UCL Computer Science
PG Project Assessment Form (2015-16)

INSTRUCTIONS: *Use the General Comments box to note the particular strengths and weaknesses of the project and any factors that are not covered by the rest of the form. Fill in a (%) mark for each of the four areas. See overleaf for more guidance. Underline key phrases in the descriptions that apply to this project where appropriate (also overleaf). Fill in an overall (%) mark. First Markers (Supervisors) should also grade students on their independence (Grades A to F). All parts of the form should be filled in.*

Student's Name	Vinit Vijay Jadhav	**Programme**	MSc Business Analytics with specialization in Computer Science
Marker's Name		**Marker's Role**	
Project Title	Predicting HIV drug resistance from genomic data using deep learning.		

General Comments (please write at least two or three sentences):		
1. Background, Aims and Organisation		**Mark (%)**
The student has not understood the aims of the project. The student has failed to place the work in context of the surrounding literature. The student has failed to identify suitable sub-goals.	The student has clearly understood and stated the aims of the project. There is a suitable literature review which relates to the task. The project is well-organised with suitable sub-goals.	
2. Difficulty Level and Achievement		
The student failed to achieve basic aims. Goals weren't sufficiently ambitious to warrant a whole project. Quality of the work is insufficient. The student has not produced sufficient deliverables.	The student has achieved all of the stated aims. Project is complex and challenging. The student has produced a considerably body of deliverables in terms of both software and write-up.	
3. Clarity		
The report is unclear or written badly. The write-up is disorganised. Figures and figure legends are of insufficient quality. The presentation is poor. It is hard to understand the core ideas.	Report is written carefully. Clear structure with a flowing, logical argument. Figures and legends are helpful for understanding the project. It is easy to understand the core ideas.	
4. Analysis / Testing		
For a software-based project there is insufficient testing. Documentation is poor. For a research-based project, there is no critical analysis of the results. Weaknesses and improvements aren't considered.	For software-based projects there is thorough testing. Analysis of strengths / weaknesses present. Detailed documentation. For research-based projects, there is critical analysis of method and results. Weaknesses and possible extensions are discussed.	
Overall Mark		
This is your overall mark given before discussion between the First and Second Markers. MSc Project Pass Mark: 50%. See overleaf for the criteria for each mark range.		

Supervision Level		**Grade (A-F)**
The student required close supervision and did not work independently (Grade F)	The student worked independently and did not overly rely on the supervisor (Grade A).	

Agreed Mark	**Mark (%)**
This is the mark agreed between the First Marker (Supervisor) and Second Marker after discussion. Please summarise how the final mark was agreed on between Examiners. This is particularly important where there is a significant difference between the first and second examiner and for borderline cases.	
Summary of Marker's Discussion (this must be completed):	

Marking Descriptors:

Range	Descriptor	Expectations
90-100% Exceptional	• Significant contribution to field, of publishable quality • Evidence of considerable extra-curricular reading, critical thought and original interpretation • Challenging goals and substantial deliverables, research level insight needed • Close to faultless in execution and write-up, a high level of independence	This represents a really outstanding achievement. The project needs to clearly stand out above others. A mark in this range is hard to achieve and rare (< 1%).
80-89% Outstanding	• Potential contribution to field, could lead on to publishable work • Evidence of extra-curricular academic reading, critical thought and original interpretation • Only very minor faults in execution, depth of understanding or write-up • Challenging project and substantial deliverables, largely self-directed	Excellent in most respects but doesn't fully meet the criteria for the top range. A small number of projects are in this range each year (2-3%).
70-79% Excellent (Distinction)	• Very well written report with a clear logical structure • Demonstration of critical thought, understanding and extra-curricular reading • Some minor faults in execution or understanding, otherwise carried out effectively • A good level of challenge, substantial deliverables, and a good level of self-direction	This represents a straightforward distinction project. Most things have been done well, but there will be some faults or criticisms. The goals have been met. A reasonable number of projects can be expected to achieve this level (~20%).
60-69% Good (Merit)	• Clear project write-up with logical structure • Evidence of understanding, and at least some evidence of extra-curricular reading and critical thought • May contain some ambiguities or faults, not all goals fully achieved • Reasonable level of challenge, good quality deliverables, satisfactory self-management, with some supervision help needed occasionally	A good result, that is well on the way to delivering most features, but is not fully complete or finished, or has a lower level of challenge. The majority of projects are likely to be at this level.
50-59% Satisfactory (Pass)	• Adequate project write-up, lacking clarity or detail in places, or containing irrelevant material • Limited evidence of extra-curricular reading or original thought, mostly demonstrates understanding of core issues • Some significant deficiencies or incomplete goals, deliverables adequate but of limited quality • Project not particularly ambitious or challenging, more significant supervision help required	A satisfactory but limited result. The core features are in place but may be buggy or not that well defined. Enough has been done to present a viable solution, of which at least some parts can be demonstrated. A minority of projects in this range (maybe 20%).
45-49% (Borderline fail, but could pass with extra work)	• Write-up is sub-standard, with noticeable errors or ommisions, but could be made passable within a reasonable time • Some clear flaws in understanding, limited or no extra-curicular reading • Actual achievements not very substantial or challenging, deliverables of lower quality or incomplete, but could be improved fairly easily • Not quite enough challenge or depth demonstrated, required significant supervision or there was a failure to attend tutorials	The project has enough substance to demonstrate it could be made into a pass in a fairly short length of time but is still missing significant features, or the write-up fails to describe what was actually achieved.
30-44% Unsatisfactory (Clear fail)	• Write-up is somewhat incoherent, rushed, contains important omissions, or irrelevant material • Some serious flaws in understanding, little or no extra-curicular reading • A lack of concrete achievements, substantial parts missing • Serious lack of challenge or depth demonstrated, required excessive supervision or there was a failure to attend tutorials	The basis of a viable project may be present but is a long way from being completed. A significant amount of additional work would be needed to reach a passable standard.
0-29% Hopeless (Unacceptable fail)	• Write-up is substantially absent, incomprehensible or wrong • No proper evidence of understanding, serious lack of extra-curicular reading • No or minimal achievements demonstrated, lack of deliverables • No depth or challenge presented, all aspects of the project have been handled badly, tutorials not attended or failed to achieve anything	Inexcusable result, that really should never happen. A complete failure to engage and carry forward the project.

Project Classification: 0-49% (Fail); 50-59% (Pass); 60-69% (Merit); 70-100% (Distinction)

Predicting HIV drug resistance from genomic data using deep learning.

Vinit Jadhav

Supervised by

Dr. Daniel Hulme
Dr. Raman Gangakhedkar

A dissertation submitted in partial fulfillment

of the requirements for the degree of

MSc Business Analytics

of

University College London.

Department of Computer Science

University College London

September 4, 2016

I, Vinit Jadhav, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work. The report may be freely copied and distributed provided the source is explicitly acknowledged.

Acknowledgements

The success of this research would not have been possible without the support of many individuals and organizations. Firstly I thank Dr. Daniel Hulme for his words of encouragement and continued support throughout the research. I thank Dr. Raman Gangakhedkar from National AIDS Research Institute, India for the opportunity to research in a domain which has a huge impact on millions of lives fighting HIV across the world. I am immensely thankful to Dr. Radhika Brahme and Dr. Swarali Kulkarni for helping me acquaint myself with essential knowledge about HIV and HIV drug resistance. Lastly I thank my family and friends for their support and encouragement.

Abstract

Antiretroviral therapy is the combination of several antiretroviral drugs to reduce the rate at which the Human Immunodeficiency Virus (HIV) multiplies in the human body. HIV however mutates to develop resistance to the antiretroviral drugs and hence hinders the success of antiretroviral therapy. Genotypic and phenotypic tests are used to measure the susceptibility of a particular strain of HIV to various antiretroviral drugs. While genotypic tests look for changes in the HIV genetic sequence compared to the wild type HIV, phenotypic tests provide an accurate measure of antiretroviral drug resistance in a controlled environment. Phenotypic tests are however complex, expensive and time consuming as compared to genotypic tests. In this study 20357 pairs of HIV genomic sequences and their corresponding phenotypic resistance values for 17 antiretroviral drugs were used to construct deep learning models to predict phenotypic drug resistance from genomic data. The accuracy of these models was measured using 10-fold cross-validation. Average prediction accuracy of 90.50% was obtained for eight protease inhibitor antiretroviral drugs. The average accuracy reduced to 84.15% for three non-nucleoside reverse transcriptase inhibitor antiretroviral drugs and further to 81.30% for six nucleoside reverse transcriptase inhibitor antiretroviral drugs. A software tool was developed to harness the predictive capability of these models in predicting the phenotypic drug resistance from the genomic information of an individuals HIV strain. The tool is now being used by National AIDS Research Institute of India to predict the phenotypic drug resistance of the 17 antiretroviral drugs included in this study and help formulate effective drug regimens to reduce the viral load in infected individuals and help them lead better healthy lives.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides a brief overview of the dissertation. The motivation behind the research topic chosen is highlighted and the objectives of the research are stated unambiguously. The proposed methodology to achieve the set research objectives has been introduced and the structure of this thesis is outlined. The chapter summarizes the feedback received from the industry sponsor National AIDS Research Institute of India and their remarks about the publication aspects and application of this research.

1.1 Research Motivation

The human immunodeficiency virus (HIV) infection resulting into acquired immunodeficiency syndrome (AIDS) has been a cause of over a million yearly deaths worldwide for the past three decades. HIV infection interferes with the human immune system and deteriorates it over time leaving the human body highly susceptible to common infections which rarely affect individuals with a normal immune system. HIV infection is transmitted among humans via sexual contact, infected blood/sexual fluids and pregnancy among others. The World Health Organization (WHO) has globally adopted numerous measures to prevent transmission of HIV infection which include awareness about the infection, awareness about protected sex, and medication during pregnancy etc. In spite of these measures there are more than 34 million people living with HIV infection globally.

Global research and development about HIV and related infections has been carried out on a massive scale in the past few decades. Although HIV cannot be completely cured, researchers have developed certain drugs (antiretroviral drugs) over the past decade which can control the multiplication of HIV in the human body and thus prolong an infected individuals life. Exposure to such antiretroviral drugs has however led to the evolution of HIV. Over the years the human immunodeficiency virus has undergone changes in its genetic material which has resulted into the prevalence of mutated strains of HIV carrying certain mutations in its genetic code which result into resistance to some antiretroviral drugs. Such mutations limit the success of antiretroviral drugs and antiretroviral therapy (drug regimen which is a combination of three or more antiretroviral drugs).

Drug resistance testing provisions for the testing of the mutations and the impact of such mutations on antiretroviral drug resistance. These tests can be broadly classified into genotypic tests and phenotypic tests. Genotypic tests simply uncover the genomic details of a particular strain of HIV and require less time and effort. Phenotypic tests on the other hand are complex in nature as they involve measuring the effect of antiretroviral drugs on that particular mutated HIV strain in an isolated environment. As a result, phenotypic tests are time consuming, labour intensive and expensive, especially in a developing and populated country like India with a huge number of infected individuals but limited infrastructure for testing and research.

Keeping this in mind, the aim is to apply contemporary machine learning methodologies learnt over the duration of the masters programme to predict the phenotypic drug resistance of a particular strain of HIV given its genomic data i.e. the mutations observed in that particular strain. A system capable of learning from a dataset of correlated genotypic-phenotypic tests data and accurately classifying a given strain of HIV as resistant/susceptible to a particular antiretroviral drug can help clinicians save time, effort and cost involved in phenotypic tests and formulate effective antiretroviral drug regimens to keep HIV multiplication in control within

an infected individuals body and help that individual live a prolonged healthy life. The system will also enable the industry sponsor to initiate HIV subtype specific drug resistance research.

1.2 Research Objectives

The primary objective of this research is to address the drug resistance classification task given the genomic data of an individuals HIV strain. The objectives are listed as below.

- Construct deep learning models which use a dataset of correlated genotypic-phenotypic test information for the drug resistance classification task.
- Evaluate the performance of these models using the metrics of accuracy, sensitivity and specificity over 10-fold cross validation.
- Given the mutations in a particular individuals strain of HIV, classify if the strain is resistant/susceptible to 17 antiretroviral drugs using these models.
- Construct a Drug Resistance Classification System capable of processing a given input file of mutations and providing a drug resistance summary report based upon the classification made by the constructed models along with their associated metrics.

