a prediction is required, the patient information mimicking the one used in the training, is fed into the system and the system will pull drug resistance data from the publicly available databases. The system then provides a prediction of the estimated viral load level for the patient over a given period of time. Figure 2.10 shows the conceptual working of the system.

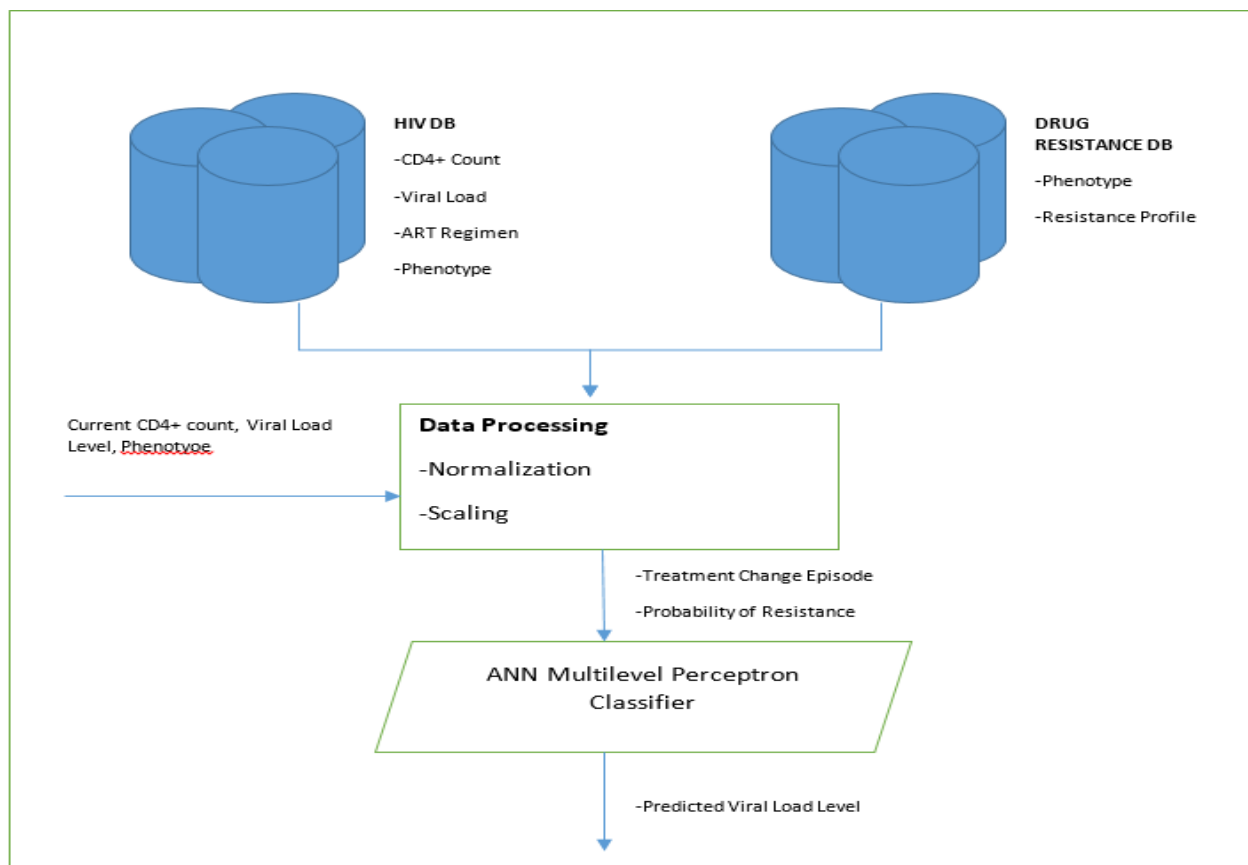


Figure 2.10: Conceptual Framework for the system

Chapter 3 : Research Methodology

3.1. Introduction

In this chapter, the research methodology used is outlined. The research design chosen, the population selected and the sampling method used will also be analyzed.

3.2. Research Design

Wanjugu (2015) defines research design as the structure that the research process follows in terms of collection, measurement, and analysis of data, with the aim of obtaining answers to the research questions. The author also states that one of the aims of research design is to combine relevance to the purpose of research with the economy in procedure.

Research design can also be described as a general plan on how the various aspects of the research will be organized together in an orderly and meaningful manner, in order to ensure that the research problem is addressed adequately, by establishing how the data will be collected, measured and analyzed (Alick, 2016).

This research is an applied research. This is because it aims to provide information that can be used and applied in an effort to help other people understand and control their environment. Applied research is more prescriptive in nature and seeks to offer potential solutions to problems (Center for Innovation in Research and Teaching, 2016).

3.3. 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) Validation of the model

3.3.1. Obtaining Data

The data that was used in the building of the neural network regressor was obtained from the publicly available HIV Stanford database. This data was in form of XML files, which were downloaded directly as a zip file. The data consisted of 1518 individual XML files, and they contained the individual TCEs.

3.3.2. Data Pre-processing

The data obtained for this study contained duplicates and gaps. The data had to be pre-processed. The TCE was first extracted from the XML file, and stored in a relational database. The extraction was done via PHP code that was developed. The duplicated were then eliminated and the complete data was then used to generate CSV files that could then be used for the training, testing and validation of the model.

3.3.3. Development of the Model

The model was developed using the sklearn library, which is an open-source Python library. The evaluation method was artificial neural network regressor with multiple layers.

3.3.4. Validation of the Model

The model generated was validated by use of the coefficient of determination (R^2) measure and the mean square error (MSE).

3.4. System Development Methodology

The study utilized Data-driven Modelling (DDM) in the development of the system. This technique is useful in computational intelligence, machine learning and data mining. Computational intelligence includes neural networks (which this study relies on), fuzzy systems and evolutionary computing (Solomatine, See, & Abrahart, 2008). The Data-driven modelling methodology relies on analysis of the data available about a system, and particularly finding connections that exist between the input, internal and output variables. This modelling is done without any explicit knowledge of the physical behavior of the system. Figure 3.1 shows the Data-driven Modelling methodology workflow.

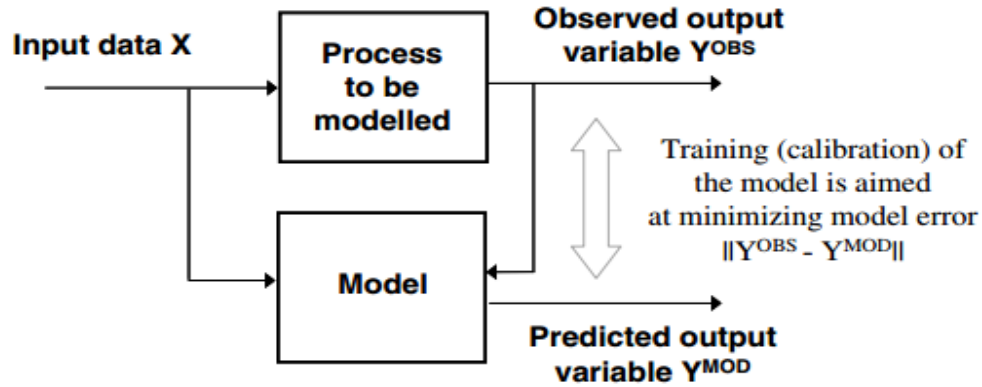


Figure 3.1: Data-driven Modelling (Adapted from (Solomatine, See, & Abrahart, 2008))

3.4.1. Application of the Data-driven Modelling Methodology

The input variables were obtained from an online source, and was categorized into training, test and validation data. The data was then processed by use of a machine learning algorithm; the artificial neural network. The data used in this model was representative of the behaviour found in the system. The model that was developed can be used to augment the operation of the Comprehensive Care Centers.

The training data was used to train the model in a bid to have the model converge by minimizing the model error. The test data was used in performance evaluation of the model. This data was not used in to modify the model in any way, but to indicate the errors that may be present in the developed model.

3.4.2. Techniques Used

Some of the techniques used in this approach as stated by Solomatine, See and Abrahart (2008) are neural networks, fuzzy rule-based systems, genetic algorithms for model optimization, support vector machines, and chaos theory. This study used neural networks with backpropagation and multiple-layers.

3.4.3. Model Testing and Deployment

The system testing was undertaken by use of test data. The test data was set from the set of data obtained online. The level of efficiency obtained by the system was assessed from the test data and the validation data.

3.5. Research Quality

Research Information Network, (2010) define quality research as research that has intellectual vigor, accurate recording and honest reporting of work, and integrity in the recognition of the work of others. This research followed quality standards by ensuring that all work that was used was properly cited and that the literature reviewed was obtained from reputable sources. The data collected and used was presented as is, except in instances of duplication of data entries and missing data, that would have affected the quality of model used. The results that were obtained were assessed based on R^2 (coefficient and determination) and Mean Squared Error (MSE).

The Mean Squared Error of prediction is the average squared difference between independent observations and the predictions from the model or equation . It incorporates both the variance of prediction and the square of the bias of the prediction. The equation used to compute this is shown in Equation 3.1. The n is the number of instances, $\hat{Y}_{\text{pred}(i)}$ is the prediction of the observation (i) and Y_i is the expected value.

Equation 3.1: Formula for Computing Mean Squared Error (Adapted from Rawlings, Pantula, & Dickey, (1998))

$$\text{MSE} = \sum_{i=1}^n (Y_i - \hat{Y}_{\text{pred}_{i(i)}})^2.$$

Coefficient of determination (R^2) is the proportion of the corrected sum of squares of Y that can be attributed from the information obtained from the independent variable (s). It usually ranges from zero to one and is the square of the product moment relation between Y_i and $\hat{Y}_{\text{pred}(i)}$.Equation 3.2 shows the formula for computing coefficient of determination. In the formula, SSR is the regression sum of squares, SSE is the error sum of squares and SSTO is the total sum of squares.

