

MACHINE LEARNING APPROACHES TO UNDERSTAND HIV DRUG RESISTANCE

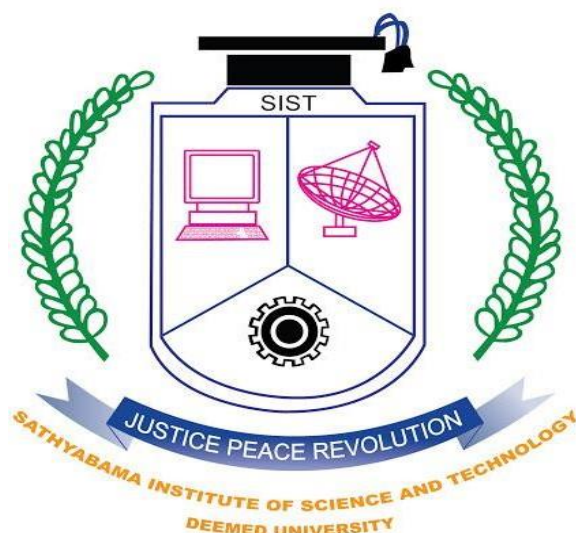
Submitted in partial fulfillment of the requirements for
the award of

Bachelor of Technology Degree in BIOINFORMATICS

by

TANIYA KANGSHA BANIK (Register No.37220024)

ARPAN JOSHI (Register No. 37220001)



DEPARTMENT OF BIOINFORMATICS

SCHOOL OF BIO AND CHEMICAL ENGINEERING

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,

CHENNAI - 600 119

APRIL 2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

www.sathyabama.ac.in



DEPARTMENT OF BIOINFORMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **TANIYA KANGSHA BANIK** (Reg. No. 37220024) and **ARPAN JOSHI** (Reg. No. 37220001) who carried out the project entitled "**MACHINE LEARNING APPROACHES TO UNDERSTAND HIV DRUG RESISTANCE** " under our supervision from December 2020 to April 2021.

Internal Guide

Dr. DANIEL ALEX ANAND

Head of the Department

Dr. JEMMY CHRISTY

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

We **TANIYA KANGSHA BANIK** (Reg. No. 37220024) and **ARPAN JOSHI** (Reg. No. 37220001) hereby declare that the Project Report entitled “**MACHINE LEARNING APPROACHES TO UNDERSTAND HIV DRUG RESISTANCE**” done by us under the guidance of **Dr. DANIEL ALEX ANAND** is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in Bioinformatics.

1.

2.

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATES

ACKNOWLEDGEMENT

We are pleased to acknowledge our sincere thanks to Board of Management of **SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement in doing this project and for completing it successfully. We are grateful to them.

We convey our thanks to **Dr. JEMMY CHRISTY MSC., MTECH., (PhD)., and Dr. SWETHA SUNKAR MSC., MTECH., (PhD)., Head of the Department Dept. of Bioinformatics** for providing us necessary support and details at the right time during the progressive reviews.

We would like to express our sincere and deep sense of gratitude to our Project Guide **Dr. DANIEL ALEX ANAND PhD., Dept of Bioinformatics** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of our project work.

We wish to express our thanks to all Teaching and Non-teaching staff members of the **Department of Bioinformatics** who were helpful in many ways for the completion of the project.