The experiments conducted while working towards the formulated research objectives have uncovered novel ways of addressing certain tasks. For instance, mixtures of nucleic acids at certain positions in the genomic information which have been ignored in similar previous researches, has been tackled in our research. Experimentation with the neural network architecture (number of hidden layers), learning algorithm, activation function and related hyper-parameters has been performed to find optimal models to address the drug resistance classification task.

1.3 Structure of Dissertation

The dissertation is structured keeping in mind the consolidation of information from different faculties of science including bio-informatics, medicine and computer sci-

ence. The work presented in this dissertation is arranged in a way such that basic background information required to briefly understand the domain of research has been presented initially and is followed by information related to the application of computer-science principles in previous research related to the same domain.

Chapter 1: Introduction

This chapter briefly introduces the task at hand and the approach selected to tackle this task. Details about the industry partner for this research and their suggestions/feedback has been summarized.

Chapter 2: Background

This chapter provides background information about HIV, genomics, drugs, drug resistance and drug resistance testing which constitutes the bio-informatics and medicine aspect of the research.

Chapter 3: Predicting Drug Resistance

This chapter summarizes previous research and the methodologies adopted by the respective researchers to tackle drug resistance classification tasks.

Chapter 4: Material, Methods, Experiments and Results

This chapter describes the crucial aspects of research including the data-sets used, methodology adopted and results and also provides a brief overview of the software tool created.

Chapter 5: General Conclusions

This chapter summarizes the findings of the research and dives into future applications of the research.

1.4 Industry Partner (NARI) and Feedback

National AIDS Research Institute, India (NARI) operating under the Indian Council of Medical Research is a premier institute dedicated to multi-disciplinary research on HIV and AIDS. NARI has been involved in active research over the past two decades exploring facets of HIV related to immunology, epidemiology, microbiol-

ogy and drug resistance. The institute has been successful in delivering cutting-edge insights into HIV treatment and HIV microbiology owing to an expert panel of research scientists and a large number of international collaborations. Owing to the large population of India, NARI enjoys large community participation and vibrant involvement in undertaken studies which makes their research community oriented and community sensitive.

The scientists at NARI have played a vital hand holding role in the initial phase of this research helping me acquaint myself with crucial concepts about HIV and bio-informatics. Availability of research journals to understand the domain and look into previous research related to drug resistance testing has been facilitated by NARI. The Director In-Charge, Dr. Raman Gangakhedkar, has supervised this research within NARI. The inputs provided by Dr. Gangakhedkar and his team have been crucial in the success of this research. The opportunity to present findings about this research to scientists and technical staff at NARI towards the end of the research has been a fantastic experience. Dr. Gangakhedkar expressed his satisfaction about the research in the below words.

Vinit has done a fine job towards applying his analytics knowledge in the HIV drug resistance testing domain. He has been comfortable operating in a domain which is vastly different from his field of study and shown great keenness and openness towards learning something entirely new. The final presentation and workshop provided to the staff at NARI was highly engrossing and I thank him for sharing his knowledge with peers at NARI.

1.5 Publication and Application

Two papers related to this research have been shaping up in the past couple of months. They are described below.

- The first paper titled ”Predicting HIV drug resistance from genomic data using deep learning” is a summary of this research.

- The second paper is yet to be titled and will focus on concordance of results from existing virtual phenotype systems which use different algorithms for the drug resistance classification task and the system constructed in our research.

These papers will be thoroughly scrutinized by NARI staff before they are put forth for review with various scientific journals. NARI also intends to publish these papers in their annual internal research journal.

1.6 Funding

The UCL student funding team has awarded an amount of GBP 1,150 under the Yusuf Ali Grant towards my travel related expenses. The Yusuf Ali grant is available for Indian, Bangladeshi or Pakistani students for conference or fieldwork costs related to the Masters programme. I am immensely thankful to the UCL funding team for considering my application and providing me this grant.

Chapter 2

Background

This chapter provides essential domain knowledge which forms the backbone of this research. A brief overview of HIV/AIDS, drugs and therapies used to control the spread of HIV, HIV evolution, drug resistance, and drug resistance testing has been provided. This chapter lays the foundation for understanding the statistical/machine-learning aspect of the research which is discussed in the following chapters.

2.1 Human Immunodeficiency Virus (HIV)

2.1.1 HIV and AIDS

Acquired Immune Deficiency Syndrome (AIDS) is a severe medical condition caused by HIV infection which severely hampers the functioning of the human immune system and makes the human body more susceptible to minor infections [1]. The detection of this medical condition dates back to 1981 [2]; however, the cause of this condition remained unknown for a brief while. A few years past its recognition, the cause of AIDS was identified as a retrovirus which is now known as the human immunodeficiency virus type 1 (HIV-1) [3]. Subsequently a similar virus was detected as the cause of AIDS in western Africa. This virus is now known as the human immunodeficiency virus type 2 (HIV-2) [4]. Phylogenetic tree representations of HIV-1 trace back its origin from chimpanzees [5] while the origins of HIV-2 trace back to the sooty mangabeys community in west Africa [6]. The transmission rate of HIV-2 is however much lower compared to HIV-1 as the viral

count in HIV-2 affected individuals is much lesser than the viral count in HIV-1 affected individuals [7]. The focus of this study is entirely on HIV-1 and any references to HIV from this point ahead in the report refer to the HIV-1 unless explicitly mentioned.

2.1.2 HIV/AIDS: A Global Pandemic

According to the Global Health Observatory (GHO) data of World Health Organisation (WHO) there were 36.9 million living with HIV/AIDS worldwide in 2014 [8]. The number of lives lost due to HIV/AIDS till now is more than 35 million. The number of deaths due to HIV related infections in 2015 alone is 1.1 million. A total of 2.1 million of new infections were reported in the year 2015 [1]. These humungous numbers of fatalities and infections related to HIV portray the threat posed by HIV since the past few decades. Moreover the presence of these infections in all continents of the world have made HIV a global threat. The sub-Saharan African region accounts to majority of the deaths as well as new infections and is followed by south and south-east Asia. The life expectancy of individuals in the countries majorly affected by HIV has reduced [9].

The economic impact of HIV on individuals as well as countries cannot be ignored. Loss of lives to HIV and related infections severely impacts the availability of human capital to a countries economy. The health and medical related costs in countries with large number of HIV infected individuals is significant [10]. Household spending of families having an infected individual has seen an increase due to expenditure on health care and medicines. At the same time, reduction in income and thereafter scarcity of income to spend on education and quality of life put the lives of such families in a vicious circle [10]. The stigma associated with HIV infection in certain developing countries and discrimination of infected individuals has long term psychological and sociological impact [11].

Since the past decade the United Nations has called for a unified effort from all countries to fight the HIV/AIDS pandemic [12]. This call has seen a substantial rise

in the number of governments and research institutions across the globe participate in and collaborate on HIV research spanning HIV phylogeny, drug development, prophylaxis, coinfection etc. These combined efforts have resulted in the formulation of effective antiretroviral drugs that can control virus multiplication and prevent transmission. As a result of the unified effort, the number of new HIV infections have fallen by 35% in the past 15 years while the number of deaths in the same period has fallen by 28% [1].

2.2 Antiretroviral Drugs and Antiretroviral Therapy

A vaccine or drug which can completely cure HIV infection is not yet available. However, the combined research effort across the globe had resulted in the development of antiretroviral drugs [13] which supress HIV infection in the human body by keeping HIV multiplication in the human body under control and thereby reducing the viral count/load. The combination of these antiretroviral drugs in order to optimally reduce the viral count in an individual is termed is antiretroviral therapy [14]. It is crucial to understand the HIV life cycle and the stage at which these antiretroviral drugs act in the HIV life cycle before we proceed to drug resistance testing and the statistical aspect of this research.

2.2.1 The HIV Life Cycle

The impact of HIV infection lies in the potential of HIV to attack and destroy CD4 cells which are responsible for the human body's response to any infections caused by bacteria or viruses. These CD4 cells play a vital role in maintaining the human immune system and loss of these cells makes the human body susceptible to minor infections. HIV makes use of the CD4 cells to replicate in the human body. HIV binds to a CD4 cell, hides its DNA inside the CD4 cells DNA and thereafter uses the CD4 cell to replicate itself. HIV carries its genetic information in two strands of RNA and after it binds to a CD4 cell the RNA is converted to DNA by a process called as reverse transcription. It is due to this aspect of HIV that it is called a retrovirus. Once the DNA is made, it is then carried to the host cell nucleus by the integrase enzyme. This DNA from the nucleus is then propagated to make new HIV.

The propagated elements need to be assembled together to make new HIV. The long strings of proteins (material to construct new HIV) are cut into smaller proteins by the protease enzyme for assembly. After this viral assembly, the new cell is cut-off from the host cell which then matures to form new viruses [15].

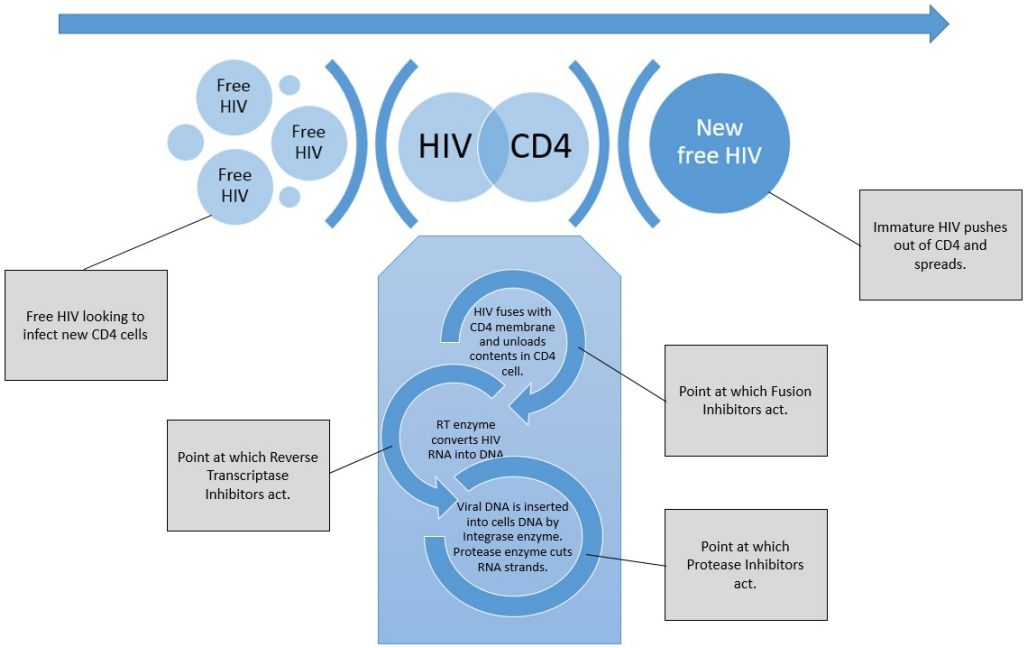


Figure 2.1: *The HIV Replication Cycle showing the stages at which each antiretroviral drug acts.*

2.2.2 Classification of Antiretroviral Drugs

Antiretroviral drugs are classified based upon the phase at which they act in the HIV lifecycle [16]. The class of antiretroviral drugs which act at the viral entry phase (binding) are termed as entry inhibitors [17]. These antiretroviral drugs are also referred as fusion inhibitors as they prohibit cell membrane fusion of the host CD4 cell and HIV [18]. Fusion inhibitors are however not commonly administered in antiretroviral therapy drug regimens.

The phase proceeding the binding phase in the HIV replication lifecycle is the reverse transcription phase where HIV RNA is converted to DNA by using the reverse transcriptase enzyme. Antiretroviral drugs which inhibit the reverse tran-

scriptase enzyme and block the reverse transcription process are called as reverse transcriptase (RT) inhibitors. These drugs are further classified into nucleoside reverse transcriptase inhibitors (NRTIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs) [19]. NRTIs and NNRTIs are widely used to formulate potent antiretroviral drug regimens.

Antiretroviral drugs which act at the phase where the HIV DNA is carried to the nucleus by the integrase enzyme are termed as integrase inhibitors. The use of integrase inhibitors in the formulation of antiretroviral drug regimens is yet limited and will gain momentum after significant clinical trials have been conducted with this class of drugs [20].

At the final stage of the HIV replication cycle the protease enzyme assembles necessary proteins to formulate a new infectious cell. Protease inhibitors act at this stage and prohibit this assembly [21]. Antiretroviral regimens make use of drugs from each of these categories for effective treatment of HIV infection. The various drugs belonging to these categories are listed in the next section.

2.2.3 List of Protease Inhibitors included in study.

This research is focused on 8 protease inhibitors for which genotypic-phenotypic data is available. The data sources are explained later in Chapter 4. The below table lists the FDA approved protease inhibitors along with their year of approval [22].