Equation 3.2: Coefficient of Determination Formula (Adapted from PennState Eberly College of Science, (2017))

$$r^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

3.6. Ethical Considerations

In the study, the researcher utilized publicly available online data, only after obtaining consent from the caretakers of the system. In addition, the researcher disclosed the use of the data as pertaining to the scope of the study and was subsequently granted access to the Sierra Web Service. Given that the data is based on patients' treatment history, the researcher ensured that the data given only contained aliases that could not be linked to any particular patient, and that the data was only used for the purpose of this research. Attribution for the use of data was given to the owners of the data. The data collected by the researcher was used in the format without modification in order to ensure its integrity.

Chapter 4 : System Design and Architecture

4.1. Introduction

In this section, we will outline the design architecture of the developed HIV/AIDS viral load prediction system. This architecture will follow from the conceptual model shown in Figure 2.10. This section shall cover the interaction between the users and the developed system, the components of the developed system and the interaction between the various components of the developed system. This was modelled by use of class diagrams, use case diagrams, data flow diagrams and system sequence diagrams.

4.2. Requirements Analysis

The requirements that were gathered for this study can be categorized into three sections; the functional requirements, non-functional requirements and usability requirements.

4.2.1. Functional Requirements

- i. The system should allow the user to input data as a CSV file, from the form or as an XML file. Any other formats of data input should be rejected.
- ii. The system should extract the viral load (RNA), CD4+ counts, the drug information and the duration from the uploaded data.
- iii. The system should predict the level of viral load (RNA) by use of the multilevel neural network.
- iv. The predicted viral load level should be valid based on the input given from the user.
- v. The system should provide a rating of the level based on the CDC guidelines.

4.2.2. Usability Requirements

The system is intended to be used at the hospital level. The main users will be the clinicians located at the hospital. The system should, therefore, be simple and straightforward to enhance user acceptance. It should also be accurate as its predictions might directly affect the patients' lives.

4.2.3. Reliability Requirements

- i. The system should always be able to interface with the existing drug resistance database.
- ii. The system should be able to extract the treatment change episode from the input data either in CSV, XML or from the form.
- iii. The administrator should be able to correctly restore the system to a functioning state in the event of a failure.

4.2.4. Supportability Requirements

The system should be accessible as a web application. The system should be accessible across all the major web browsers, and across all the major desktop platforms.

4.3. System Architecture

The architecture for the system is as shown in Figure 4.1. The main input to the system will be the treatment change episode (TCE) of the patient. This is data showing the previous, baseline and future levels of RNA, mutations of the virus across time and the regimens used during the period. These parameters will be standardized before being fed into the neural network. The data is split into the training and test data, which will be used in the training phase. Once the training has been done, the subsequent data input will utilize the trained model in the prediction of the viral load level. The viral load level given as output shall then be compared to the set CDC guideline and the system will give the recommendation based on the guidelines.

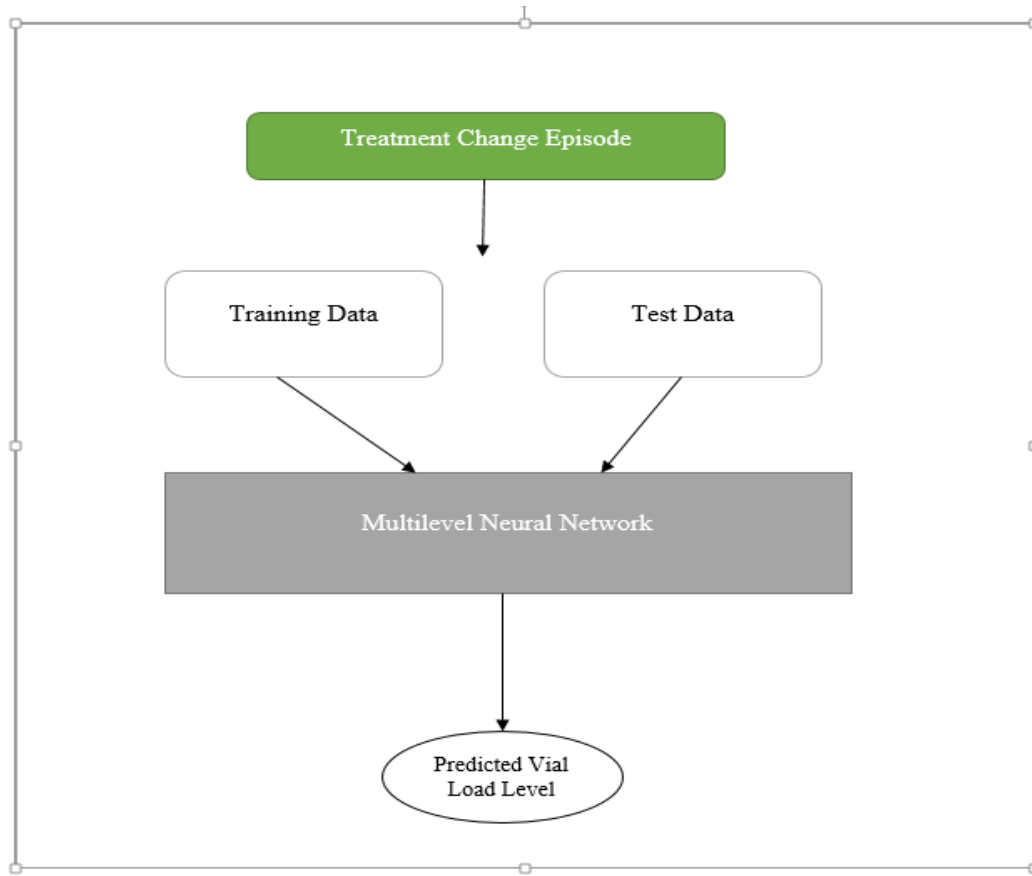


Figure 4.1: System Architecture

4.4. Use Case Diagram

The actors in this system are the administrator, the clinician, and the sierra drug resistance database. The administrator performs data pre-processing and model training. The data pre-processing includes the standardization of the data to prepare it for use in the training of the model. It also includes feature extraction where the features with the most information gain are selected. In addition to this, the percentage of drug resistance is obtained from the sierra system (drug resistance database) which is also standardized. Under model training, the administrator utilizes the data obtained from the data pre-processing stage to train the neural network. The sierra system is used to perform the computation of the drug resistance percentage, which is used as an input in the data pre-processing and prediction stages.

The users who use the system have been identified and depicted in Figure 4.2. The clinician interacts with the system by providing the data that will be used in the prediction. S/he will be performing prediction of the viral load level. This step utilizes the functionality of the model that has been saved from the model training, and the resistance percentage obtained from the sierra system. This would, thus, output the predicted viral load level and the recommendation. Table 4.1, Table 4.2 and Table 4.3 give more details on the main use cases in the system and their success scenarios.