ABSTRACT

Human immunodeficiency virus (HIV) is one of the fastest developing pathogens known. The human immunodeficiency virus (HIV)/AIDS pandemic is one of the most important challenges facing humanity. More than 36 million people are currently living with HIV. The prediction of HIV resistance to a drug or a combination of drugs is an important issue for the development of new potent and safe antiretroviral drugs. The recommended treatment for HIV is a combination of three antiretroviral (ARV) drugs from two or more different drug groups and is known as highly active antiretroviral therapy (HAART). The Stanford HIV Drug Resistance Database (HIVDB) is an essential resource for public health officials monitoring ADR and TDR, for scientists developing new ARV drugs, and for HIV care providers managing patients with HIVDR. Developing a computational prediction of drug resistance phenotype will enable efficient and timely selection of the best treatment regimens. Machine learning methods are widely and successfully used in new drug discovery. Using machine learning approach that takes advantage of the hierarchical nature in the relationships among the sequence, structure, and function, an integrative analysis of these trajectories reveals key details of the resistance mechanism, including changes in the protein structure, hydrogen bonding, and protein-ligand contacts. Machine learning algorithms can learn from electronic registry data and help to predict HIV status with a fairly high level of accuracy. Machine learning algorithms are being implicated here to measure the mutation over genes of the affected person. Machine learning approach, like by using tool it's easier to interpret drug resistance dataset and predicting resistance scores of drugs and actually find out the best drug with less resistibility which performs better in the treatment of HIV. Mutation selection here plays a major role to categorize the inputs of the model and also helps to calculate the drug resistance for each mutation in the model.

Keywords: HIV-AIDS, HIV drugs, Antiretroviral therapy, Drug resistance, Machine learning techniques

INDEX

S.NO	TITLE	PAGE NO
1	CHAPTER 1- INTRODUCTION 1.1. OUTLINE OF THE PROJECT 1.1.1. HIV AIDS 1.1.2. HIV-DRUGS 1.1.3. DRUG RESISTANCE 1.1.4 HIV DRUG RESISTANCE PREDICTION 1.2. OUTSOURCE 1.2.1. TOOLS AND SOFTWARE USED TO PREDICT 1.2.2. HIV DRUG RESISTANCE STANFORD DATASET 1.2.3. MACHINE LEARNING APPROACH IN PREDICTION	1-4
2	CHAPTER 2- LITERATURE SURVEY 2.1. REVIEW-LITERATURE WORK 2.1.1. WHAT IS DRUG RESISTANCE IN GENERAL 2.1.2. HOW DOES A DRUG GET RESISTANT PERIODICALLY 2.1.3. STUDY DRUG RESISTANCE DATABASE ONLINE 2.1.4. WHAT MACHINE LEARNING APPROACH USE TO PREDICT DRUG RESISTANCE	5-14
3	CHAPTER 3- AIM AND OBJECTIVES	15

4	CHAPTER 4- METHODOLOGY 4.1. HIV DRUG RESISTANCE DATABASES USED 4.1.1. STANFORD HIV DRUG RESISTANCE DATABASE (HIVDB) 4.1.2. HIV DRUG RESISTANCE DATABASE 4.1.3. WHO HIV DRUG RESISTANCE DATABASE 4.2. MACHINE LEARNING APPROACH 4.2.1. FUZZY LOGIC TOOLBOX 4.2.1.1. TRAINING DATA 4.2.1.2. GENERATE OR LOAD FIS STRUCTURE 4.2.1.3. INTERACTIVELY SPECIFY FIS STRUCTURE 4.2.1.4. VIEW FIS STRUCTURE 4.2.1.5. TRAIN FIS 4.2.1.6. SURFACE VIEWER 4.2.1.7. RULE VIEWER	16-26
5	CHAPTER 5-RESULT AND DISCUSSION	27-50
6	CHAPTER 6- SUMMARY AND CONCLUSION 6.1. SUMMARY 6.2. CONCLUSION	51-53
6	REFERENCES	54

LIST OF FIGURES

FIGURE No.	FIGURE NAME	PAGE
2.1	HIV Drug resistance process in general	6
2.2	Online databases as reference	8
4.1	Source of the dataset	16
4.2	MATLAB Command Window	18
4.3	'anfisedit' command opens ANFIS Designer window	19
4.4	Loading data from a file	20
4.5	Data loaded in Neuro- Fuzzy Designer	21
4.6	Loading FIS Structure	22
4.7	View of model structure	23
4.8	Trained FIS	24
4.9	Surface viewer of NFV	25
4.10	Rule viewer represents the drug resistance scores	26
5.1	Excel sheet of protease table with 8 drugs	27
5.2	Excel sheet with 46 mutation and 8 drugs	28
5.3	Final excel sheet for analysis	28
5.4	Rule viewer of different dataset with NVF	29
5.5	Input data	29
5.6	Plot points	30
5.7	Rule viewer of different dataset with SQV	30
5.8	Different input values	31