Abbreviation	Generic Name	Drug Name	Approval Year
ATV	Atazanavir	Atazanavir Sulfate	2003
DRV	Darunavir	Darunavir Ethanolate	2006
FPV	Fosamprenavir	Fosamprenavir Calcium	2003
IDV	Indinavir	Indinavir Sulfate	1996
NFV	Nelfinavir	Nelfinavir Mesylate	1997
SQV	Saquinavir	Saquinavir Mesylate	1995
TPV	Tipranavir	Tipranavir Disodium	2005
LPV	Lopinavir/Ritonavir	Lopinavir / Ritonavir	1996

Table 2.1: *Details of Protease Inhibitors included in this research.*

2.2.4 List of Reverse Transcriptase Inhibitors included in study.

This research is focused on 9 reverse transcriptase inhibitors for which genotypic-phenotypic data is available. The data sources are explained later in Chapter 4. The 9 reverse transcriptase inhibitors include 6 nucleoside reverse transcriptase inhibitors (NRTIs) and 3 non-nucleoside reverse transcriptase inhibitors (NNRTIs). The below table lists the FDA approved reverse transcriptase inhibitors along with their year of approval [22].

Type	Abbreviation	Generic Name	Drug Name	Approval Year
NRTI	AZT	Zidovudine	Azidothymidine	1987
NRTI	D4T	Stavudine	Stavudine	1994
NRTI	TDF	Tenofovir	Tenofovir Disoproxil	2001
NRTI	ABC	Abacavir	Abcavir Sulfate	1998
NRTI	3TC	Lamivudine	Lamivudine	1995
NRTI	DDI	Didanosine	Dideoxyinosine	1991
NNRTI	EFV	Efavirenz	Efavirenz	1998
NNRTI	NVP	Nevirapine	Nevirapine	1996
NNRTI	ETR	Etravirine	Etravirine	2008

Table 2.2: *Details of reverse transcriptase inhibitors included in this research.*

2.3 Mutations and Drug Resistance

The formulation of effective drug regimens involving three or more antiretroviral drugs to control the rate at which HIV multiplies in the human body is hampered by the capability of HIV to develop mutations in its genetic structure which render the mutated HIV resistant to certain antiretroviral drugs [24]. These mutations are developed in the viral proteins which are targeted by antiretroviral drugs as a result of selective pressure [25]. The HIV genetic sequence containing no mutations is commonly referred to as the HIV wild type sequence. Any mutations in the HIV wild type sequence may have an impact on one of the antiretroviral drugs from the classes discussed in the previous section [26]. Furthermore, the transmission of mutated HIV strains and continual evolution of HIV further complicates the choice of antiretroviral drugs to form an effective drug regimen in order to limit the viral load [25].

It is crucial to understand the representation of HIV mutations as all the data used in this research deals with genomic sequences which include these mutations and the effect on these mutations on resistance towards an antiretroviral drug. The HIV sequence is a stream of amino acids in which each amino acid is determined by three participating nucleotides (A,T,C,G) from the DNA sequence. This group of three nucleotides, called as a codon, represent an amino acid. The HIV wild type virus is represented by a series of these amino acids present at various codons. A change in one of the three participating nucleotides at a codon changes the amino acid at that particular codon. This change in amino acid at a particular codon compared to that at the wild type codon is termed as a mutation. An example of such a mutation is as below.

- In the wild type HIV, the Pol gene codon number 184 contains the amino acid methionine (M) represented by the nucleotide base sequence ATA.
- If the base sequence changes from ATA to GTA, the amino acid changes from methionine (M) to valine (V).
- This is a mutation which is represented as M184V signifying that the amino acid methionine (M) at codon 184 has mutated to valine (V).

A list of all amino acids which and their respective single letter codes are provided in the table 2.3.

2.4 Drug Resistance Testing

The mutations present in an individuals HIV strain affect the way antiretroviral drugs act upon that HIV strain. An individual may host wild type HIV strains as well as mutated strains and while an antiretroviral regimen may control the replication of the wild type virus, it may foster the multiplication of a mutated strain which is called selective pressure. Hence it is crucial to thoroughly understand the mutations in the individuals HIV and the impact such mutations may have on resistance

Abbrevation	Corresponding Amino Acid
I	Isoleucine
L	Leucine
V	Valine
F	Phenylalanine
M	Methionine
C	Cysteine
A	Alanine
G	Glycine
P	Proline
T	Threonine
S	Serine
Y	Tyrosine
W	Tryptophan
Q	Glutamine
N	Asparagine
H	Histidine
E	Glutamic acid
D	Aspartic acid
K	Lysine
R	Arginine

Table 2.3: *List of all amino acids which occupy codons in an HIV genomic sequence along with their single digit codes used in the master dataset.*

towards antiretroviral drugs before a drug regimen is administered to that particular individual. The main types of tests which cater to this purpose are described below.

2.4.1 Genotypic Testing

This type of testing studies the structure of the HIV genetic sequences in detail. The baseline to compare these structural changes is the genetic structure of the wild type HIV. As explained in the previous section a comparison is made between the amino acids at particular codons in the HIV wild type sequence and the mutated strain. The test results give a set of observed mutations. However, only a few of these observed mutations may cause resistance to certain antiretroviral drugs and it is at the discretion of a specialist doctor to interpret the test results and formulate effective drug regimens. It should however be remembered that the impact of a mutation at a specific codon on one drug may be countered by another mutation at another codon. This makes the interpretation of genotypic results complex. Also

the possibility of encountering a new combination of mutations complicates the interpretation further. Genotypic tests are however inexpensive and the results are available within a week of testing.

2.4.2 Phenotypic Testing

Phenotypic tests do not look at the HIV structural details. The focus of phenotypic tests is to directly observe the effect of antiretroviral drugs on an individuals HIV strain in an isolated in-vitro environment and compare that to the antiretroviral drugs effect on an HIV wild type virus. In order to perform this test, an individuals HIV is isolated and replicated in a test tube. The impact of an antiretroviral drug is then measured on this isolated HIV and compared to the impact of that antiretroviral drug on a wild type HIV. The impact is measured in terms of fold change. Fold change tells how much the quantity of an antiretroviral drug should be increased in order to have the same effect on the mutated HIV strain as on the wild type HIV. So a 5 fold resistance for an antiretroviral drug means that if an antiretroviral drug of quantity x limits the replication of wild type HIV then a quantity 5x would be necessary to limit the replication of the mutated HIV. This is also known as fold resistance.

Every drug has a threshold beyond which it cannot be administered as the drug may induce toxins in the human body and cause further side effects. This threshold is termed as the fold resistance cut-off and is different for every antiretroviral drug. The terms fold resistance and fold resistance cut-off are of utmost importance in this research as most of our data munging and data transformation activities explained in Chapter 4 are based around these concepts.

2.4.3 Virtual Phenotypic Testing

Virtual phenotypic tests provide a mechanism of predicting phenotypic test results based upon the results obtained from a genotypic test. The prediction mechanisms are supported by databases of matched genotypic-phenotypic test information and

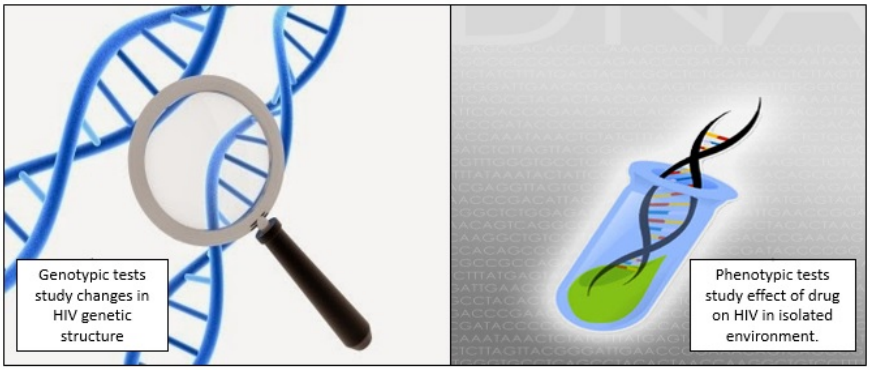


Figure 2.2: *An illustratory image showing difference between genotypic and phenotypic testing.*

driven by statistical techniques. This research focuses on establishing a similar mechanism using contemporary techniques in deep learning. Virtual phenotypic tests are of prime importance to the industry sponsor NARI, as a system capable of predicting phenotypic results will allow clinicians at NARI to use the system to formulate effective drug regimens for a huge population of HIV infected individuals in a country where majority of the infected individuals are not able to afford an actual phenotypic test due to economic constraints.

Chapter 3

Predicting Drug Resistance

The previous chapter provided essential knowledge about the HIV drug resistance testing domain. In this chapter we take a detailed look at previous research related to this domain. The topics discussed in this chapter will offer a detailed understanding of how genotypic-phenotypic data can be used to predict phenotypic drug resistance from genomic data. A summary of various methods employed by previous researchers for the drug resistance prediction task are discussed.

3.1 Rule Based Drug Resistance Prediction

An approach used by a number of researchers in the drug resistance prediction task is to formulate a set of complex rules involving mutations and the antiretroviral drugs affected by these mutations and then use this set of complex rules to determine if the mutations in a particular genomic sequence will impact its susceptibility to an antiretroviral drug. An important research adopting this approach is the work by K Van Laethem et al. [27]. from Rega Institute for Medical Research and University Hospitals, Belgium. The Rega algorithm uses a pre-categorized set of mutations and their impact on 15 antiretroviral drugs to predict therapy response. These rules have been set up by an individual expert and take into consideration the effects of combinations of mutations as well. The rules and mutation tables are used to identify if a sequence can be classified as susceptible, intermediate resistant or resistant. The rules and mutation tables behind this algorithm have been published for use and further research [27].

The ANRS algorithm [28] developed by the French National Agency for AIDS Research is a rule based algorithm providing the impact of drug resistance mutations on 22 antiretroviral drugs including protease inhibitors, reverse transcriptase inhibitors, fusion inhibitors as well as integrase inhibitors. The rules are based on tables containing details about the antiretroviral drugs and mutations associated with resistance or possible resistance towards these antiretroviral drugs. The table of rules has been made available for public use and further research by the agency [29].

An advanced algorithm using a score based approach has been developed by the researchers at Stanford University based upon their HIV Drug Resistance database [30]. This approach assigns a drug penalty score to every mutation. While categorizing an HIV sequence, the score of each mutation in that sequence is added to get a cumulative score. Based upon this score the sequence is categorized as susceptible, potential low resistant, low resistant, intermediate resistant and high resistant. The classification is available for 16 antiretroviral drugs [30].

The use of such rule based systems to interpret genomic data and provide a summary of their impact on the choice of antiretroviral treatment regimen is beneficial to physicians in most cases. However these systems offer limited insights into new mutations which surface due to the high evolution rate of HIV. The formulation of exhaustive rules to ensure coverage of these mutations is an exhaustive task. Furthermore, a discordance is observed between various algorithms which cater to this task [31]. An alternative approach to avoid these drawbacks is to use a bio informatics based approach to tackle the drug resistance prediction task.

3.2 Bioinformatics Based Drug Resistance Prediction

Bioinformatics based drug resistance prediction systems are not dependent upon a set of rules or scoring patterns created by expert users. These systems used sta-

tistical/computational techniques to find patterns between the genomic data (mutations). The backbone of these systems is the training data which is a dataset of matched genotype-phenotype results. Algorithms are trained based upon this data which are then used to predict a phenotypic result given a genomic sequence. A generic flowchart of the way these systems work is shown below.

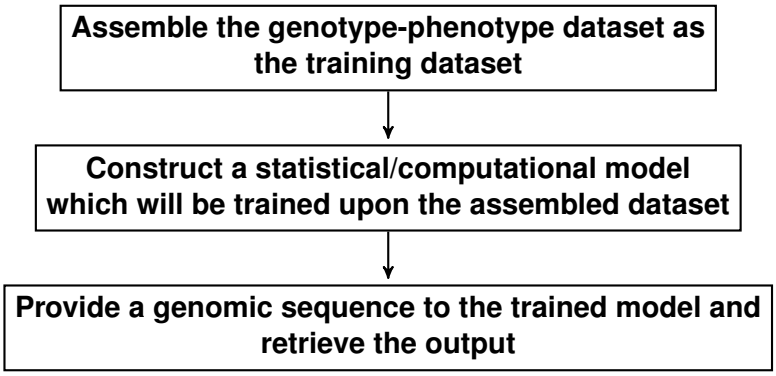


Figure 3.1: *Generic flowchart of bio-informatics based drug resistance prediction systems.*

Few bioinformatics based approaches adopted by researchers in the drug resistance prediction task are summarized below.