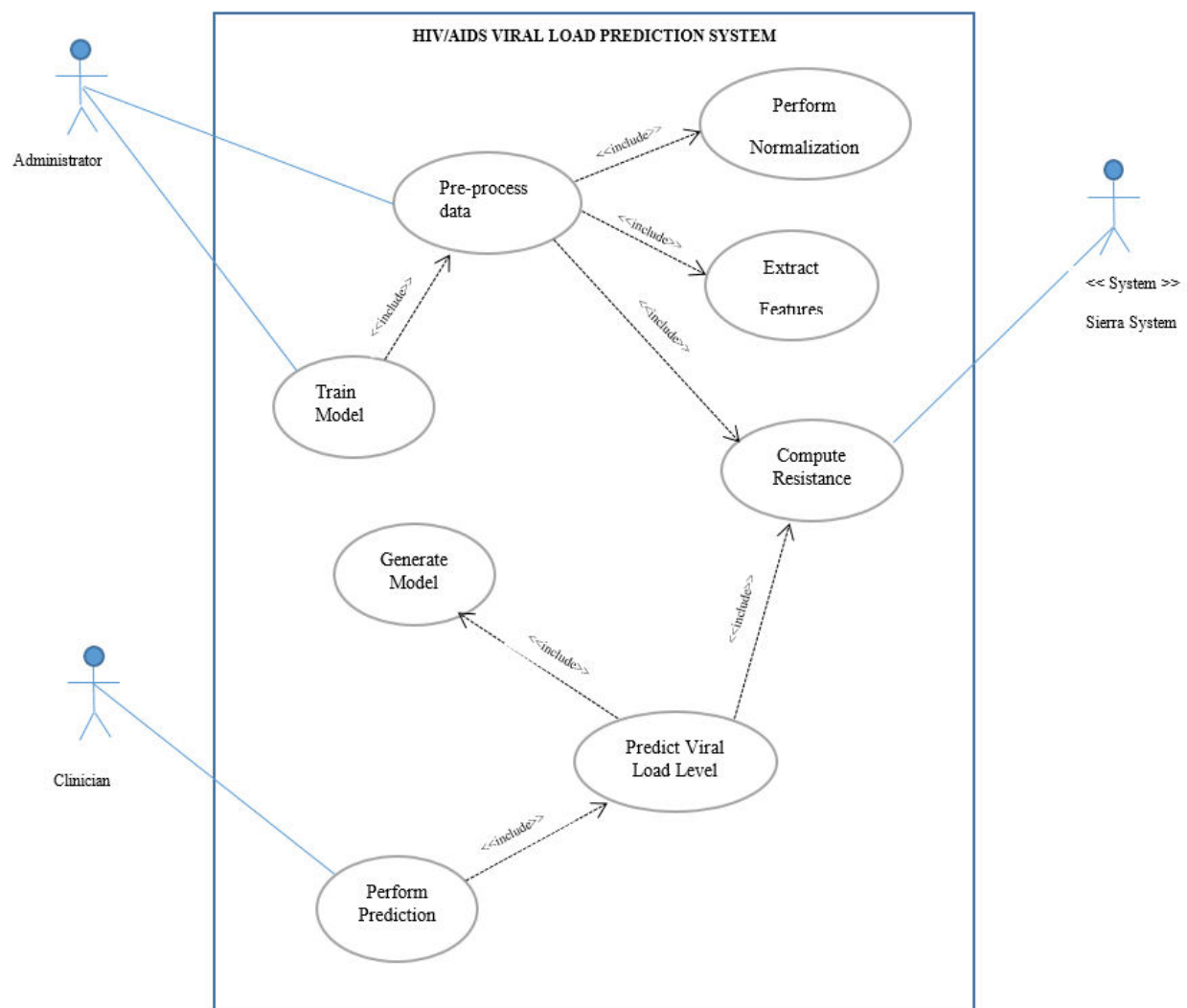


Figure 4.2: System Use Case Diagram

Table 4.1: Data Pre-processing, Standardization and Feature Extraction

Use Case: Data Pre-processing, Standardization, and Feature Extraction	
Primary Actors: Administrator	
Precondition: Feature Selection System Available, Standardization Function	
Post-condition: Treatment Change Episode obtained should have all required features	
Main Success Scenarios	
Actor Intention	System Responsibility
1. Administrator inputs dataset	
2. Administrator selects feature extraction parameters	
	3. System runs feature extraction
	4. System saves extracted features

Table 4.2: Model Training and Testing

Use Case: Model Training and Testing	
Primary Actors: Administrator	
Precondition: Pre-processed Treatment Change Episode	
Post-condition: A multilevel neural network model for viral load level prediction	
Main Success Scenarios	
Actor Intention	System Responsibility
1. Administrator selects the neural network features	
2. Administrator selects the number of epochs to run	
3. Administrator enters the target output for the data	
	4. System trains the model based on the number of epochs
	5. System generates neural network model and saves it as pickle file
	6. System tests generated model using test data
	7. System generates prediction analysis

Table 4.3: Obtaining the Predicted Viral Load Level

Use Case: Get Viral Load Level	
Primary Actors: Clinician	
Precondition: A multilevel neural network model for viral load level prediction, Valid Treatment Change Episode (TCE), Drug Resistance Percentage	
Post-condition: Predicted Viral Load Level and Recommendation	
Main Success Scenarios	
Actor Intention	System Responsibility
1. Clinician Enters the TCE	
	2. Perform Prediction from data
	3. Return predicted level, recommendation, and accuracy
4. Clinician views predicted level, recommendation, and accuracy	

4.5. System Sequence Diagram

Figure 4.3 shows the sequence diagram for the system. The Clinician inserts data, which is basically a treatment change episode either in a CSV, XML or form. The uploaded data undergoes feature extraction to get the features with the most information gain. This data is split into the test and training cases and fed into the multilevel neural network to create the model. The system then performs the prediction for the level and then compares the results to the guidelines from CDC. Once this information has been obtained, the result is given back to the clinician as a recommendation with the predicted level and the percentage of accuracy.

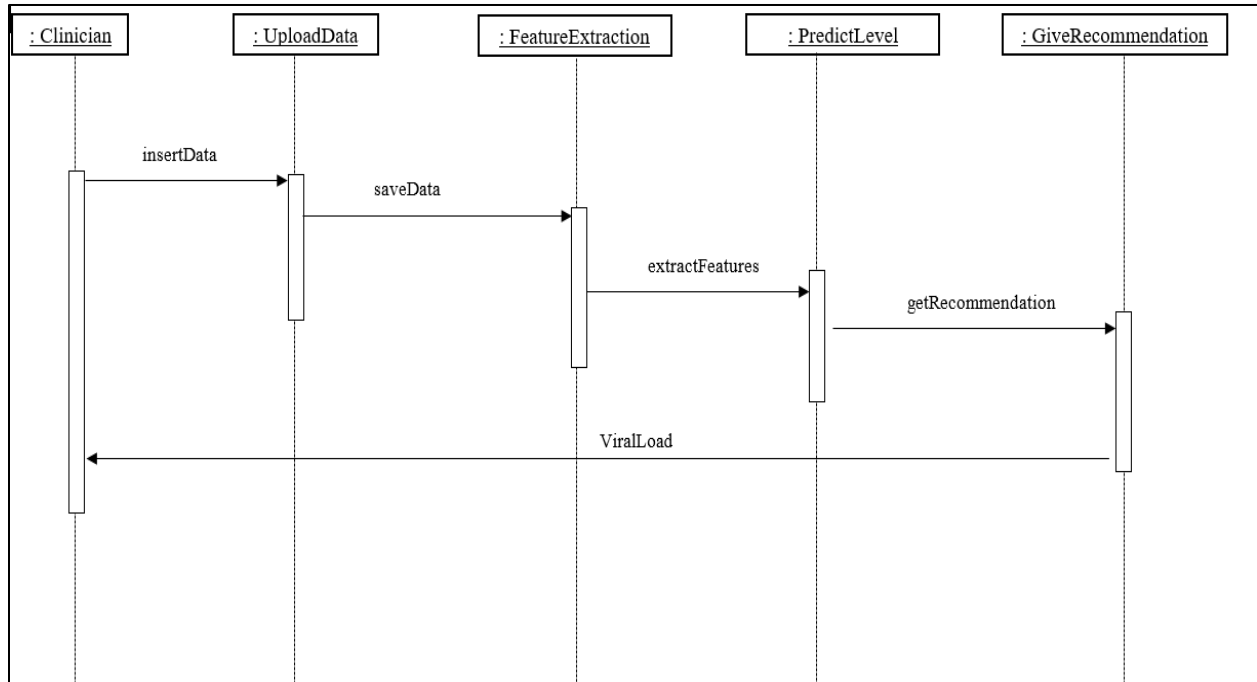


Figure 4.3: System Sequence Diagram

4.6. Flow Chart

The process flow for the training of the model is as shown in Figure 4.4. The treatment change episode (TCE) is entered and data pre-processing is undertaken on it. The data in it is tested with a generated model based on the number of epochs. If the model is not considered optimal (lack of convergence) then it checks if the stopping condition has been met. This stopping condition is basically the number of epochs that has to be run while training the system. By this time, the model has not been saved. If the specified number of epochs has been met, then the system stops.

If the specified number of epochs has not been reached, a new model is generated by readjusting the weights and training the model. The trained model is then reevaluated for its optimality, and there is convergence, the model is saved as a pickle file. The system then checks to see if the stopping condition has been met. This process is repeated until the stopping condition is met. The final saved model is what will be used for the system during runtime.

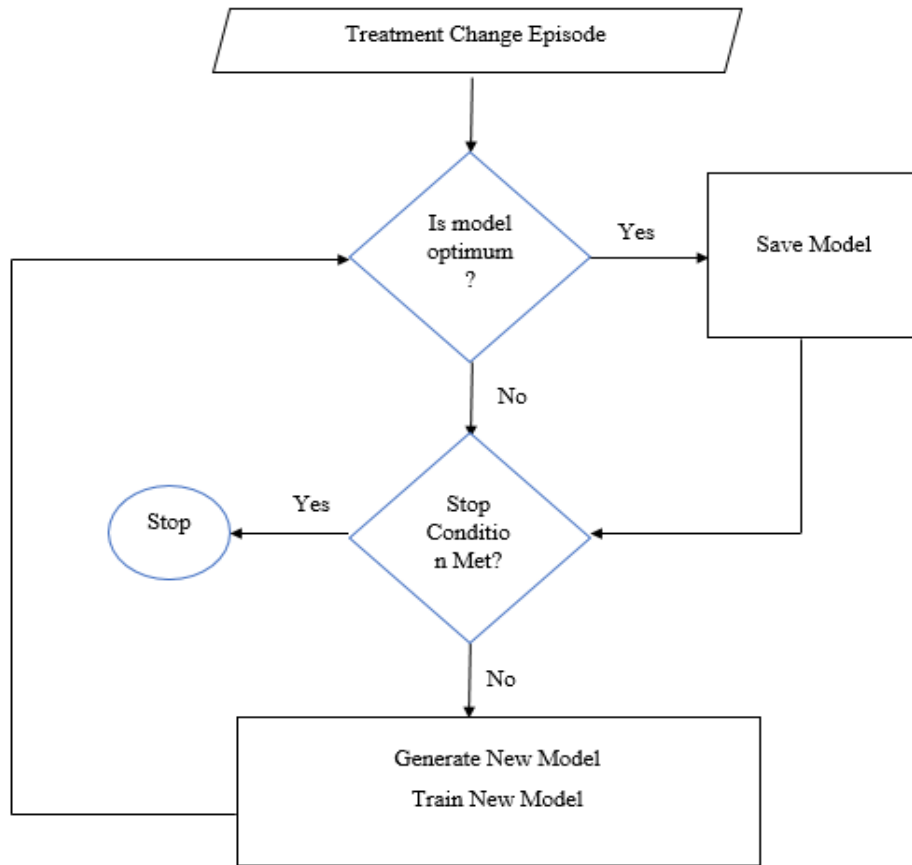


Figure 4.4: Flow Chart for the System

4.7. Database Schema

The system has a database as defined in Figure 4.5. The database contains a total of 11 normalized tables. The types table contains the duration within which the treatment, viral load (RNA) or CD4+ counts were taken, and is either Past, Baseline (beginning of testing) and Follow-up. These types are also contained in the isolates table.