5.9	Plot points	31
5.10	Rule viewer of different dataset with IDV	31
5.11	Different input values	32
5.12	Plot points	32
5.13	Rule viewer of different dataset with FPV	32
5.14	Different input values	33
5.15	Plot points	33
5.16	Rule viewer of different dataset with ATV	33
5.17	Different input values	34
5.18	Plot points	34
5.19	Rule viewer of different dataset with LPV	34
5.20	Different input values	35
5.21	Plot points	35
5.22	Rule viewer of different dataset with TPV	35
5.23	Different input values	36
5.24	Plot points	36
5.25	Rule viewer of different dataset with DRV	36
5.26	Different input values	37
5.27	Plot points	37
5.28	Surface plot of NFV	38
5.29	Surface plot of SQV	38
5.30	Surface plot of IDV	39
5.31	Surface plot of FPV	39
5.32	Surface plot of ATV	40
5.33	Surface plot of LPV	40
5.34	Surface plot of TPV	41
5.35	Surface plot of DRV	41
5.36	Excel sheet of protease table with 4 drugs after grouping	42
5.37	Final excel sheet for analysis	43
5.38	Rule viewer of different dataset with NFV	44
5.39	Different input values	44
5.40	Plot points	44

5.41	Rule viewer of different dataset with SQV	45
5.42	Different input values	45
5.43	Plot points	46
5.44	Rule viewer of different dataset with IDV	46
5.45	Different input values	46
5.46	Plot points	47
5.47	Rule viewer of different dataset with FPV	47
5.48	Different input values	47
5.49	Plot points	48
5.50	Surface plot of NFV	48
5.51	Surface plot of SQV	49
5.52	Surface plot of IDV	49
5.53	Surface plot of FPV	50

ABBREVIATION

HIV	- Human Immunodeficiency Virus
AIDS	- Acquired Immunodeficiency Syndrome
WHO	- World Health Organization
ART	- Anti Retroviral Therapy
NRTI	- Nucleoside Reverse Transcriptase Inhibitor
NNRTI	- Non-Nucleoside Reverse Transcriptase Inhibitor
INSTI	- Integrase Strand Transfer Inhibitor
PI	- Protease Inhibitor
SHDRDB	- Stanford HIV Drug Resistance Database
ADR	- Acquired Drug Resistance
TDR	- Transmitted Drug Resistance
PDR	- Pre-treatment Drug Resistance
FIS	- Fuzzy Inference System
MF	- Membership Function
NFV	- Nelfinavir
SQV	- Saquinavir
IDV	- Indinavir
FPV	- Fosamprenavir
ATV	- Atazanavir
LPV	- Lopinavir
TPV	- Tipranavir
DRV	- Darunavir
APV	- Amprenavir
RTV	- Ritonavir

CHAPTER-1

INTRODUCTION

1.1. OUTLINE OF THE PROJECT

1.1.1. HIV AIDS:

HIV is a virus that damages the immune system. The immune system helps the body fight off infections. Untreated HIV infects and kills CD4 cells, which are a type of immune cell called T cells. Over time, as HIV kills more CD4 cells, the body is more likely to get various types of infections and cancers.

Without treatment, a person with HIV is likely to develop a serious condition called AIDS. AIDS is a disease that can develop in people with HIV. It's the most advanced stage of HIV. However, treatment with antiretroviral drugs can prevent AIDS from developing.

1.1.2. HIV-DRUGS:

There's currently no known cure for HIV. However, medications can help people living with HIV manage the condition and lead healthy lives. These drugs work by stopping HIV from replicating.