3.2.1 Statistical Learning Methods

Statistical learning methods are widely used in prediction tasks. In the drug resistance prediction context these learning methods can be viewed as regression problems where the phenotypic drug resistance is predicted based upon the mutation data in the genomic sequence.

3.2.1.1 Bayesian Variable Partitioning

A study by Zhang et. al. [32]. adopts a Bayesian statistical modelling approach based upon the genomic data to detect mutation patterns and their association with drug treatment instead of phenotypic resistance. The dataset of genomic sequences is first categorised into sequences belonging to treated individuals and sequences belonging to untreated individuals which are then assessed based upon the prior mutation data. This study however focuses only on the use of 3 antiretroviral drugs

and a full-fledged system capable of providing a holistic drug resistance summary report is not available [32].

3.2.1.2 Linear Regression Models

The study by Vermeiren et. al. [33] approaches the drug resistance prediction task by employing the linear regression modelling principles on the genotype-phenotype dataset. The phenotypic resistance in this case is calculated as the weighted sum of individual mutations in the genomic sequences. The predictions are made for 17 antiretroviral drugs by constructing a linear regression model for each drug. The study also identifies mutations which are significant in determining the phenotypic fold resistance values. These models show a high concordance to measured in-vitro fold change [33].

3.2.1.3 Cluster Analysis, Recursive Partitioning and Linear Discriminant Analysis

Another interesting research by Sevin et. al [34] employs the concepts of clustering, recursive partitioning and linear discriminant analysis to investigate the relationship between the HIV genome and phenotype fold changes for 2 antiretroviral drugs, Indinavir and Saquinavir. The research considers sequences from 72 individuals and their corresponding phenotypic test results for assessment. The first step in this research was to cluster the genomic sequences similar in nature in order to study the impact of these similar genomic sequences on drug susceptibility. Further recursive partitioning was used to best understand variables which distinctly split the clusters. The last step was to determine which genetic mutations best predict susceptibility to drugs by using linear discriminant analysis. The results of this research highlighted the mutations which have a high impact on resistance to Indinavir and Saquinavir. [34]

3.2.2 Computational Classification Methods

Classification methods use algorithms to classify if a genomic sequence with mutations are susceptible or resistant to a particular antiretroviral drug. This makes

it necessary to transform the phenotypic fold resistance values to a categorical variable depending upon the number of classes the classification is to be made. A study employing such a technique is the study by Beerenwinkel et. al. [35] which employs support vector machines (SVM) and decision trees for the classification task across 17 antiretroviral drugs. The study further proceeds to engineer a system Geno2Pheno which has been made available for public use. [35].

This research aims at developing a similar system using a machine learning approach and constructing neural network models capable of learning from labelled genotypic-phenotypic sequences for 17 antiretroviral drugs. The datasets used, experiments set up, results achieved and system developed are explained in detail in the next chapter.

Chapter 4

Material, Methods, Experiments and Results

This chapter gives detailed insights into the experiments carried out during the research and the outcome of these experiments. The initial section provides details about the data used and the methods adopted to transform this data into a format and structure which is usable for the experiments. The source of this data is acknowledged and novel aspects of experimentation based upon this data are described. The chapter then describes in detail the construction of deep learning models adopted for our classification task and the rationale behind the metrics / methodology chosen to validate these models. The HIV Drug Resistance Predictor Tool created based upon this research is described in detail at the end of this chapter.

4.1 Dataset

The prime aspect of this research is the genotypic-phenotypic dataset of HIV sequences across 17 antiretroviral drugs based upon which the neural network models will learn and predict if a given HIV sequence is resistant or susceptible to the antiretroviral drugs. The availability of this data for the research is ensured from two sources which are described below.

4.1.1 The Stanford HIV DB

The primary source of the required genotypic-phenotypic data is Stanford University's HIV Drug Resistance Database. This database has been created over a decade

from HIV drug resistance initiatives across the globe with an aim for provisioning data to researchers across the globe studying HIV drug evolution and resistance. In addition to data, it also provides various other services including HIV subtyping tools, mutation specific resistance summaries, and detailed mutation lists among others. HIV DB has their own drug resistance classification system which has employed various algorithms for the drug resistance prediction task, the results for which have been made available on their website and various research papers. These results have been used further for a comparative analysis of our results and are available in the appendix. The below image shows the home page of the HIV DB programme.

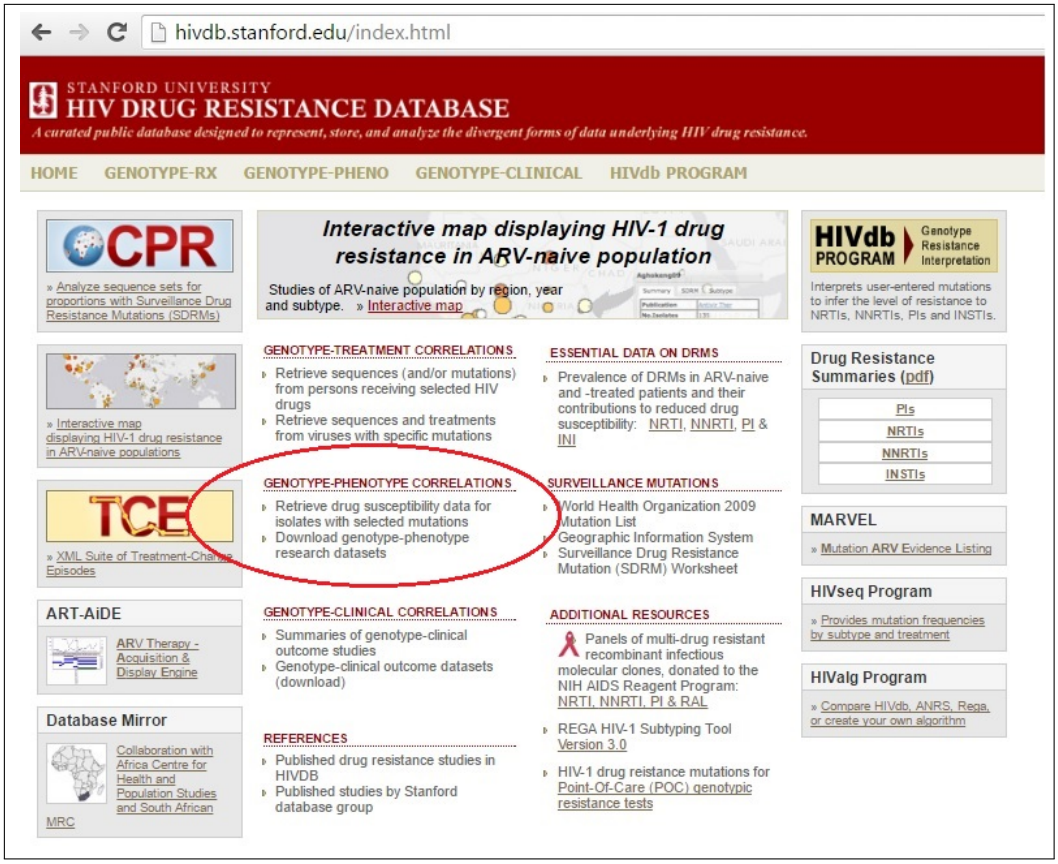


Figure 4.1: *Stanford University's HIV DB homepage showing the source of data used for this research.*

The genotype-phenotype correlation dataset from HIV DB provides us sequences of HIV strains belonging to various subtypes and their corresponding phe-

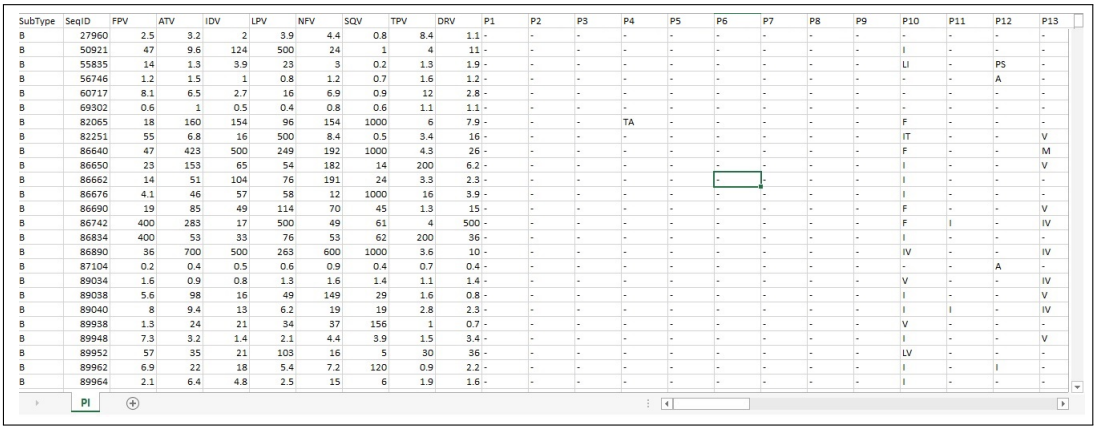
notypic values for 17 antiretroviral drugs. There are over 20,000 phenotypic results available which have been employed to train the neural network models in this research. The raw data in this dataset however needs to be treated before it can be used to train neural networks.

4.1.2 NARI Internal DB

The industry sponsor NARI has its own sequences and corresponding phenotypic results which have been accumulated over years of internal research. The genotypic-phenotypic data has been maintained in excel sheets and is used along with the Stanford HIV DB dataset to create a cumulative dataset for this research. This data is however less as compared to the Stanford HIV DB dataset with 367 sequences all belonging to HIV subtype C.

4.2 Data Transformation

The data available from the two sources listed above has representations in different formats and a single data source having a unified representation is essential to proceed with the research. In the data transformation task, a master dataset combining the two sources was created and stored in three separate CSV files each corresponding to a type of inhibitor namely protease inhibitor (PI), non-nucleoside reverse transcriptase inhibitor (NNRTI) and nucleoside reverse transcriptase inhibitor (NRTI). These CSV files formulate the entire data backbone of the research. The below image shows a glimpse of the master CSV after treatment from the raw format.



SubType	SeqID	FPV	ATV	IDV	LPV	NFV	SQV	TPV	DRV	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
B	27960	2.5	3.2	2	3.9	4.4	0.8	8.4	1.1	-	-	-	-	-	-	-	-	-	-	-	-	-
B	50921	47	9.6	124	500	24	1	4	11	-	-	-	-	-	-	-	-	-	I	-	-	-
B	55835	14	1.3	3.9	23	3	0.2	1.3	1.9	-	-	-	-	-	-	-	-	-	LI	-	PS	-
B	56746	1.2	1.5	1	0.8	1.2	0.7	1.6	1.2	-	-	-	-	-	-	-	-	-	-	-	A	-
B	60717	8.1	6.5	2.7	16	6.9	0.9	12	2.8	-	-	-	-	-	-	-	-	-	-	-	-	-
B	69302	0.6	1	0.5	0.4	0.8	0.6	1.1	1.1	-	-	-	-	-	-	-	-	-	-	-	-	-
B	82065	18	160	154	96	154	1000	6	7.9	-	-	-	TA	-	-	-	-	-	F	-	-	-
B	82251	55	6.8	16	500	8.4	0.5	3.4	16	-	-	-	-	-	-	-	-	-	IT	-	-	V
B	86640	47	423	500	249	192	1000	4.3	26	-	-	-	-	-	-	-	-	-	F	-	-	M
B	86650	23	153	65	54	182	14	200	6.2	-	-	-	-	-	-	-	-	-	I	-	-	V
B	86662	14	51	104	76	191	24	3.3	2.3	-	-	-	-	-	-	-	-	-	I	-	-	-
B	86676	4.1	46	57	58	12	1000	16	3.9	-	-	-	-	-	-	-	-	-	I	-	-	-
B	86690	19	85	49	114	70	45	1.3	15	-	-	-	-	-	-	-	-	-	F	-	-	V
B	86742	400	283	17	500	49	61	4	500	-	-	-	-	-	-	-	-	-	F	I	-	IV
B	86834	400	53	33	76	53	62	200	36	-	-	-	-	-	-	-	-	-	I	-	-	-
B	86890	36	700	500	263	600	1000	3.6	10	-	-	-	-	-	-	-	-	-	IV	-	-	IV
B	87104	0.2	0.4	0.5	0.6	0.9	0.4	0.7	0.4	-	-	-	-	-	-	-	-	-	-	-	A	-
B	89034	1.6	0.9	0.8	1.3	1.6	1.4	1.1	1.4	-	-	-	-	-	-	-	-	-	V	-	-	IV
B	89038	5.6	98	16	49	149	29	1.6	0.8	-	-	-	-	-	-	-	-	-	I	-	-	V
B	89040	8	9.4	13	6.2	19	19	2.8	2.3	-	-	-	-	-	-	-	-	-	I	I	-	IV
B	89938	1.3	24	21	34	37	156	1	0.7	-	-	-	-	-	-	-	-	-	V	-	-	-
B	89948	7.3	3.2	1.4	2.1	4.4	3.9	1.5	3.4	-	-	-	-	-	-	-	-	-	I	-	-	V
B	89952	57	35	21	103	16	5	30	36	-	-	-	-	-	-	-	-	-	LV	-	-	-
B	89962	6.9	22	18	5.4	7.2	120	0.9	2.2	-	-	-	-	-	-	-	-	-	I	-	I	-
B	89964	2.1	6.4	4.8	2.5	15	6	1.9	1.6	-	-	-	-	-	-	-	-	-	I	-	-	-