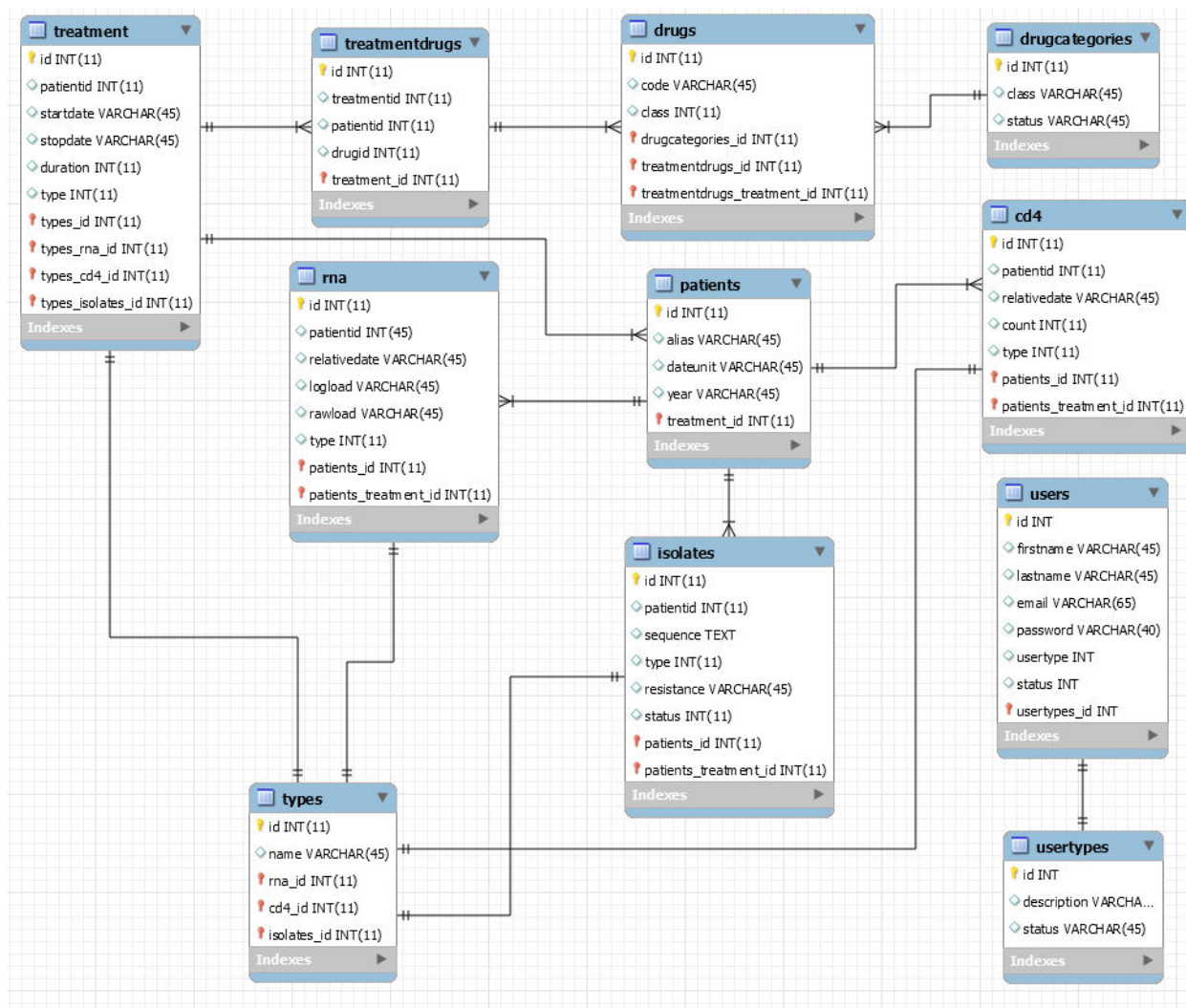


Figure 4.5: Database Schema of the System

Chapter 5 : Implementation and Testing

5.1. Introduction

The prototype was developed through the use of neural network multi-layer perceptron regressor. The model was implemented as a standalone system, consisting of a PHP web application, a python backend API and a neural network developed in python. The data used in the system was stored in MySQL database.

The data was entered through the web interface, and was saved in the database. This data consists of fasta files, the nucleotide sequences or the treatment profile of the new patient. The nucleotide sequence and the fasta file was then used to get the drug specific resistance level from the system. The treatment profile (TCE) was used in the prediction of their viral load level. The multi-layer neural network regressor, trained by use of backpropagation, was used in the prediction of the viral load levels for the HIV/AIDS positive patient. The neural network was trained by use of training set of data, and the result of the prediction are displayed to the user on the web interface with the recommendation.

5.2. Model Components

5.2.1. System Components

The web system developed contained two sections of data input. The first section was the drug resistance interface which is a public interface that allows anyone with the knowledge of the website to easily obtain resistance reports of the HIV/AIDS regimens. It contains two inputs sections; the nucleotide and the fasta file. The fasta file is a text file that contains the nucleotide sequences in a list, separated by a character, in this case, it is '>'. This interface is shown in Figure 5.1 below. The output of the analysis is displayed in a pop-up interface.

The private section can be accessed by logging into the system with valid credentials, either as an administrator or as a clinician. The administrator is able to undertake administrative tasks such as addition, editing and deletion of users. In addition, s/he is able to compute the drug resistance for a set of inputs in the system based on the nucleotide sequence already existing. The clinician is able to enter a number of parameters, which includes the past CD4+ counts, the past viral load levels, the drug regimens that the patient has been on in the past, and the nucleotide sequence that was there in the past. The clinician is also be able to add the baseline (current during prediction) levels of the viral load, and the CD4+ counts, and the nucleotide if present.

The baseline year is the current year if none is entered/ selected, and the unit of time measure used (i.e. either weeks, days, months) has to be selected. The duration under consideration for prediction has to be selected too. If none is selected, the default prediction period of 8 weeks is used. This input form is shown in Figure 5.2.

The image shows the Medipredict website interface. At the top, there is a header with the Medipredict logo and the text "FREE HIV/AIDS VIRAL LOAD PREDICTION". Below this, a sub-header says "Enter your Nucleotide Sequence or Fasta File to Check Resistance". A "Check Resistance" button is visible. The main content area is divided into two sections: "NUCLEOTIDE SEQUENCE" and "FASTA FILE". The "NUCLEOTIDE SEQUENCE" section has a text input field labeled "Nucleotide Sequence" and an "Analyze Sequence" button. The "FASTA FILE" section has a "Choose File" button and an "Analyze Sequences" button. Below the "FASTA FILE" section, there is a note: "A fasta file is a file containing nucleotide sequences with the sequences separated by a > symbol".

Figure 5.1: Public Input Section, Showing Nucleotide Input field, and Fasta file Upload Section

Add Treatment Change Episode

Patient Details

Patient Alias #
12345

Time Unit of Measure
Weeks

Past Treatment Date
+/- no of weeks

Past RNA Level
100000

Past RNA Log Level
5

Baseline Treatment Date
+/- no of weeks

Baseline RNA Level
100000

Baseline RNA Log Level
5

Nucleotide Sequence / Isolate
CGGGTC....

Treatment Past

Drugs
☐ AZT
☐ SQV
☐ TDF
☐ LPV

Start Date
5

Stop Date
5

Treatment Baseline

Drugs
☐ AZT
☐ SQV
☐ TDF
☐ LPV

Start Date
5

Stop Date
5

Cancel
Submit

Figure 5.2: Add New Treatment Change Profile Interface

5.2.2. Neural Network Components

The multilevel perceptron neural network includes several components as listed below

5.2.2.1 Input Layer

The input layer is the first layer in the neural network. It contains neurons that correspond to the number of attributes that are used by the neural network to predict the values. A bias node is also included in this layer to ensure that the neural network will be able to fit the data. The model used consists of 8 nodes. The inputs to the system are the past CD4+ count, the baseline CD4+ count, the past viral load log level, the baseline viral load log level, past average drug resistance, the baseline average drug resistance, the duration between the past treatment start date and the baseline treatment stop date, and the duration within which the viral load is to be predicted.

5.2.2.2 Hidden Layer

The hidden layer is a series of neural network layers of nodes which serve to enhance the prediction accuracy of the network. This is done by having a series of nodes with activation functions. These neurons would also serve to prevent overfitting of data to the model, and under fitting of data to the model. Karsoliya, (2012) stipulates that the number of hidden layers should be less than twice the number of neurons in the input layer. This prevents underfitting and overfitting. Thus the formula chosen for this model was $2n-1$. The model developed used 15 hidden layers, given that it is a prediction system with a sizeable number of variables.

5.2.2.3 Output Layer

This is the final layer of the neural network, and it serves to produce the output from the neural network. The model selected was a neural network regressor which had multiple output values from one output node.

5.3. Model Implementation

5.3.1. Data Input

The data to be input is in form of either XML, CSV, fasta file or text. The clinician uploads the file on the system, and runs an analysis of the same. The fasta file contains a list of nucleotide sequences as shown in Appendix A. The XML file used needs to have at least one past viral load (RNA) level count and its log, the past CD4+ count, the baseline (current) RNA and CD4+ counts, the patient alias, the unit of measure and the regimens used in the management of the virus. A sample of the XML can be seen in Appendix B.

The data is extracted from the various input fields and inserted into the MySQL database based on the patient alias as the linking attribute. If the logarithmic value for the viral load level is not entered, the system computes its logarithmic equivalent using Equation 5.1. This value is stored in the database.