The method used to treat and fight off the HIV infection is ARV (Anti Retro Viral therapy). There are six main types ('classes') of antiretroviral drugs. Each class of drug attacks HIV in a different way. Generally, drugs from two (or sometimes three) classes are combined to ensure a powerful attack on HIV.

Antiretroviral drugs are listed below with example:

a) Nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs):

Example: Abacavir, Lamivudin, Emtricitabine.

b) Non-nucleoside reverse transcriptase inhibitors (NNRTIs):

Example: Doravirine, Etravirine, Nevirapine.

c) *Integrase inhibitors:*

Example: Bictegravir

.

d) *Entry inhibitors:* Entry inhibitor stops HIV from entering human cell,

There are two types- CCR5 inhibitor, CD4 inhibitor.

Example: Maraviroc.

e) *Protease inhibitor:*

Example: Atazanavir.

f) *Booster drugs:* Booster drugs are used to boost the effect of protease inhibitor. A small dose of booster drug effectively works good. It stays in the body for long time or at higher level. Without booster drug, the primary drugs are ineffective.

1.1.2.1. Protease Inhibitor:

Antiretroviral drugs are designed to treat HIV. Different drugs have different mechanisms of action. Protease inhibitors work by blocking the activity of HIV protease, which is an enzyme that HIV needs to multiply.

List of protease inhibitor drugs

- **Nelfinavir (NFV)**
- **Saquinavir (SQV)**
- **Indinavir (IDV)**
- **Fosamprenavir (FPV)**
- **Atazanavir (ATV)**
- **Lopinavir (LPV)**
- **Tipranavir (TPV)**
- **Darunavir (DRV)**

1.1.2.2. Nucleoside reverse transcriptase inhibitors:

Nucleoside reverse transcriptase inhibitors (NRTIs) block reverse transcriptase (an HIV enzyme). HIV uses reverse transcriptase to convert

its RNA into DNA (reverse transcription). Blocking reverse transcriptase and reverse transcription prevents HIV from replicating.

List of Nucleoside reverse transcriptase inhibitors drugs

- Zidovudine (AZT)
- Tenofovir disoproxil fumarate (TDF)
- Abacavir sulfate (ABC)
- Lamivudine (3TC)

1.1.3. DRUG RESISTANCE:

Drug resistance is the reduction in effectiveness of a medication to cure a disease or condition.

- HIV DRUG RESISTANCE occurs when the virus starts to make changes (mutations) to its genetic make-up (RNA) that are resistant to certain HIV drugs, or classes of HIV drugs.
- Once a person has HIV, the virus begins to multiply in the body. As HIV multiplies, it sometimes changes form (mutates). Some HIV mutations that develop while a person is taking HIV medicines can lead to drug-resistant strains of HIV.
- Once drug resistance develops, HIV medicines that previously controlled the person's HIV are no longer effective. In other words, the HIV medicines can't prevent the drug-resistant HIV from multiplying. Drug resistance can cause HIV treatment to fail.

1.1.4 HIV DRUG RESISTANCE PREDICTION:

Which is briefing the process of prediction where drug resistance scores occurs of each drugs taken and predicted scores provide the concept of the better drug to be used against HIV infection. Resistance scores in this HIV drug resistance prediction process is in numerical format.

1.2. OUTSOURCE:

1.2.1. TOOLS AND SOFTWARE USED TO PREDICT:

The tool here used for HIV drug resistance prediction is ANFIS algorithm present in the MATLAB software, package of variety of tools. ANFIS is one of the good algorithm to analyse and predict.

1.2.2. HIV DRUG RESISTANCE-STANFORD DATASET:

1453 patterns of major drug resistance mutations from 163680 clinical isolates obtained from 155477 persons infected with HIV-1 virus are arranged in decreasing prevalence in the table (updated on 2018). For each pattern, drug susceptibility data on isolates matching the specific mutation pattern is shown on the left side. On the right side shown the fold resistance scores of drugs according to the mutation patterns.