PI

Figure 4.2: *An image showing phenotypic results and mutations in the master CSV.*

4.2.1 Mutation Encoding

The master CSVs created show the mutation present in a sequence from codons 1 to 99 in case of protease inhibitors and codons 1 to 240 in case of reverse transcriptase (NRTI and NNRTI) inhibitors. These mutations which correspond to a changed amino acid at that particular codon are represented by the abbreviation for that particular amino acid and need to be converted to an equivalent numeric code before they can be fed to the neural network model for training. Also only those sequences which have phenotypic fold change values for a drug are to be considered when training a neural network for that particular drug. These tasks are accomplished by encoding the mutations with numeric values and analysing the sequences for the presence of a phenotypic fold change value for the particular model respectively. Codons where no mutations are observed are represented by a hyphen and need to be converted to an equivalent integer code as well. A sample R dataframe containing sequences filtered for constructing a neural network model for the drug Forsemprenavir (FPV) can be seen in the image below.

	FPV	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
1	2.5	-	-	-	-	-	-	-	-	-	I	-	-	-
2	0.7	-	-	-	-	-	-	-	-	-	-	-	-	-
4	1.5	-	-	-	-	-	-	-	-	-	-	-	-	V
5	3.9	-	-	-	-	-	-	-	-	-	V	-	-	-
6	9.5	-	-	-	-	-	-	-	-	-	F	-	-	-
8	3.1	-	-	-	-	-	-	-	-	-	I	-	-	-
9	4.9	-	-	-	-	-	-	-	-	-	I	-	-	-
10	1.2	-	-	-	-	-	-	-	-	-	-	-	-	-
12	8.3	-	-	-	-	-	-	-	-	-	-	-	-	V
13	2.7	-	-	-	-	-	-	-	-	-	I	-	-	-
14	2.1	-	-	-	-	-	-	-	-	-	-	-	-	-

Figure 4.3: *An R dataframe showing phenotypic fold resistance values and the corresponding genomic sequences for Forsemprenavir.*

R provides a function *utf8ToInt* which converts the UTF8 encoded character to its equivalent integer vector. With the help of this function the data in all the codons comprising of amino acid abbreviations in case of mutations and hyphens in case of no mutations are converted to an equivalent integer value. The code snippet to

accomplish this task can be seen below.

Listing 4.1: Code snippet for Sequence Analysis and Mutation Encoding

```
for (i in 1:nrow(data_drug)){
        for (j in 2:NCOL(data_drug)){
                x=utf8ToInt(data_drug[i,j])
                data_drug[i,j] = as.numeric(x)
        }
}
```

On conversion of the UTF8 encoded character to their corresponding integer vectors the R dataframe takes the below format which can then be further used for training neural network models.

	FPV	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
1	2.5	45	45	45	45	45	45	45	45	45	73	45	45	45
2	0.7	45	45	45	45	45	45	45	45	45	45	45	45	45
4	1.5	45	45	45	45	45	45	45	45	45	45	45	45	86
5	3.9	45	45	45	45	45	45	45	45	45	86	45	45	45
6	9.5	45	45	45	45	45	45	45	45	45	70	45	45	45
8	3.1	45	45	45	45	45	45	45	45	45	73	45	45	45
9	4.9	45	45	45	45	45	45	45	45	45	73	45	45	45
10	1.2	45	45	45	45	45	45	45	45	45	45	45	45	45
12	8.3	45	45	45	45	45	45	45	45	45	45	45	45	86
13	2.7	45	45	45	45	45	45	45	45	45	73	45	45	45
14	2.1	45	45	45	45	45	45	45	45	45	45	45	45	45

Figure 4.4: *An R dataframe showing phenotypic fold resistance values and the corresponding genomic sequences converted to integer vectors.*

4.2.2 Treating Mixture Mutations

A novel part of this research is the inclusion of mixture mutations while analysing the HIV sequences and including these sequences while training the neural network models for each antiretroviral drug. Mixtures are simply the presence of multiple amino acids at a particular codon in the HIV sequence. The impact mixtures at a particular codon have on the resistance developed to a particular antiretroviral drug is different than the impact of individual amino acids comprising that mixture.

The conversion of these mixtures to equivalent integer vectors however cannot be achieved as described in the previous section and needs an alternative approach. The approach adopted in the research to treat these mixtures is to consider all the amino acids participating in the mixture at a particular codon and sum up the individual integer vectors to formulate a composite integer vector representing the particular mixture. The code snippet performing this task is seen below.

Listing 4.2: Code snippet for Treating Mixtures

```
for (i in 1:nrow(data_drug)){
        for (j in 2:NCOL(data_drug)){
                x=utf8ToInt(data_drug[i,j])
                sum = 0
                if(length(x)>1){
                        no_mut[i]=no_mut[i]+1
                }
                if(length(x)>max_mix[i]){
                        max_mix[i]=length(x)
                }
                for(k in 1:length(x)){
                        sum = sum + x[k]
                }
                data_drug[i,j] = as.numeric(sum)
        }
}
```

The mutations in codons 1 to 99 in case of protease inhibitors and 1 to 240 in case of reverse transcriptase inhibitors are the features used to train the neural network models. In addition to these it is seen that the above snippet harnesses two more features. The first one is the number of mutations present in the given genomic sequence while the second one is the maximum length of mixture in the given sequence in case a mixture is present. The length of the mixture is the number of different amino acids present in the mixture.

4.2.3 Transforming phenotypic resistance to binary variable

The next crucial step in data transformation is to convert the phenotypic fold resistance cut-off values to binary values. This task is accomplished by simply comparing the fold resistance cut-off values corresponding to each antiretroviral drug in each sequence to the cut-off provided by Stanford HIV DB for the respective antiretroviral drugs. If the fold resistance value corresponding to each drug is less than the cut-off provided for that particular drug then the binary variable takes value 0 signifying that the sequence is susceptible to that drug, else, the binary variable takes value1 signifying that the sequence is resistant to that particular drug. This is achieved by the ifelse function in R.

Table 4.1 on the next page summarizes the cut-offs for the 17 antiretroviral drugs used in this research. The table shows a lower fold resistance cut-off as well as an upper fold resistance cut-off. The values for lower fold resistance cut-offs are used for the binary transformation. The application of upper fold resistance cut-offs is discussed in the future work section of the last chapter.

The data transformation task described in the above sections has addressed two major concerns.

- The feature data, which is the mutations observed in codons for each sequence has been converted to a numeric equivalent including mixtures if any at codons. The data at codons (1 to 90 for protease inhibitors, 1 to 240 for reverse transcriptase inhibitors) are now our predictor variables which will be fed to the neural network.

- The phenotypic fold resistance values corresponding to each drug in all sequences have been converted to their binary equivalent based upon the cut-off values provided by Stanford HIV DB. This new binary variable is now the target variable.

The next section describes how neural network models will be constructed and configured to harness the above transformed data for the drug resistance classification task.

Type	Abbreviation	Generic Name	Lower Cutoff	Upper Cutoff
PI	ATV	Atazanavir	3	15
PI	DRV	Darunavir	10	90
PI	FPV	Fosamprenavir	4	11
PI	IDV	Indinavir	3	15
PI	NFV	Nelfinavir	3	6
PI	SQV	Saquinavir	3	15
PI	TPV	Tipranavir	2	8
PI	LPV	Lopinavir/Ritonavir	9	55
NRTI	AZT	Zidovudine	3	10
NRTI	D4T	Stavudine	1.5	2
NRTI	TDF	Tenofovir	1.5	4
NRTI	ABC	Abacavir	3	6
NRTI	3TC	Lamivudine	3	20
NRTI	DDI	Didanosine	1.5	2
NNRTI	EFV	Efavirenz	3	10
NNRTI	NVP	Nevirapine	3	10
NNRTI	ETR	Etravirine	3	10

Table 4.1: *The lower and upper phenotypic fold resistance cut-off values for all drugs as reported on Stanford HIV DB.*

4.3 Deep Learning Models for Prediction

The data munging and transformation tasks described in the above sections have created a numeric data set which can be used for training neural network models for the drug resistance classification task. This section describes the construction and optimization of the neural network models to achieve a robust model for classification. The deep learning aspect of the neural network model lies in the number of hidden layers used in the network architecture to achieve enhanced accuracy in classification.

4.3.1 Artificial Neural Networks

The combination of mutations in the genomic sequences and the impacted phenotypic fold resistance values are non-linear in nature. While a mutation may increase

the fold resistance towards a particular antiretroviral drug, the presence of that mutation with another mutation in a genomic sequence may decrease the fold resistance towards the same antiretroviral drug. Artificial Neural Networks (ANN) models help in uncovering these non-linear aspects of the relationship between mutations and phenotypic fold resistance. The architecture of an ANN is similar to that of the brain where neurons are substituted by nodes arranged in different layers [36]. A simple three layer architecture of a neural network can be seen in Figure 1 where an input layer, a single hidden layer and an output layer is seen.

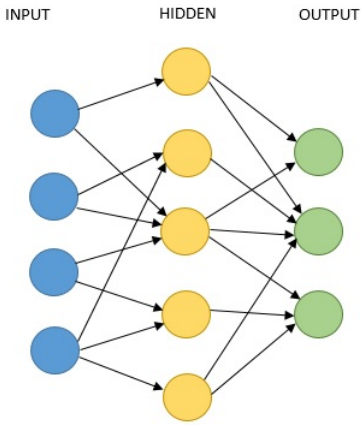


Figure 4.5: *A neural network showing the input, hidden and output layer.*

As seen in Figure 4.5, the leftmost layer in the diagram is called the input layer and the neurons are called input neurons. The rightmost layer in the network is called the output layer and the neuron called the output neurons. We can have one or many input / output neurons. The middle layer having neurons which are neither input nor output is termed as the hidden layer. A neural network architecture may have one or many hidden layers. In the experimental setup for this research we have a neural network model for each antiretroviral drug with varying number of hidden layers. The deep learning terminology of this research stems from the depth (number of layers + nodes in each layer) of the hidden layers used for each neural network model. The specifications of the neural networks we put to use and details about the same are explained in the following sections.

4.3.2 Architecture

The example of an artificial neural network provided in the previous section is trivial in nature. The classification of genomic sequences as resistant or susceptible to an antiretroviral drug requires processing a large number of inputs propagated through the multiple hidden layers [37] to reach the desired output while training. The inputs (features/predictors) in this context are the mutations at each codon and the output (target) is the phenotypic binary variable which indicates the sequence as resistant if 1 and susceptible if 0. The number of nodes in the input layer are 99 in case of models corresponding to protease inhibitors and 240 in case of models corresponding to reverse transcriptase inhibitors. The output layer has a single node for the binary variable.