Equation 5.1: Logarithmic Formulae for Finding the logload value

$$\text{logload} = \log_{10} (\text{Viral Load Count})$$

5.3.2. Drug Resistance Computation

Each of the regimen that the HIV positive patient is on consists of a cocktail of a number of drugs. These drugs have a specified resistance that can be estimated. The Stanford HIV drug resistance database is able to provide the predicted resistance values. The HIV/AIDS viral load prediction system utilizes this functionality to obtain the resistance. This is done through the use of the Sierra web service. The resistance is based on the specific nucleotide and the approximated drug resistance as identified from the mutations in the nucleotide. The predicted drug resistance contains a summary of all regimen drugs with their associated resistance. However, a patient may not be on all the drugs and as such the resistance would be a function of the average resistance based on the number of drugs. The function is defined in Equation 5.2.

Equation 5.2: Determining Drug Resistance for a Patient

$$\frac{\sum_1^n X}{n}$$

where: n is the number of drugs the patient is on
X is the drugs the patient is on

5.3.3. Scaling Data

The data obtained from the clinician needs to be fed into the neural network during the training, testing, validation and model use phases. This data needs to fit within the range [0, 1] for it to be valid in its use. Given the variance of the viral load logload and the actual values, the need for normalizing the data is apparent. This is done based on the formula on Equation 5.3. The standard scaler removes the mean of the data and scaled the data to unit variance. x is the original vector, μ is the mean of the vectors, and δ is the standard deviation. This is useful as it helps improve the prediction performance of the model.

Equation 5.3: Standard Scaler Formula

$$z = \frac{x - \mu}{\sigma}$$

The tables Table 5.1 and Table 5.2 show sample data that was used in the training of the system. The variation between the values for the CD4+ count is very large, but once standardized, this variation decreases largely.

Table 5.1: Table Showing Unstandardized Data

Past CD4 Count	Past RNA Log
80	4.1
513	4.8
284	2.8
290	2.6

Table 5.2: Table Showing Standardized Data

Past CD4 Count	Past RNA Log
0.87974797	1.131719603
-0.930151099	-0.438385407
0.492899314	-1.910358854
-0.718305406	-1.321569475

5.3.4. Software Flow

A web application was developed that allowed the clinician to add the treatment change episode for a particular patient. The clinician has to log in first, and add the treatment profile (TCE) to the system and save it. The TCE contains information on the past treatment, the present treatment, viral load levels for the past and present, the CD4+ counts for the past and the present, the nucleotide sequence(s) and the duration for which the prediction is expected to cover. The default duration is 8 weeks.

Once the profile has been saved, the clinician is taken to another interface from where they can perform the prediction of the viral load. Here the system queries for the drug resistance from the Stanford HIV database through the Sierra web system via an API call.

The resistance to the specific drugs is compared with the drugs that the user is on and the average stored. The resistance is a predicted value hence the average is taken to simplify the viral load prediction process and ensure that the number of attributes remains constant. The clinician is then given an interface with the final predicted value for the viral load level, and the recommended course of action. The environment used in the development of the system are as follows:

- i. Windows 10 Home Edition
- ii. Sublime Text 3
- iii. Python 2.7
- iv. PHP 5.6
- v. Numpy version 1.11.3
- vi. Scikit learn version 0.18.1
- vii. JetBrains Pycharm Student Edition

5.4. Model Architecture

The model that has been developed in this research is as depicted in Figure 5.3. The data once uploaded will be saved into a local database. The data is then completed by obtaining the resistance for each patient based on their nucleotide sequence and the drug resistance. This data is then used in the prediction. The data is preprocessed to extract some specific attributes and then scaled to fit within a given range. This is done by use of the standard scaler function. The scaled data is then passed through the neural network regression model which outputs a value. The value is the predicted viral load level for that particular patient.

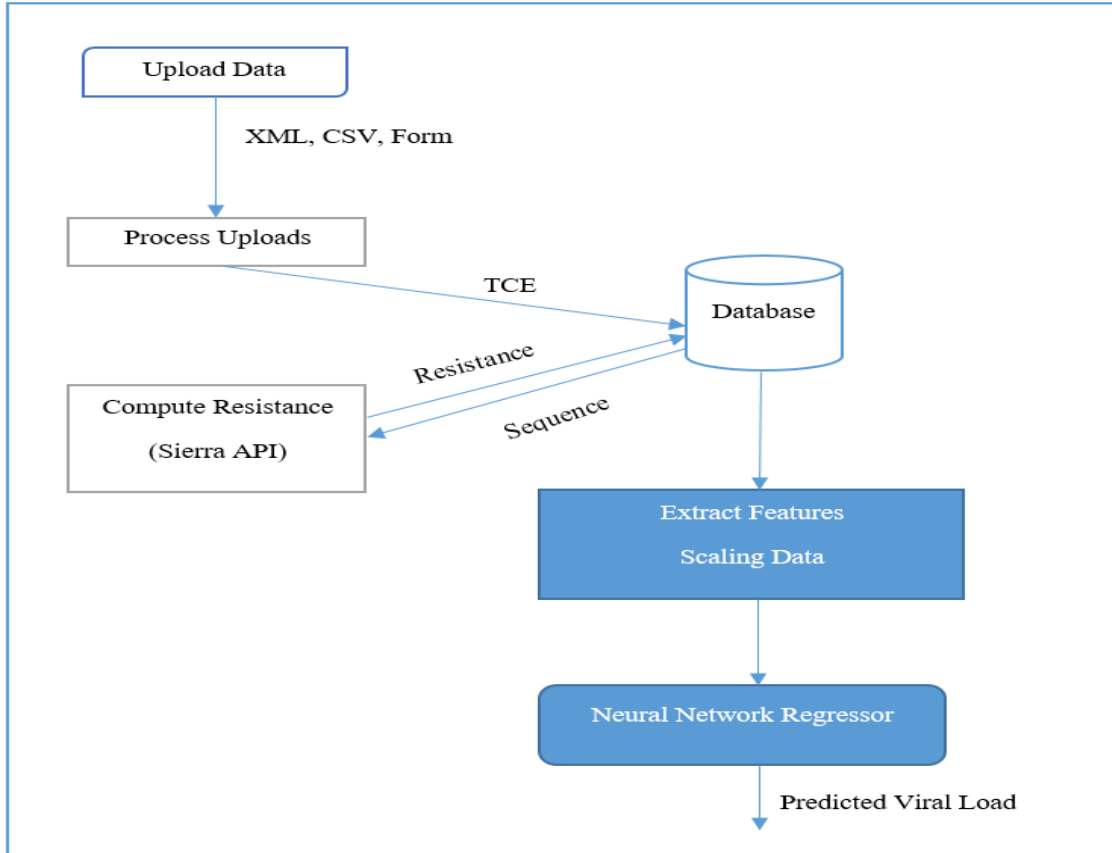


Figure 5.3: Model Architecture

5.5. Model Validation and Testing

The University of Edinburgh, (2017) defines model validation as the task of showing that the model created is representative of the actual system to a reasonable degree. This means that the model produces results similar to what the actual system would, with enough confidence to allow analysis to be performed on the results produced.

NIST/SEMATECH, (2017) asserts that model validation is one of the most important aspects of building a model. This importance is underscored by the assertion that this is also one of the most overlooked aspects in the entire model generation process. NIST/SEMATECH, (2017) mentions the use of the R^2 statistic as a commonly used measure in the analysis of models. This performance measure was used in the analysis of the developed model as is discussed in chapter 6. The functional requirements, usability requirements, and supportability requirements are what were used to look at the validity of the developed model.

5.5.1. Functional Requirements

Table 5.3 shows the functional requirements for the system. The elements assessed were valid predictions, extraction of data and allowing multiple input forms.

Table 5.3: Functional Requirements

Requirements	Priority	Result
Allow user to input data as CSV, form or XML file	High	The system allowed the data to be input from the specified sources only, and a sample CSV file was provided in order to ensure the data was properly captured. A valid XML format was also provided.
System should be able to extract the data from the input sources	High	The system was able to get the various data from the specified sources and store the value
Predicted viral load level should be valid based on the user input	High	The system was able to predict the viral load levels from the given user input to a given degree of confidence
The system should provide a rating based on the Centers for Disease Control (CDC) guidelines	Medium	The system compared the predicted viral load levels to the guidelines given by the CDC and give the recommendation

5.5.2. Usability Requirements

Table 5.4 shows the usability requirements for the system. The simplicity of use, access to the system and speed of use of the system were analyzed.

Table 5.4: Usability Requirements

Requirements	Priority	Result
Simplicity of Use	High	The system had few interfaces which were straightforward to ensure that the users were using the system optimally
Access to the system	Medium	<p>The system contained some aspects that could be accessed without the need to log in. This would allow the users to perform routine tasks with ease.</p> <p>The system was also web based thus can be easily scaled and accessed widely.</p> <p>The system also allowed the user to perform multiple analysis at a time, therefore reducing the time the user while using the system</p>
Speed of the system	High	The speed of the system was wanting but could still be used effectively. This was as a result of the request to an external API for drug resistance. The performance varied based on the availability of stable internet

5.5.3. Reliability Requirements

Table 5.5 shows the reliability requirements for the system. The ability of the system to interface with other systems, restoration of the system and extraction of the TCE were assessed.

Table 5.5: Reliability Requirements

Requirements	Priority	Result
Ability to interface with the drug resistance database	High	The system is able to interface appropriately with the drug resistance database via the Sierra web system API
Extracting the Treatment Change episode from input data	High	The system was able to successfully extract the treatment change episode from the input methods that followed the correct procedure
Administrator restoring system to functioning state in case of error	Medium	The codebase and database was routinely backed up on Github and Firebase respectively to ensure ease of restoration

5.5.4. Supportability Requirements

Table 5.6 shows the supportability requirements for the system. The ability of the system to run across multiple browsers and to run on multiple desktop platforms were assessed.