1.2.3. MACHINE LEARNING APPROACH IN PREDICTION:

- An adaptive neuro-fuzzy inference system or adaptive network-based fuzzy inference system (ANFIS) is a kind of artificial neural network that is based on Takagi–Sugeno fuzzy inference system.
- The algorithm used, ANFIS help out to predict the drug susceptibility datas in the dataset.
- When using the ANFIS function, create or load the input data and pass it to the training Data input argument.
- When using Neuro-Fuzzy Designer, in the Load data section, select Training, and then: To load data from a file, select file. To load data from the MATLAB workspace, select workspace.

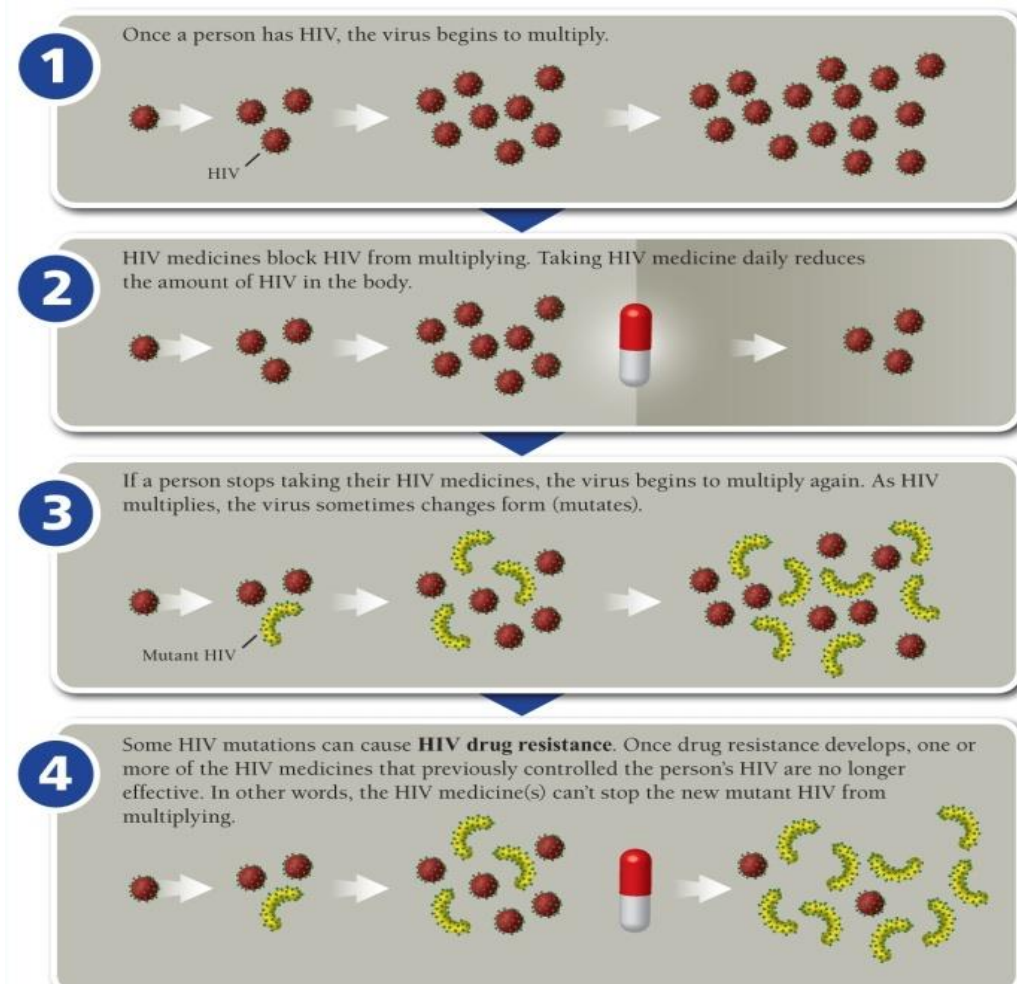
CHAPTER-2

LITERATURE SURVEY

2.1. REVIEW-LITERATURE WORK:

2.1.1. What is Drug Resistance in general:

When a bacteria, virus, or other microorganism mutates (changes form) and becomes insensitive to (resistant to) a drug that was previously effective. When a drug is resistant cause of microorganism stops the activity of this so, the drug can't be effective towards the microorganisms (virus, bacteria etc). Drug resistance can be a cause of HIV treatment failure. Drug resistance process is shown below in figure – 2.1.