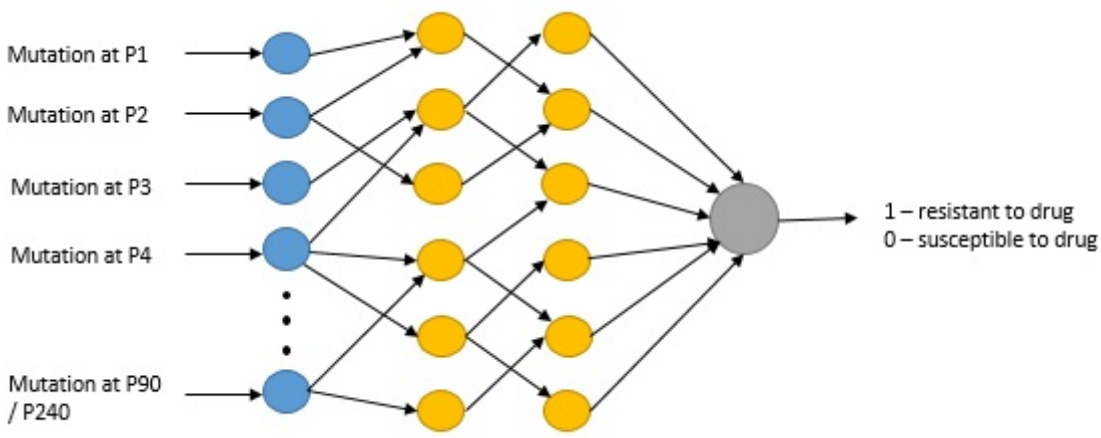


Figure 4.6: *Illustratory image of NN architecture for PI/RT models.*

4.3.3 Backpropagation Algorithm for training.

The development of artificial neural networks started from the creation of the perceptron [38], a system which determines the output based upon the inputs provided and the weights associated to these inputs. Further research to the development of perceptron in order to create networks capable of learning complex relations between input and output were undertaken and the formulation of the delta rule [39] for learning further accelerated the research around neural networks. The delta rule drives the learning phase by using the error i.e. the difference between the target

value and actual value. All nodes in the neural network except the nodes in the output layer have a weight associated with it. The input given to the nodes at the input layer are multiplied by their respective weights and passed to every node in the next layer. The node in the next layer receives the output from previous layer and passes it through an activation function before it is propagated further. This can be represented as,

$$S_j = \sum w_{ij} a_i \tag{4.1}$$

where S_j is the sum of the products of the weights and outputs of nodes from previous layer . The output/activation of the current node in layer j is,

$$a_j = f(S_j) \tag{4.2}$$

The function f here is the activation function. The activation function used in all the models is the sigmoid activation function given by,

$$f(x) = 1/(1+e^{-x}) \tag{4.3}$$

During the training of the network, we start with a set of arbitrary weights and the inputs are provided to the input nodes. These inputs are propagated with the weight values and activation function acts at each node to give the activation at the output node which is the value calculated by the network. The error is calculated at the output nodes based upon the known target value. The change of weights associated with the nodes in the preceding layers are adjusted based upon the calculated error. The error calculated in this case is the sum of squared errors (SSE). The adjustment of the weights is performed using gradient descent. The idea is that if a change in weight will increase the error then we want to decrease the weight and vice versa. This is done by taking the derivative of the error with respect to weight given by,

$$\delta_{ij} = -\lambda \frac{dE_i}{dw_{ij}} \tag{4.4}$$

The parameter λ here is the rate of learning which allows one to control the rate at which the errors are adjusted. While training of the neural network the training data is presented iteratively to the network for updating the weights. The weights are updated over every iteration as the error is propagated from the output node till the input node for adjusting the weights. The error in this case reduces after every iteration and finally converges. The below table summarizes the converged error, number of iterations till convergence and the number of hidden layer in each of the 17 neural network models. The number of iteration till convergence and the SSE is an outlier for the neural network model for Etravirine as the number of sequences available of traning are very low for this drug.

Type	Drug	Hidden Layers	λ	Iterations	SSE
PI	FPV - Fosamprenavir	3	0.1	110	276.43
PI	NFV - Nelfinavir	2	0.1	140	176.22
PI	ATV - Atazanavir	3	0.1	160	109.65
PI	DRV - Darunavir	3	0.1	170	47.34
PI	SQV - Saquinavir	3	0.1	130	183.63
PI	IDV - Indinavir	4	0.1	120	181.47
PI	LPV - Lopinavir	2	0.1	120	161.97
PI	TPV - Tipranavir	3	0.1	130	115.97
NRTI	AZT - Zidovudine	3	0.1	300	149.75
NRTI	D4T - Stavudine	4	0.1	300	235.44
NRTI	TDF - Tenofovir	4	0.1	360	197.44
NRTI	ABC - Abacavir	4	0.1	340	190.95
NRTI	3TC - Lamivudine	4	0.1	230	97.78
NRTI	DDI - Didanosine	4	0.1	290	403.07
NNRTI	EFV - Efavirenz	4	0.1	370	152.5
NNRTI	NVP - Nevirapine	4	0.1	390	175.17
NNRTI	ETR - Etravirine	4	0.1	1460	8.26

Table 4.2: *Model parameters for the neural netowrk models.*

4.4 Metrics, Model Validation and Results

The performance of the models constructed above is tested using metrics based upon the confusion matrix [40] shown below.

	PREDICTED: Susceptible	PREDICTED: Resistant
ACTUAL: Susceptible	True Negative (TN)	False Positive (FP)
ACTUAL: Resistant	False Negative (FN)	True Positive (TP)

Figure 4.7: *Confusion matrix to explain choice of metrics.*

4.4.1 Accuracy

Accuracy is the overall correctness of the model in classifying resistant sequences as resistant when they are actually resistant and susceptible sequences as susceptible when they are actually susceptible. The accuracy is calculated as,

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{4.5}$$

4.4.2 Sensitivity

The sensitivity also termed as the true positive rate tells how often the model predicts that the sequence is resistant to the antiretroviral drug when it is actually resistant. It is calculated as,

$$Sensitivity = \frac{TP}{TP+FN} \tag{4.6}$$

4.4.3 Specificity

The specificity also termed as the true negative rate tells how often the model predicts that the sequence is susceptible to the antiretroviral drug when it is actually susceptible. It is calculated as,

$$Specificity = \frac{TN}{TN+FP} \tag{4.7}$$

4.4.4 10-fold Cross Validation

The performance of the 17 neural network models in terms of the above metrics is reported based on 10-fold cross validation. In this technique the entire training data is divided into 10 folds. There are 10 instances of training the model. At each instance 9 folds are used for training the model and the 10th fold is used as test data

on which the metrics are calculated. At every instance the test fold is changed and the values for accuracy, sensitivity and specificity are recorded. The performance of the model is the average of the values of these metrics over the 10 instances.

4.4.5 Results

The below table summarizes the results of the 17 neural network models in terms of the metrics defined in the previous section after 10-fold cross validation.

Type	Drug	Accuracy	Sensitivity	Specificity
PI	FPV - Fosamprenavir	0.90	0.91	0.86
PI	NFV - Nelfinavir	0.93	0.89	0.92
PI	ATV - Atazanavir	0.90	0.88	0.89
PI	DRV - Darunavir	0.89	0.93	0.58
PI	SQV - Saquinavir	0.91	0.92	0.87
PI	IDV - Indinavir	0.92	0.91	0.90
PI	LPV - Lopinavir	0.92	0.91	0.90
PI	TPV - Tipranavir	0.86	0.89	0.73
	Average PI	**0.90**	**0.91**	**0.83**
NRTI	AZT - Zidovudine	0.84	0.83	0.81
NRTI	D4T - Stavudine	0.83	0.82	0.79
NRTI	TDF - Tenofovir	0.78	0.83	0.63
NRTI	ABC - Abacavir	0.81	0.79	0.80
NRTI	3TC - Lamivudine	0.86	0.78	0.89
NRTI	DDI - Didanosine	0.76	0.74	0.75
	Average NRTI	**0.81**	**0.80**	**0.78**
NNRTI	EFV - Efavirenz	0.87	0.87	0.85
NNRTI	NVP - Nevirapine	0.87	0.86	0.87
NNRTI	ETR - Etravirine	0.77	0.86	0.56
	Average NNRTI	**0.84**	**0.86**	**0.76**

Table 4.3: *Results in terms of accuracy, sensitivity and specificity of the models.*

4.5 The HIV Drug Resistance Predictor Tool

The insights gained from the above experiments have created a pathway to construct a tool capable of predicting if a given strain of an individuals HIV is resistant or susceptible to the 17 antiretroviral drugs taken into consideration in this research. A detailed description of the HIV Drug Predictor Tool is given below.

4.5.1 Tool Overview

At heart of the HIV Drug Resistance Predictor Tool is the genotype-phenotype correlation dataset which is a composite dataset of the genotype-phenotype data available from Stanford HIV DB and genotype-phenotype data available from NARIs internal database. The brain of the HIV Drug Resistance Predictor tool is a combination of 17 neural network models, each corresponding to one of the 17 antiretroviral drug, trained on the composite genotype-phenotype dataset. The model hyper-parameters constituting the number of hidden layers and rate of learning are tuned in order to achieve optimal accuracy, sensitivity and specificity values.

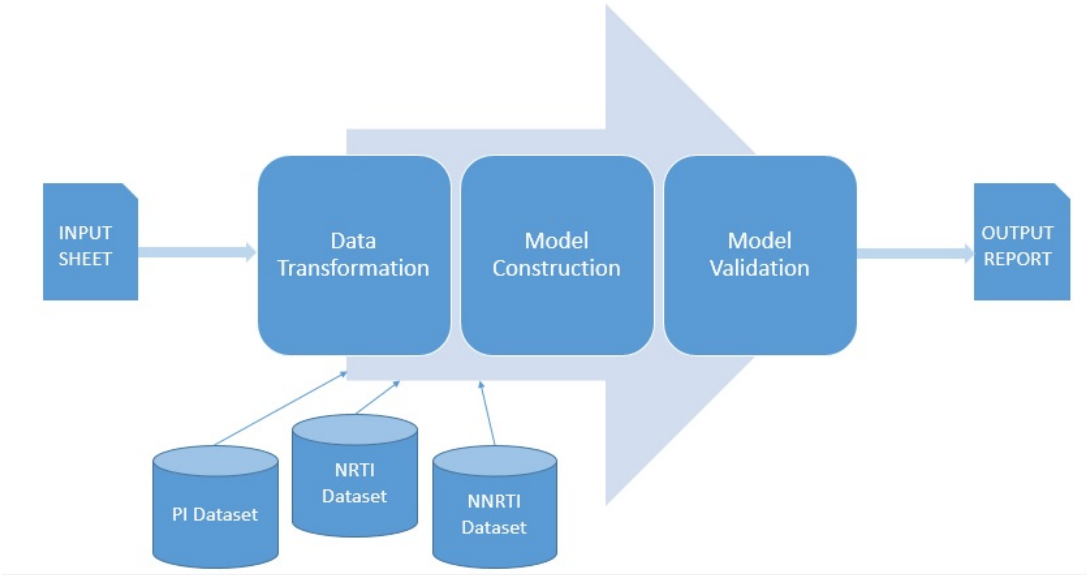


Figure 4.8: *Architectural Overview of HIV Drug Resistance Predictor Tool.*

The tool is essentially a script written in R, capable of reading input from an excel-sheet, transforming the data into a format which can be used to construct and train neural networks, optimizing the neural networks and using these neural networks to classify if the strain of an individuals HIV, which is characterized by the mutations provided in the input sheet, is resistant or susceptible to the antiretroviral drugs which are a part of this study. Finally, the tool provides the user with a comprehensive report about the drug resistance/susceptibility of the HIV strain under consideration along with other metrics which can allow clinicians to formulate

effective drug regimens. The theory and experimentation concerning the data transformation, neural network model construction and model validation has already been explained in the previous section. In the following sections we will discuss the significance of fields in the input sheet on model construction and summarize the fields provided in the output report and how this information can be beneficial to clinicians.

4.5.2 Input File

The input file serves two purposes. Firstly the input file allows the user to key in the genomic information about an individuals HIV strain based upon which the strain will be classified as resistant / susceptible to the 17 antiretroviral drugs. Secondly, the input file allows to select the specific subtype for which the neural network models are to be trained.