Table 5.6: Supportability Requirements

Requirements	Priority	Result
Accessibility across all browsers	High	The system was accessible across major browsers; Google Chrome, Safari, Mozilla Firefox and Microsoft Edge. This fact will be disclosed when deploying the system and will form part of the system requirements
Accessibility across all desktop platforms	Medium	The system can be deployed on Windows and Linux systems. The codebase was developed to cater for the two major systems that are used for servers

Chapter 6 : Discussion

6.1. Introduction

The HIV/AIDS viral load system was implemented by use of data from the online HIV Stanford Database. This data comprised of the patient's treatment change episode (TCE), which contains the drugs used in the past, the drugs currently being used, the CD4+ counts and viral load levels for the patient before the change of regimen, and during the new treatment. The TCE also contained the nucleotide sequences for the virus, which was used in the determination of the drug resistance levels for the medication the patient was on. This drug resistance was obtained from the Sierra Service. The models performance was evaluated based on the mean square error (MSE) and the coefficient of determination (R^2). The model was developed by use of a backpropagation artificial neural network with multiple-layers. The use of neural networks provided a faster way of performing the predictions for the HIV/AIDS viral load levels.

6.2. Model Validation

The prediction accuracy for the developed model was determined by the coefficient of determination and the mean square error. A cross validation of 10 folds was used to validate the model. The prediction accuracy of the model was set to 93.76%. This was done on a data set of 374 individual TCEs. The TCEs were selected from a group of 1518 TCEs, and involved removing any duplicates and TCEs with incomplete data. The mean square error for the trained model was 0.0323. A lower MSE denotes a better model that has been trained whereas a R^2 value closer to 1 denotes a better model that has been developed. The optimal learning rate for the model was 0.005, the momentum was set to 0.9.

6.3. Model Implementation Outputs

6.3.1. Training Outputs

The training outputs from the model trained were as captured in Table 6.1 This data was obtained after 298 iterations due to convergence of the model. The model had a mean square error (MSE) of 0.02587 with a learning rate of 0.0005. The coefficient of determination of 0.97412.

Table 6.1: Sample Expected Output and Predictions on Training

Expected Output	Predicted Output	Error
-0.843767364	-0.816364	-0.027403364
-0.137077927	-0.207077	0.069999073
1.197779898	1.393897	-0.196117102
1.11925885	1.291918	-0.17265915
0.962216753	1.061611	-0.099394247
-0.843767364	-0.805855	-0.037912364
-0.843767364	-0.804018	-0.039749364
-0.686725267	-0.638511	-0.048214267
-0.686725267	-0.71183	0.025104733
0.412569413	0.345061	0.067508413
Mean Square Error (MSE)	0.025876	
Coefficient of Determination (R^2)	0.974124	

6.3.2. Testing Outputs

The testing outputs from the model trained were as captured in

Table 6.2. This data was obtained after 620 iterations until convergence of the model. The model had a mean square error (MSE) of 0.087833 with a learning rate of 0.0005. The coefficient of determination of 0.912167.

Table 6.2: Sample Expected Outputs and Predicted Outputs on Testing

Expected Output	Predicted Output	Error
-0.84031933	-0.566701	-0.27362
-0.685787445	-0.577036	-0.10875
1.091329229	0.970283	0.121046
0.009606036	-0.326198	0.335804
0.318669805	0.082967	0.235703
0.395935748	0.180087	0.215849
-0.84031933	-0.741415	-0.0989
0.009606036	0.00037	0.009236
-0.685787445	-0.691041	0.005254
-0.685787445	-0.508442	-0.17735
Mean Squared Error (MSE)	0.087833	
Coefficient of Determination (R^2)	0.912167	

6.3.3. Validation Outputs

The validation outputs from the model trained were as captured in Table 6.3. This data was obtained after 620 iterations until convergence of the model. The model had a mean square error (MSE) of 0.062349 with a learning rate of 0.0005. The coefficient of determination of 0.937650.

Table 6.3: Sample Expected Outputs and Predicted Outputs on Validation

Expected Output	Predicted Output	Error
-0.845635997	-0.768517	-0.077118997
-0.683381765	-0.670127	-0.013254765
-0.115491953	-0.496783	0.381291047
-0.845635997	-0.839473	-0.006162997
-0.683381765	-0.642442	-0.040939765
-0.358873301	-0.325721	-0.033152301
-0.683381765	-0.682973	-0.000408765
1.993813061	1.615	0.378813061
-0.358873301	-0.599878	0.241004699
1.101414786	1.131859	-0.030444214
Mean Squared Error (MSE)	0.062349	
Coefficient of Determination (R2)	0.937650	

6.4. Contributions to Research

The model developed provided a solution to estimating viral load levels in HIV-positive patients, especially in limited resource settings. This model helps the clinician estimate the likelihood of a patient's prescribed treatment not being effective, and thus can institute further monitoring on the patient. This also reduces the need for the continual measurement of viral load of HIV-patients, especially in limited resource settings.

6.5. Challenges

Availability of quality data was a challenge during the study. The data was obtained from secondary sources, and contained gaps and duplicates. This reduced the available training data that could be used, and affected the model accuracy.

Chapter 7 : Conclusions and Recommendations

7.1. Conclusions

The main objective of this research was to develop a HIV/AIDS viral load prediction system that utilizes the artificial neural network algorithm. This objective was split into five major objectives in order to adequately achieve the main objective. These major objectives are as follow:

- i) To investigate the factors influencing the HIV/AIDS progression in human beings
- ii) To analyze the methods used in the measurement of HIV/AIDS progression
- iii) To review the existing prediction algorithms and models used in the prediction of HIV/AIDS progression

The first three objectives were met by reviewing existing literature. These algorithms helped the researcher to identify the factors affecting the progression, and enabled the researcher to understand which attributes contribute to the final viral load of the HIV-positive patient. This was covered in chapter 2.

- iv) To develop an HIV/AIDS viral load prediction system

The system was developed using secondary data. Data-driven modelling was used, and this is discussed in chapter 3. The design for the system included the use case, flow chart and database schema. These designs were used to model how the system components interact, and therefore guide the development of the system. This objective was met and is covered in chapter 4 and chapter 5. The tools used in the design of the system included Microsoft Word.

- v) To test the functionality of the developed HIV/AIDS viral load prediction system

Evaluation of the functionality of the developed system was done by validating the outputs generated from the developed system against the expected output. The ease of use of the system was ensured by use of very simple interfaces. This objective was met and is discussed in chapter 5.

Wasti, et al., (2012) and Beer, et al., (2012) identified that a high level of patient adherence to the prescribed antiretroviral therapy (ART) regimen. This means that the viralological success of the treatment has to fall within the prescribed 95% that Wasti, et al., (2012) suggested. This implies that the patient would require to be strictly following the regimen as indicated in order to have an optimal viral suppression to prevent superinfection or transmission of the virus.

This study has also shown that the use of multi-layer back-propagated artificial neural networks in prediction of a HIV-positive patient's viral load level is possible. The study showed the relatively high accuracy levels that were obtained from the developed model, with the data pre-processed first and applied to the neural network model. The pre-processing ensured that the model was able to perform with high accuracy and eliminated bias.

The model created was able to produce predictions that had high accuracy and low errors as indicated by the findings in the study. This illustrates the ability of neural networks to be used in dispensing centers and comprehensive care centers to perform prediction of HIV/AIDS viral load levels. The study findings also pointed out the need to have more data attributes collected from the patients in order to better improve the model generated. This would help in the monitoring and follow-up of the patients' to ensure appropriate care is given.

7.2. Recommendations

From the results obtained, the following recommendations can be made:

- i. The degree of adherence to the prescribed ART regimen by the patient can be utilized in the model to improve the outputs of the prediction
- ii. The model can also be trained with data from known sources so as to evaluate the effectivity based on location
- iii. More data to be collected in order to further improve the model's prediction performance

7.3. Suggestions for Future Research

More research could be done on how to estimate the patient's adherence to the prescribed regimen by use of their lifestyle. This could provide a basis for computing the adherence percentage which could be used as an input to the developed model. The developed system should also be interlinked with the currently existing systems to provide seamless integration of the systems. This will eliminate the need to have to manually transfer data, and ultimately

eliminate the errors attributable to human intervention. The data could be used to train the model periodically thus resulting to improved prediction accuracy.

Other types of Artificial Neural Networks such as associative neural networks can be used to develop the model. This can ultimately lead to the determination of the most optimal neural network to be used in training and developing the model for the system.

References

- AIDS Gov. (2016, 11 28). *HIV LIFECYCLE*. Retrieved from AIDS Gov: <https://www.aids.gov/hiv-aids-basics/just-diagnosed-with-hiv-aids/hiv-in-your-body/hiv-lifecycle/>
- AIDS INFO. (2016, 12 8). *HIV Treatment, Drug Resistance*. Retrieved from AIDS INFO: <https://aidsinfo.nih.gov/education-materials/fact-sheets/21/56/drug-resistance>
- Albu, A., & Stanciu, L. (2015). Benefits of Using Artificial Intelligence in Medical Predictions. *The 5th IEEE International Conference on E-health and Bioengineering*. Iasi: IEEE.
- Ali, J., Khan, R., Ahmed, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science*, 9(5), 272-278.
- Alick, R. S. (2016). *Real-time solution for automated inventory monitoring of antiretroviral medicines: case of Nairobi County*. Nairobi: Strathmore University.
- Archer, J. P. (2008). *The Diversity of HIV -I*. Manchester: University of Manchester.
- 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>
- Averting HIV and AIDS. (2016, 11 25). *HOW HIV INFECTS THE BODY AND THE LIFECYCLE OF HIV*. Retrieved August 14, 2016, from Averting HIV and AIDS: <http://www.avert.org/about-hiv-aids/how-infects-body>
- Aylien, N. B. (2016, 07 01). *Support Vector Machines: A simple Explanation*. Retrieved from KDNuggets: <http://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html>
- Beer, L., Heffelfinger, J., Frazier, E., Mattson, C., Roter, B., Barash, E., . . . Valverde, E. (2012). Use of and Adherence to Antiretroviral Therapy in a Large U.S. Sample of HIV-infected Adults in Care, 2007-2008. *The Open AIDS Journal*(6), 213-223.
- Bertagnolio, S., Perno, C. F., Vella, S., & Pillay, D. (2013). The Impact of HIV Drug Resistance on the Selection of First- and Second-Line ART in Resource-Limited Settings. *Journal of Infectious Diseases*, 207, S45-S48.