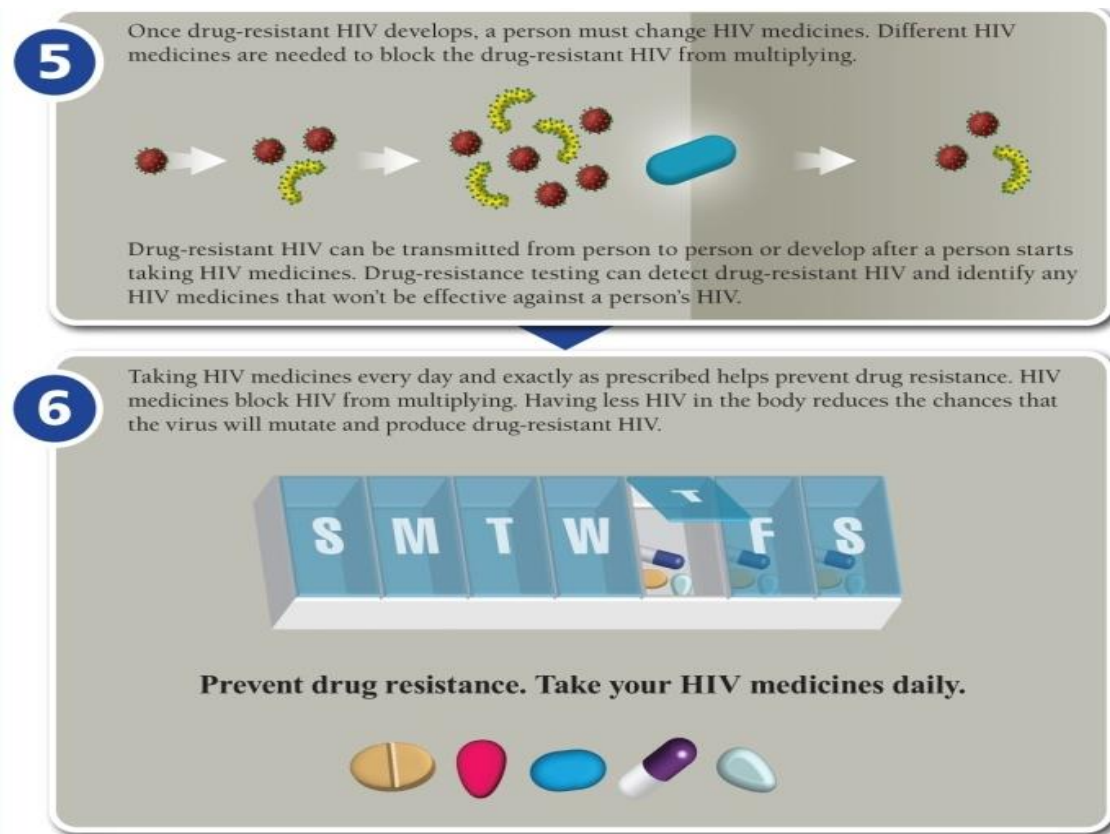


Figure – 2.1: HIV Drug resistance process in general

2.1.2. How does a drug get resistant periodically:

- Drug resistance can happen when HIV changes, or mutates, in a person's body. Viruses do this to try and survive. When it mutates, it may become resistant to the medicines a person is taking. It is the reduction in effectiveness of a drug in curing a disease or condition.
- There are two types of virus particle in a infected person's body – wild type (non-mutated virus particle) and mutated virus particle. Mutation confers resistant to antiretroviral drugs so that the virus can replicate in the presence of ARV.
- Wild type of viruses are more fit to replicate than mutant virus.
- If the HIV viral load is increased under treatment, then HIV can replicate under selective condition and can develop resistance to treatment.
- In some cases, a person may contract a drug-resistant strain of HIV. This means the type of HIV they have does not respond to certain medications.

- Drug resistance is more likely in certain condition like treatment with less than three drugs, inappropriate selection of drugs, adding one drug to a failing regimen.

2.1.3. Study drug resistance database online:

Since the AIDS was pandemic first identifies in the early 1980s, HIV has killed more than 30 million individuals. HIV research ushered in a golden age of antiviral drug development. National HIV treatment programs saved millions of lives, prevented millions of infections. Highly effective antiretroviral (ARV) drugs that inhibit HIV replication have been developed and made available at cost in many parts of the world. However, one of the main challenges in treating HIV it mutates at high rates and develops drug resistance. This poses an omnipresent threat of acquired HIV drug resistance (HIVDR) in persons receiving treatment. Both acquired (ADR) and transmitted drug resistance (TDR) increase the risk of treatment failure.

WHO is already developed five-year global action plan for 2017-2021 to support a coordinated international effort to prevent, monitor and respond to the emergence of HIV drug resistance, and to strengthen country efforts to achieve the global HIV targets. WHO recommends ART for all people with HIV as soon as possible after diagnosis without any restrictions of CD4 counts. It also recommends offer of pre-exposure prophylaxis to people at substantial risk of HIV infection as an additional prevention choice as part of comprehensive prevention. As a reference HIV Drug Resistance report 2019 made by scientist declared the concept of ADR (Acquired HIV drug resistance), TDR (Transmitted HIV drug resistance), PDR (Pre-treatment HIV drug resistance).