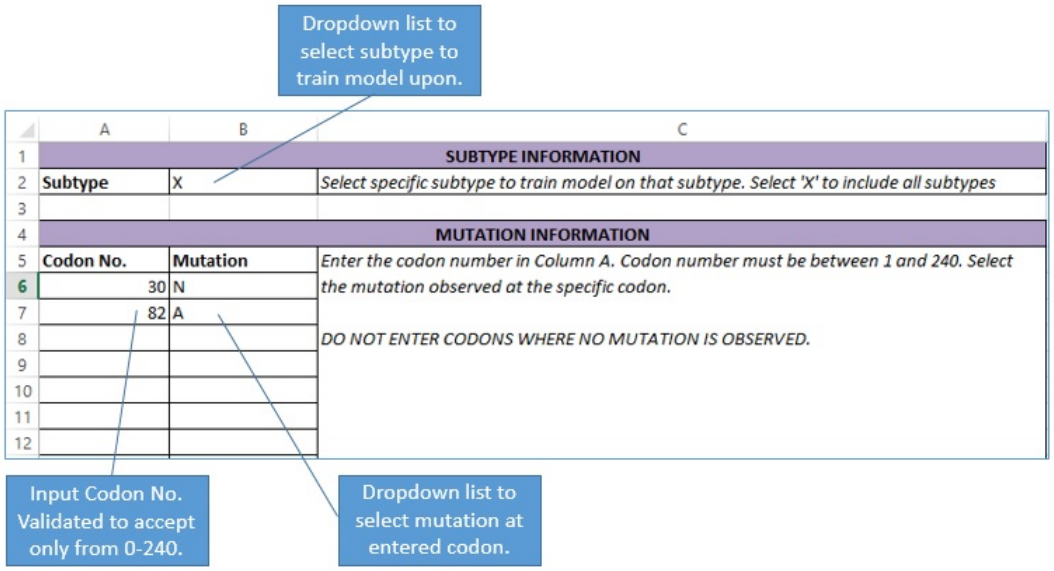


Figure 4.9: *Fields in the Input File.*

The top section of the input file is the Subtype Information section which contains a field labelled Subtype and allows the user to select one subtype from a dropdown containing the values A, B, C, D, F, G and X. Selection of any value apart from X indicates the training of the neural network models based upon genotype-phenotype information from the master dataset pertaining to that subtype. The

value X or no selection indicates training of the neural network models based upon the genotype-phenotype data of all subtypes. For instance, selecting value B in this dropdown will result in the system filtering the genotypic sequences belonging to subtype B and their corresponding phenotypic resistance values from the master data and only this filtered data will be used to train the neural network model.

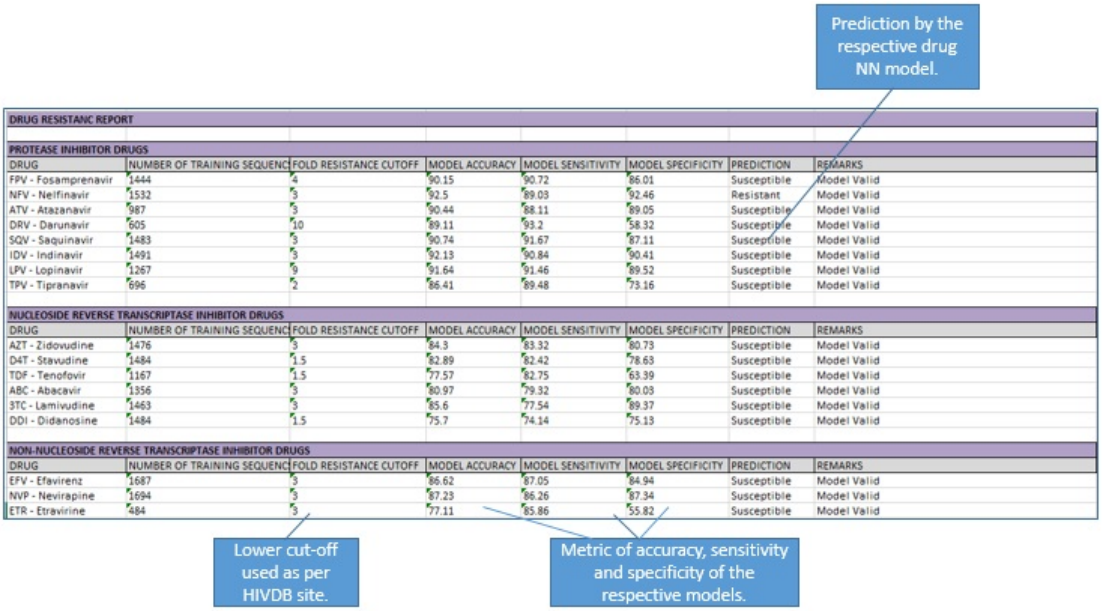
The other section titled Mutation Information allows the user to key in the mutation data of an individuals HIV strain. The data is to be entered in two columns. The first column allows the user to key in a numeric value from 1 to 240 which corresponds to the codon number in the genetic sequence of the individuals HIV. For every codon number entered an equivalent amino acid which signifies the mutation is to be selected from a dropdown list. The dropdown allows selection of the values each corresponding to an amino shown previously in table 2.3.

Only codons where a mutation is observed along with the observed mutation data (amino acid) need to be entered in the input sheet. Codons where no mutation is observed need not be entered in the input sheet. Codons where mixtures are observed will have multiple entries (length of the mixture) in the input sheet. For example a mixture of two amino acids at codon 30 needs to be represented by two separate entries in the mutation information section. The first entry will be codon 30 with the first amino acid in the mixture while the second entry will be codon 30 with the second amino acid observed in the mixture. A sample input file indicating an HIV strain with two mutations and specifying neural network training on all subtypes is seen in figure 4.9.

4.5.3 Drug Resistance Report

The Drug Resistance Report summarizes the predictions made by the neural network models for the input HIV strain (mutations) to each antiretroviral drug and the metrics of accuracy, sensitivity and specificity associated with these models. The number of training sequences used to train the neural network models are also specified along with the phenotypic cut-off resistance values which were adopted

for classification. In case of absence of sufficient training sequences for a neural network corresponding to a specific drug, a remark indicating the same is present in the remarks column of the report. A sample output report can be seen in the image below.



DRUG RESISTANC REPORT

PROTEASE INHIBITOR DRUGS

DRUG	NUMBER OF TRAINING SEQUENC	FOLD RESISTANCE CUTOFF	MODEL ACCURACY	MODEL SENSITIVITY	MODEL SPECIFICITY	PREDICTION	REMARKS
FPV - Fosamprenavir	1444	4	90.15	90.72	86.01	Susceptible	Model Valid
NFV - Nelfinavir	1532	3	92.5	89.03	92.46	Resistant	Model Valid
ATV - Atazanavir	987	3	90.44	88.11	89.05	Susceptible	Model Valid
DRV - Darunavir	605	10	89.11	93.2	58.32	Susceptible	Model Valid
SQV - Saquinavir	1483	3	90.74	91.67	87.11	Susceptible	Model Valid
IDV - Indinavir	1491	3	92.13	90.84	90.41	Susceptible	Model Valid
LPV - Lopinavir	1267	9	91.64	91.46	89.52	Susceptible	Model Valid
TPV - Tipranavir	696	2	86.41	89.48	73.16	Susceptible	Model Valid

NUCLEOSIDE REVERSE TRANSCRIPTASE INHIBITOR DRUGS

DRUG	NUMBER OF TRAINING SEQUENC	FOLD RESISTANCE CUTOFF	MODEL ACCURACY	MODEL SENSITIVITY	MODEL SPECIFICITY	PREDICTION	REMARKS
AZT - Zidovudine	1476	3	84.3	83.32	80.73	Susceptible	Model Valid
D4T - Stavudine	1484	1.5	82.89	82.42	78.63	Susceptible	Model Valid
TDF - Tenofovir	1167	1.5	77.57	82.75	63.39	Susceptible	Model Valid
ABC - Abacavir	1356	3	80.97	79.32	80.03	Susceptible	Model Valid
3TC - Lamivudine	1463	3	85.6	77.54	89.37	Susceptible	Model Valid
DDI - Didanosine	1484	1.5	75.7	74.14	75.13	Susceptible	Model Valid

NON-NUCLEOSIDE REVERSE TRANSCRIPTASE INHIBITOR DRUGS

DRUG	NUMBER OF TRAINING SEQUENC	FOLD RESISTANCE CUTOFF	MODEL ACCURACY	MODEL SENSITIVITY	MODEL SPECIFICITY	PREDICTION	REMARKS
EFV - Efavirenz	1687	3	86.62	87.05	84.94	Susceptible	Model Valid
NVP - Nevirapine	1694	3	87.23	86.26	87.34	Susceptible	Model Valid
ETR - Etravirine	484	3	77.11	85.86	55.82	Susceptible	Model Valid

Figure 4.10: *A glimpse of the Drug Resistance Summary Report.*

The sample report shows the predictions made by each of the models to the input provided in previous section. The mutations in the input strain are at codons 10 and 82 where the amino acids have changed to N (Asparagine) and A (Alanine) respectively. The report shows that the neural network model corresponding to Nelfinavir (NFV) has predicted that the input sequence with the given mutations is resistant to the drug. This model has an accuracy of 92.5%, sensitivity of 89.03% and a specificity of 92.46% based on which the clinician can safely exclude Nelfinavir while formulating an antiretroviral regimen for the individual with this HIV strain. Predictions made by neural network models corresponding to other antiretroviral drugs are also seen in the report. As per the predictions the HIV strain is susceptible to all other antiretroviral drugs. The clinician can choose the antiretroviral drugs for the regimen which the HIV strain is best susceptible to by considering the metrics associated with these models.

Chapter 5

General Conclusions

This chapter provides an overview of the entire research and highlights the tasks which have been accomplished. Avenues which have opened for further exploration based upon this research are discussed.

5.1 Summary

This research was aimed at exploring a novel technique to approach the drug resistance prediction task and developing a system based upon this approach to be used by the industry sponsor. A deep learning approach harnessing the power of artificial neural networks to uncover non-linear relationships between the mutations in an HIV genomic sequence and their impact on resistance to various antiretroviral drugs was adopted. Models constructed during experimentation have provided more than satisfactory results in terms of the metrics which were formulated to assess the robustness of these models. A classification accuracy of 90.50% was obtained for eight protease inhibitor antiretroviral drugs after 10-fold cross validation on the dataset. The accuracy reduced to 84.15% for three non-nucleoside reverse transcriptase inhibitor antiretroviral drugs and further to 81.30% for six nucleoside reverse transcriptase inhibitor antiretroviral drugs.

The data munging and transformation approach adopted has facilitated the inclusion of mixture mutations for the prediction task. This is a novel aspect of the study. The dataset assembled for this study as well as for the developed system can

be appended with new sequences and their corresponding phenotypic results whenever they are available in the future. The current dataset contains sequences for all HIV subtypes, however the system developed provides a mechanism to choose data belonging to a particular subtype for training the neural network models. This provision is vital for the industry sponsor due to the prevalence of HIV subtype C in India where the industry sponsor is based.

A system harnessing the power of the deep learning models created to provide a summary of resistance/susceptibility towards 17 antiretroviral drugs used in this study has been developed. Templates to provide the genomic details of a sequence with validations to control error have been designed and used. An unambiguous report providing the prediction made by the 17 models along with necessary metrics and other required information has been designed and put to use. This system will be used by clinicians at the industry sponsors field clinics to formulate potent antiretroviral drug regimens to control viral count in infected individuals.

5.2 Impact

There are more than 1 million people living with HIV infections in India. A majority of this population belong to an economically weaker section of the society and cannot afford sophisticated phenotypic diagnostic tests. The virtual phenotypic prediction system created in this research will allow field clinicians working with the industry sponsor to formulate drug regimens for the infected individuals bypassing the expensive phenotypic test. This will help the infected under-privileged population to live longer healthy lives. The research also lays the foundation for further work in this domain which can be undertaken by the industry sponsor.

5.3 Future Work

The classification accuracy achieved for reverse transcriptase inhibitor drugs is inferior to that of protease inhibitor drugs. A possible reason for the reduction in the accuracy is the number of highly complex mutations and mixture mutations prevalent in the reverse transcriptase sequences. Exploring further non-bioinformatics

based features based upon clinical history and regimen change can enhance the accuracy of the models related to these antiretroviral drugs.

The model currently classifies the resistance in two level: resistant and susceptible. The models can be enhanced further for classification in more than two levels. The lower as well as upper phenotypic cut-off values provided can be used to classify the resistance in three levels. This will provide further flexibility to clinicians in formulating antiretroviral drug regimens.

The experiments and the system built are based upon the genotype-phenotype data for 17 antiretroviral drugs. Currently there is no available genotype-phenotype data for entry inhibitors and integrase inhibitors. Availability of data for these antiretroviral drugs will allow replication of the models created in our research to accommodate the new antiretroviral with minimum changes to the code.

The inclusion of phylogenetic as well as clinical features along with the mutation details is another aspect to be explored. The inclusion of such features might help enhance the prediction accuracy.