- Boutayeb, A. (2009). The impact of HIV/AIDS on human development in African countries. *BMC Public Health*, 9.
- Cai, Y.-D., Liu, X.-J., Xu, X.-b., & Zhou, G.-P. (2001). Support Vector Machines for predicting protein structural class. *BMC Bioinformatics*.
- Center for Innovation in Research and Teaching. (2016, 12 02). *Basic Research Designs*. Retrieved from Center for Innovation in Research and Teaching: <https://cirt.gcu.edu/research/developmentresources/tutorials/researchdesigns>
- Charan, J., & Biswas, T. (2013). How to Calculate Sample Size for Different Study Designs in Medical Research? *Indian Journal of Psychological Medicine*, 2, 121-126.
- Chebet, H. M., Orero, J., & Luvanda, A. (2014). A Knowledgebase Model for Islamic Inheritance. *Information and Knowledge Management*, 4.
- Chesney, M. A. (2000). Factors Affecting Adherence to Antiretroviral Therapy. *Clinical Infectious Diseases*, 30(2), S171-S176.
- Chou, D., Iu, R., Krishna, R., & Liang, A. (2012). *An Analysis on the Prediction of HIV Progression*.
- Cohen, M. S. (2007). Preventing Sexual Transmission of HIV. *Clinical Infectious Diseases*, 287-292.
- Craenenbroek, V. E., Vermeiren, H., Muyldermans, G., Borgt, V. K., Alen, P., Bachelor, L., & Lecocq, P. (2007). Prediction of HIV-1 Susceptability Phenotype from Viral Genotype using Linear Regression Model. *Journal of Virological Methods*, 60.
- Doorn, J. v. (2014). Analysis of Deep Convolutional Neural Network. *21st Twente Student Conference on IT* (pp. 1-7). Enschede: University of Twente.
- Emuoyibofarhe, O. J., Oladosu, J. B., Omotosho, I. O., Popoola, O. P., & A, J. (2016, 12 05). *Prediction of HIV/AIDS Status using Artificial Neural Network*. Retrieved from University of Lagos: [www.unilag.edu.ng/opendoc.php?sno=1635&doctype=doc...\\$](http://www.unilag.edu.ng/opendoc.php?sno=1635&doctype=doc...$)
- Fevrier, M., Dorgham, K., & Robello, A. (2011). CD4+ T Cell Depletion in Human Immunodeficiency Virus (HIV) Infection: Role of Apoptosis. *Viruses*, 586-612.

- Freyder, C. W. (2014). *USING LINEAR REGRESSION AND MIXED MODELS TO PREDICT HEALTH CARE*. Pittsburgh: University of Pittsburgh.
- Gharai, A. (2017, 01 17). *Incremental Model*. Retrieved from Testing Excellence: <http://www.testingexcellence.com/incremental-model/>
- Government of Quebec. (2015). *Expert Consensus: Viral Load and the Risk of HIV Transmission*. Quebec: National Institute of Pulmonary Health of Quebec.
- Grant, R. M., & McConnell, J. J. (2017, 01 08). *What is HIV Superinfection and How Do I Prevent It?* Retrieved from HIV plus mag: <http://www.hivplusmag.com/treatment/2014/04/10/what-hiv-superinfection-and-how-do-i-prevent-it>
- 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/the-difference-between-hiv-viral-load-and-cd4-tests.html>
- Institut National de Sante' Publique du Quebec. (2014). *Expert Consensus: Viral Load and the Risk of HIV Transmission*. Quebec: Institut National de Sante' Publique.
- Karsoliya, S. (2012). Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture. *International Journal of Engineering Trends and Technology*, 3(6), 714-717.
- Kataria, A., & Singh, M. D. (2013). A Review of Data Classification Using K-Nearest Neighbour. *International Journal of Emerging Technology and Advanced Engineering*, 354-360.
- Kumar, J. R. (2014, December 30). *Understanding Linear Regression*. Retrieved from Data Science Central: <http://www.datasciencecentral.com/profiles/blogs/understanding-linear-regression>
- Levi, J., Raymond, A., Pozniak, A., Vernazza, P., Kohler, P., & Hill, A. (2016). Can the UNAIDS 90-90-90 target be achieved? A systematic analysis of national HIV treatment cascades. *BMJ Global Health*, 1-10.

- Levy, J. A. (2007). *HIV and the Pathogenesis of AIDS*. Washington: ASM Press.
- Liaw, A., & Wiener, M. (2017, 02 22). *UNC Gillings School of Public Health*. Retrieved from Classification and Regression by randomForests: <http://www.bios.unc.edu/~dzeng/BIOS740/randomforest.pdf>
- 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.
- Miller, V., Phillips, A. N., Clotet, B., Mocroft, A., Ledergerber, B., Kirk, O., . . . Lundgren, J. D. (2002). Association of Virus Load, CD4 Cell Count, and Treatment with Clinical Progression in Human Immunodeficiency Virus - Infected Patients with very Low CD4 Cell Counts. *The Journal of Infectious Diseases*, 186(2), 189-197.
- myMVC. (2016, 11 26). *Antiretroviral Therapy (Anti-HIV Drugs)*. Retrieved from my Virtual Medical Centre: <http://www.myvmc.com/treatments/antiretroviral-therapy-anti-hiv-drugs/>
- National AIDS & STI Control Programme. (2016, 12 02). National ACT Dashboard. Nairobi.
- National AIDS and STI Control Programme. (2014). *KENYA HIV ESTIMATES*. Nairobi: National AIDS and STI Control Programme.
- NIST/SEMATECH. (2017, March 25). *How can I tell if my model fits my data?* Retrieved from Engineering Statistics Handbook: <http://www.itl.nist.gov/div898/handbook/pmd/section4/pmd44.htm>
- Nyaga, R. K., Kimani, D. N., Mwabu, G., & Kimenyi, M. S. (2004). *HIV/AIDS in Kenya: A Review of Research and Policy Issues*. Nairobi: Kenya Institute for Public Policy Research and Analysis.
- Panel on Antiretroviral Guidelines for Adults and Adolescents. (2016, 11 26). *Guidelines for the Use of Antiretroviral Agents in HIV-1-Infected Adults and Adolescents*. Retrieved from AIDS INFO: <https://aidsinfo.nih.gov/guidelines/html/1/adult-and-adolescent-arv-guidelines/30/adherence-to-art>

- PennState Eberly College of Science. (2017, March 31). *The Coefficient of Determination, r-squared*. Retrieved from PennState Eberly College of Science: <https://onlinecourses.science.psu.edu/stat501/node/255>
- Pinheiro, J. V., Lemos, J. M., & Vinga, S. (2011). Nonlinear MPC of HIV-1 infection with periodic inputs. *50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)* (pp. 65-70). Orlando: IEEE.
- Puren, A., Gerlach, J. L., Weigl, B. H., Kelso, D. M., & Domingo, G. J. (2010). Laboratory Operations, Specimen Processing, and Handling for Viral Load Testing and Surveillance. *Journal of Infectious Diseases*, 1491-1496.
- Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1998). *Applied Regression Analysis: A Research Tool* (2nd ed.). New York: Springer.
- Research Information Network. (2010). *Quality Assurance and Assessment of Scholarly Research: A guide for researchers, academic administrators and librarians*. London: Research Information Network.
- Revell, A. D., Ene, L., Duiculescu, D., Wang, D., Youle, M., Pozniak, A., . . . Larder, B. A. (2012). The use of computational models to predict response to HIV therapy for clinical cases in Romania. *GERMS*, 6-11.
- 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.
- Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence A Modern Approach* (3 ed.). Upper Saddle River: Pearson Education.
- Sayad, S. (2017, March 21). *K Nearest Neighbors*. Retrieved from University of Toronto: <http://chem-eng.utoronto.ca/~datamining/Presentations/KNN.pdf>
- Schneider, A., Hommel, G., & Blettner, M. (2010). Linear Regression Analysis: Part 14 of a Series on Evaluation of Scientific Publications. *Deutsches Arzteblatt International*, 776-782.

- 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.
- Shen, C., Yu, X., Harrison, R. W., & Weber, I. T. (2016). Automated prediction of HIV drug resistance from genotype data. *BMC Bioinformatics*, 17, 278.
- Smith, D. M., Richman, D. D., & Little, S. J. (2005). HIV Superinfection. *The Journal of Infectious Diseases*, 192, 438-444.
- Solomatine, D., See, L. M., & Abrahart, R. J. (2008). Practical Hydroinformatics. In D. Solomatine, L. M. See, & R. J. Abrahart, *Computational Intelligence and Technological Developments in Water Application* (pp. 17-29). New York: Springer.
- Statistica. (2017, March 21). *K-Nearest Neighbors*. Retrieved from Statistica: <http://www.statsoft.com/textbook/k-nearest-neighbors>
- Stayerberg, E. W. (2009). Statistics for Biology and Health. In E. W. Stayerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* (pp. 11-30). New York: Springer Science + Business Media LLC.
- Tastan, O., Qi, Y., Carbonell, J. G., & Kleinseetharaman, J. (2009). PREDICTION OF INTERACTIONS BETWEEN HIV-1 AND HUMAN PROTEINS BY INFORMATION INTEGRATION. *Pacific Symposium on Biocomputing* (pp. 516-527). Pacific Symposium on Biocomputing.
- UNAIDS. (2017, March 28). *HIV and AIDS estimates (2015) - Kenya*. Retrieved from UNAIDS: <http://www.unaids.org/en/regionscountries/countries/kenya>
- 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>
- Venkatesan , P. (2006). A comprehensive back calculation Framework for Estimation and Prediction of HIV/AIDS in India. *Journal of Communicable Diseases*, 40-56.