Appendix A

Research Timeline

Activity Start Date	Activity End Date	Activity Description	Key deliverables
01-Jun-16	13-Jun-16	Domain / Requirement Analysis	Knowledge Transfer about HIV and drug resistance. Analysis of similar work / research. Formulating a research question / objectives. Detailed review of literature.
14-Jun-16	28-Jun-16	Data staging and munging	Finalizing datasets. Aggregating data from multiple sources. Data transformation.
29-Jun-16	10-Aug-16	Experimentation	Model construction. Model optimization. Validation and Testing.
11-Aug-16	15-Aug-16	Discussion and Conclusion	Final Results. Summarizing findings. Comparitive analysis.
16-Aug-16	2-Sep-16	Final Report Writing	Final Report and Review

Figure A.1: *A brief overview of the scheduled activities for the research and their date of completion.*

Appendix B

Comparison with existing classification algorithms.

A comparitive analysis with other classification algorithms is provided.

Drug	Our Accuracy	SVM	OLS Regression
Protease Inhibitors			
ATV	0.90	0.69	0.68
IDV	0.92	0.77	0.78
LPV	0.92	0.8	0.79
NFV	0.93	0.79	0.79
SQV	0.91	0.81	0.81
FPV	0.90	N/A	N/A
DRV	0.89	N/A	N/A
TPV	0.86	N/A	N/A
Nucleoside RT Inhibitors			
3TC	0.86	0.84	0.83
ABC	0.81	0.65	0.63
AZT	0.84	0.7	0.64
D4T	0.83	0.68	0.66
DDI	0.76	0.67	0.61
TDF	0.78	0.69	0.46
Nonnucleoside RT Inhibitors			
ETR	0.77	N/A	N/A
EFV	0.87	0.82	0.78
NVP	0.87	0.78	0.74

Table B.1: *Comparison of the accuracy of our models with Support Vector Machine (SVM) classification and Ordinary Least Squares (OLS) Regression.*

Tables show the accuracy of various statistical learning methods used for the classification task and the accuracy achieved by the methedology in this research. The datasets used are almost consistent with the exception of few more sequecnes provided by the industry sponsor in this research. The results used are the ones based on the complete set of mutations in the sequences [41].

Drug	Our Accuracy	LARS	Decision Trees
Protease Inhibitors			
ATV	0.90	0.76	0.71
IDV	0.92	0.77	0.75
LPV	0.92	0.83	0.77
NFV	0.93	0.8	0.76
SQV	0.91	0.82	0.75
FPV	0.90	N/A	N/A
DRV	0.89	N/A	N/A
TPV	0.86	N/A	N/A
Nucleoside RT Inhibitors			
3TC	0.86	0.88	0.9
ABC	0.81	0.77	0.69
AZT	0.84	0.76	0.7
D4T	0.83	0.78	0.75
DDI	0.76	0.75	0.74
TDF	0.78	0.7	0.68
Nonnucleoside RT Inhibitors			
ETR	0.77	N/A	N/A
EFV	0.87	0.87	0.84
NVP	0.87	0.87	0.91

Table B.2: *Comparison of the accuracy of our models with Least Angle Regression (LARS) and Decision Trees Classification.*

Bibliography

[1] ”HIV AIDS Fact Sheet.” *WHO.* n.p. Jul. 2016. Web. *http://www.who.int/mediacentre/factsheets/fs360/en/*

[2] ”A history of AIDS: looking back to see ahead.” *Greene WC.* Eur J Immunol. Nov. 2007. *http://www.ncbi.nlm.nih.gov/pubmed/17972351/*

[3] ”Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS.” *Popovic M, Sarngadharan MG, Read E, Gallo RC* Science. May 1984. *http://www.ncbi.nlm.nih.gov/pubmed/6200935/*

[4] ”Isolation of a new human retrovirus from West African patients with AIDS.” *Clavel F, Gutard D, Brun-Vzinet F, Chamaret S, Rey MA, Santos-Ferreira MO, Laurent AG, Dauguet C, Katlama C, Rouzioux C* Science. Jul 1986. *http://www.ncbi.nlm.nih.gov/pubmed/2425430/*

[5] ” Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes.” *Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM, Sharp PM, Hahn BH* Nature. Feb. 1999 *http://www.ncbi.nlm.nih.gov/pubmed/9989410/*

[6] ” An African primate lentivirus (SIVsm) closely related to HIV-2.” *Hirsch VM, Olmsted RA, Murphey-Corb M, Purcell RH, Johnson PR* Nature. Jun. 1989. *http://www.ncbi.nlm.nih.gov/pubmed/2786147/*

[7] ”Sixteen years of HIV surveillance in a West African research clinic reveals divergent epidemic trends of HIV-1 and HIV-2.” *Van der Lo-*

eff MF, Awasana AA, Sarge-Njie R, van der Sande M, Jaye A, Sabally S, Corrah T, McConkey SJ, Whittle HC Int J Epidemiol. Oct. 2006. *http://www.ncbi.nlm.nih.gov/pubmed/16543363/*

[8] "Global Health Observatory (GHO) data." *WHO.* n.p. 13 Jul. 2016. Web. *http://www.who.int/gho/hiv/en/*

[9] " The first postmodern pandemic: 25 years of HIV/ŁAIDS." *L. O. Kallings.* Journal of Internal Medicine. Jan. 2008. *http://onlinelibrary.wiley.com/doi/10.1111/jim.2008.263.issue-3/issuetoc*

[10] "AIDS and Macroeconomic Impact." *Robert Greener.* International AIDS-Economics Network. July 2002. *http://pdf.usaid.gov/pdf_docs/PNACP969.pdf*

[11] "AIDS Stigma and sexual prejudice." *Herek GM, Capitanio JP.* American Behavioral Scientist. 1999. *http://facultysites.dss.ucdavis.edu/ gmherek/rainbow/html/abs99_sp.pdf*

[12] "General Assembly adopts Declaration of Commitment on HIV/AIDS. Global Crisis - Global Action." *United Nations special session on HIV/AIDS* "A call to action" from UN Secretary-General in fight against HIV/AIDS. Jun. 2001. *http://www.un.org/ga/aids/conference.html*

[13] "HIV/AIDS Treatment and Care." *WHO.* n.p. 13 Jul. 2016. Web. *http://www.who.int/hiv/topics/treatment/en/*

[14] "Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection. Recommendations for a public health approach - Second edition. Chapter 4." *WHO.* Jun. 2016. *ISBN: 978 92 4 154968 4*

[15] "The HIV Life Cycle." *POZ Magazine* n.p. Feb. 2016. *https://www.poz.com/basics/hiv-basics/hiv-life-cycle*

[16] "HIV-1 Protease and Reverse Transcriptase Control the Architecture of Their Nucleocapsid Partner" *Gilles Mirambeau, Sbastien Lyonnais, Dominique*

Coulaud, Laurence Hameau, Sophie Lafosse, Josette Jeusset, Isabelle Borde, Michle Reboud-Ravaux, Tobias Restle, Robert J. Gorelick, Eric Le Cam. Aug. 22, 2007. *http://dx.doi.org/10.1371/journal.pone.0000669*

[17] "HIV entry inhibitors: mechanisms of action and resistance pathways." *Journal of Antimicrobial Chemotherapy* Oxford Journals Apr. 2006. *http://jac.oxfordjournals.org/content/57/4/619.full*

[18] "Inhibition of HIV-1 by fusion inhibitors." *Eggink D1, Berkhout B, Sanders RW.* PubMed 2010. *http://www.ncbi.nlm.nih.gov/pubmed/21128887*

[19] "HIV treatments directory - Reverse transcriptase inhibitors" *aidsmap.* n.p. Jul. 2016. Web. *http://www.aidsmap.com/Reverse-transcriptase-inhibitors/page/1729413/*

[20] "Raltegravir: first in class HIV integrase inhibitor." *Zelalem Temesgen. Dawd S Siraj.* PubMed Apr. 2008. Web. *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2504063/*

[21] "HIV treatments directory - Protease inhibitors" *aidsmap.* n.p. Jul. 14, 2016. Web. *http://www.aidsmap.com/Protease-inhibitors/page/1729414/*

[22] "AIDSinfo - Education Materials. HIV Treatment FDA-Approved HIV Medicines." (Last updated 5/24/2016; last reviewed 3/1/2016)" *AIDSinfo.* May. 24,2016. Jul. 14, 2016. Web. *https://aidsinfo.nih.gov/education-materials/fact-sheets/21/58/fda-approved-hiv-medicines*

[23] "Antiretroviral Therapy for HIV Infection." *R Chris Rathbun* Apr. 06, 2016. Jul. 14, 2016. Web. *http://emedicine.medscape.com/article/1533218-overview*

[24] "Molecular basis of human immunodeficiency virus type 1 drug resistance: overview and recent developments." *Menndez-Arias.* Apr. 2013. *Antiviral Res. 98(1):93-120.*

[25] "HIV Drug Resistance." *Franois Clavel, M.D., and Allan J. Hance, M.D.* The New England Journal of

Medicine. 2004. *http://www.usherbrooke.ca/microbiologie-infectiologie/fileadmin/sites/microbiologie-infectiologie/documents/Cours-residents/04-11-2009/APP1.pdf*

[26] "The Genetic Basis of HIV-1 Resistance to Reverse Transcriptase and Protease Inhibitors." *Robert W. Shafer, Rami Kantor and Matthew J. Gonzales.* 2000. *AIDS Rev 2000; 2: 211-228*

[27] "A genotypic drug resistance interpretation algorithmthat significantly predicts therapy response in HIV-1-infected patients." *Kristel Van Laethem et al.* International Medical Press. *Antiviral therapy 7(2):123-9. July 2002.*

[28] "HIV-1 genotypic drug resistance interpretations algorithms." *National Agency for AIDS Research.* online. *http://www.hivfrenchresistance.org/index.html*

[29] "HIV-1 genotypic drug resistance interpretations algorithms. TABLES OF RULES - SEPTEMBER 2015" *National Agency for AIDS Research.* online. *http://www.hivfrenchresistance.org/table.html*

[30] "Human immunodeficiency virus reverse transcriptase and protease sequence database." *Soo-Yon Rhee, Matthew J. Gonzales, Rami Kantor, Bradley J. Betts, Jaideep Ravela, and Robert W. Shafer.* Nucleic Acids Res. 2003 Jan 1; 31(1): 298303. *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC165547/*

[31] "Comparative evaluation of three computerized algorithms for prediction of antiretroviral susceptibility from HIV type 1 genotype." *Zazzi, M., et al.* Journal of Antimicrobial Chemotherapy, 2004. 53(2): p. 356-360. *https://jac.oxfordjournals.org/content/53/2/356?related-urls=yes&legid=jac;53/2/356*

[32] "Detecting and understanding combinatorial mutation patterns responsible for HIV drug resistance." *Jing Zhanga, Tingjun Houb, Wei Wangc, and Jun S. Liua.* Harvard University DASH Repository. 2010. *https://dash.harvard.edu/bitstream/handle/1/14169384/49700694.pdf?sequence=1*

[33] ”Prediction of HIV-1 drug susceptibility phenotype from the viral genotype using linear regression modeling.” *Vermeiren H. et. al.* Journal of Virological Methods. 2007. *http://www.ncbi.nlm.nih.gov/pubmed/17574687*

[34] ”Methods for Investigation of the Relationship between Drug-Susceptibility Phenotype and Human Immunodeficiency Virus Type 1 Genotype with Applications to AIDS Clinical Trials Group 333.” *Sevin. et. al.* The Journal of Infectious Diseases. 2000. *http://www.ncbi.nlm.nih.gov/pubmed/10882582*

[35] ”Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes.” *Beerenwinkel. et. al.* Nucleic Acid Research. 2003. *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC168981/*

[36] ”Artificial Intelligence (2nd ed).” *Elaine Rich and Kevin Knight* McGraw Hill. 1991. *ISBN 0-07-100894-2*

[37] ”Capabilities of a four-layered feedforward neural network: four layers versus three.” *Tamura, S., and Tateishi, M.* IEEE Transactions on Neural Networks. 1997.

[38] ”Perceptrons.” *Minsky, M., and Papert, S.* MIT Press, Cambridge. 1969.

[39] ”An Introduction to Neural Networks” *Kevin Gurney* UCL Press. 1997. *ISBN 0203451511*

[40] ”Selecting and interpreting measures of thematic classification accuracy” *Stehman, Stephen V.* Elsevier Inc. 1997. *http://www.sciencedirect.com/science/article/pii/S0034425797000837*

[41] ”Genotypic predictors of human immunodeficiency virus type 1 drug resistance.” *Soo-Yon Rhee, Jonathan Taylor, Gauhar Wadhera, Asa Ben-Hur, Douglas L. Brutlag, and Robert W. Shafer* Stanford HIV DB. 2005. *http://hivdb.stanford.edu/pages/pdf/Rhee.PNAS2006.pdf*