- Wanjugu, S. W. (2015). *Data Centric Decision Making Healthcare Prototype*. Nairobi: Strathmore University.
- Wasti, S. P., Simkanda, P., Randall, J., Freeman, J. V., & Teijlingen, E. v. (2012). Factors Influencing Adherence Treatment in Nepal: A mixed-methods study. *PLoS ONE*, 7(5), 1-11.
- Wilson, D. P., Law, M. G., Grulich, A. E., Cooper, D. A., & Kaldor, J. M. (2008). Relation between HIV Viral Load and infectiousness: a model based analysis. *Lancet*, 314-320.
- Zoya, K., & Sezerman, O. U. (2016). Prediction of HIV Drug Resistance by Combining Sequence and Structural Properties. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, PP, 1-1.

Appendices

Appendix A: Originality Report

Turnitin Originality Report

A HIV/AIDS VIRAL LOAD PREDICTION SYSTEM USING ARTIFICIAL NEURAL NETWORKS by Titus Tunduny Kipkosgei



From 2016 Plagiarism Check (GS) (Library Services Plagiarism Checker (2016+))

- Processed on 06-Apr-2017 7:40 PM EAT
- ID: 795412114
- Word Count: 16049

Similarity Index

18%

Similarity by Source

Internet Sources:

14%

Publications:

8%

Student Papers:

12%

sources:

- 1 < 1% match (student papers from 20-Mar-2015)
[Submitted to Strathmore University on 2015-03-20](#)

- 2 < 1% match (student papers from 31-Mar-2017)
[Submitted to Sabanci Universitesi on 2017-03-31](#)

- 3 < 1% match (Internet from 09-Mar-2017)
<http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1114-6>

- 4 < 1% match (Internet from 04-Jan-2017)
<http://www.aidslaw.ca/site/download/14825/?lang=en>

Appendix B: Fasta Input File

```
>seq0
FQIWEEFSRAAEKLYLADPMKVRVVLKYRHVDGNLCIKVTDDLVLVYRTDQAQDVKKIEKF
>seq1
KYRTWEEFTRA AEKLYQADPMKVRVVLKYRHCDGNLCIKVTDDVVCLLYRTDQAQDVKKIEKFHSQLMRLME
LKVTDNKECLKFKTDQAQEAKKMEKLNNIFFTL
>seq2
EEYQIWEEFARAAEKLYLTDPMKVRVVLKYRHCDGNLCMKVTDDAVCLQYKTDQAQDVKKVEKLHGK
>seq3
MYQVWEEFSRAVEKLYLTDPMKVRVVLKYRHCDGNLCIKVTDNSVCLQYKTDQAQDVK
>seq4
EEFSRAVEKLYLTDPMKVRVVLKYRHCDGNLCIKVTDNSVVS YEMRLFVQKDNFALEHSL
>seq5
SWEEFAKAAEVLYLEDPMKCRMCTKYRHVDHKL VVKLT DNHTVLKYVTDMAQDVKKIEKLTTLLMR
>seq6
FTNWEEFAKAAERLHSANPEKCRFVTKYNHTKGELVLKLTDDVVCLQYSTNQLQDVKKLEKLSSTLLRSI
>seq7
SWEEFVERSVQLFRGDPNATRYVMKYRHCEGKLV LKVTDDRECLKFKTDQAQDAKKMEKLNNIFF
>seq8
SWDEFVDRSVQLFRADPESTRYVMKYRHCDGKLV LKVTDNKECLKFKTDQAQEAKKMEKLNNIFFTL
>seq9
KNWEDFEIAAENMYMANPQNCRYTMKYVHSGHILLKMSDNV KCVQYRAENMPDLKK
>seq10
FDSWDEFVSKSVELFRNHPDTTRYVVKYRHCEGKLV LKVTDNHECLKFKTDQAQDAKKMEK
```

Nucleotide Sequence

CCTCAGATCACTCTTTGGCAACGACCCATCGTCACAATAAAGATAGGGGGGCAATTA
AAGGAAGCTCTATTAGATACAGGAGCAGATGATACAGTATTAGAAGAAATGAATTT
GCCAGGAAGATGGAAACCAAAAATGATAGGGGGAATTGGAGGTTTTATCAAAGTAA
GACAGTATGAGGAGGTACCCATAGAAATCTGTGGACATAAAGCTATAGGTACAGTA
TTAATAGGACCTACACCAGYCAACATAATTGGAAGAAATCTAATGACTCAGCTTGGT
TGCACTCTAAACTTT

Appendix C: Sample XML

Part of the Sample XML file. *Note: The XML file is very long hence the cropping.*

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<?xml:stylesheet type="text/xsl" href="http://www.w3.org/2001/XMLSchema.xsl">
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="tce">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="schemaVersion" type="xs:string" minOccurs="0" maxOccurs="1"/>
        <!--
          "inclusionCriteria" (optional): description of TCE collection criteria such as new regimen should include NfV or boosted PI
        -->
        <xs:element name="inclusionCriteria" type="xs:string" minOccurs="0" maxOccurs="unbounded"/>
        <!--
          "requiredTCEComponents" (optional): for example, baseline genotype is required within 24 weeks from baseline
        -->
        <xs:element name="requiredTCEComponents" type="xs:string" minOccurs="0" maxOccurs="unbounded"/>
        <!--
          "patient" (optional): patient alias, age or race, etc. can be included
        -->
        <xs:element name="patient" type="Patient" minOccurs="0" maxOccurs="1"/>
        <!--
          either of "baselineDate" (a date in YYYY-MM-DD format) or "baselineYear" (just a year in YYYY format) is required
        -->
        <xs:choice>
          <xs:element name="baselineDate" type="xs:date"/>
          <xs:element name="baselineYear" type="xs:nonNegativeInteger"/>
        </xs:choice>
        <!--
          either of day, week, or month which is the time unit to produce the time-points relative to a baseline
        -->
        <xs:element name="dateUnit" type="DateUnitType"/>
        <!--
          baselineRNA and baselineCD4 are required, only one instance for each.
        -->
        <xs:element name="baselineRNA" type="RNAMeasurement" minOccurs="1" maxOccurs="1"/>
        <xs:element name="baselineCD4" type="CD4Measurement" minOccurs="1" maxOccurs="1"/>
        <!--
          "baselineNewTreatment" (required): list of drugs given at the baseline
        -->
        <xs:element name="baselineNewTreatment" type="Regimen" minOccurs="1" maxOccurs="1"/>
        <!--
          "baselineIsolate" (required): sequence at the baseline, each gene separately, one protease and one RT at the minimum
        -->
        <xs:element name="baselineIsolate" type="Isolate" minOccurs="2" maxOccurs="3"/>
        <!--
          "followupRNA" (required): HIV-1 plasma RNA levels during the new therapy
        -->
        <xs:element name="followupRNA" type="RNAMeasurement" minOccurs="1" maxOccurs="unbounded"/>
        <!--
          "followupCD4" (optional): CD4 counts during the new therapy
        -->
        <xs:element name="followupCD4" type="CD4Measurement" minOccurs="0" maxOccurs="unbounded"/>
        <!--
          the drugs received prior to baseline can be expressed in two ways: "pastRegimenTreatments" and "pastCumulativeDrugTreatments"
        -->
        <!--
          using "pastRegimenTreatments", the drugs received can be listed regimen by regimen
        -->
        <!--
          using "pastCumulativeDrugTreatments", the list of drugs can be listed drug by drug
        -->
        <xs:element name="pastRegimenTreatments" type="Regimen" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element name="pastCumulativeDrugTreatments" type="CumulativeDrug" minOccurs="0" maxOccurs="unbounded"/>
        <!--
          "pastRNA" and "pastCD4" (both optional): list of plasma HIV-1 RNA levels or CD4 counts measured prior to "baselineRNA" or "baselineCD4"
        -->
        <xs:element name="pastRNA" type="RNAMeasurement" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element name="pastCD4" type="CD4Measurement" minOccurs="0" maxOccurs="unbounded"/>
        <!--
          past genotypes (optional) can be expressed in two ways: "pastIsolate" or "pastCumulativeMutation"
        -->
        <xs:element name="pastIsolate" type="Isolate" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element name="pastCumulativeMutation" type="CumulativeMutations" minOccurs="0" maxOccurs="unbounded"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:complexType name="Patient">
    <xs:all>
      <xs:element name="patientAlias" type="xs:string" minOccurs="0" maxOccurs="1"/>
      <xs:element name="geographicRegion" type="xs:string" minOccurs="0" maxOccurs="1"/>
      <xs:element name="CD4NadirBeforeTCE" type="xs:nonNegativeInteger" minOccurs="0" maxOccurs="1"/>
      <xs:element name="age" type="xs:float" minOccurs="0" maxOccurs="1"/>
      <xs:element name="gender" type="xs:string" minOccurs="0" maxOccurs="1"/>
      <xs:element name="race" type="xs:string" minOccurs="0" maxOccurs="1"/>
    </xs:all>
  </xs:complexType>
  <xs:complexType name="Regimen">
    <xs:sequence>
```