The Stanford HIV Drug Resistance Database (HIVDB) is an essential resource for public health officials monitoring different drug resistance, for scientists developing new ARV drugs, and for HIV care providers managing patients with HIVDR. The presence of HIVDR prior to the start of an ARV-treatment regimen is a good predictor of the success or failure of that regimen. HIVDB contains a free publicly available genotypic resistance interpretation system that interpretation system has been integrated into the workflows of the World Health Organization's HIV drug resistance network, the NIH-funded AIDS Clinical Trials Group, and several hundred research and more worldwide.

SHDRDB collaborated with researchers from around the world to track global trends in drug resistance and to characterize the genetic mechanisms of HIVDR to the ARVs for which more data are needed including those belonging to the nucleoside RT inhibitor (NRTI), nonnucleoside RT inhibitor (NNRTI), protease inhibitor (PI), and integrase strand transfer inhibitor (INSTI) class of drugs.

1453 patterns of major drug resistance mutations from 163680 clinical isolates obtained from 155477 persons infected with HIV-1 Group M virus are arranged in decreasing prevalence in the table. Sequences containing a mixture at one of the major drug resistance positions were excluded from the table. For each pattern, drug susceptibility data on isolates matching the specific mutation pattern is shown on the left side. The analysis was originally published in [Rhee et al. 2004](#). The data was described in [Rhee et al. 2010](#).

The number of sequences with each pattern is indicated, as well as the median fold resistance, as assayed by the Virologic Pheno-Sense assay, to each of several individual members of the selected drug class. The databases used as reference is given below in figure – 2.2.



Figure - 2.2: Online databases as reference

2.1.4. What machine learning approach use to predict drug resistance:

Several antiretroviral drugs have been approved for use in HIV infected people. Despite efforts made by the scientific community, an effective drug that kills the virus has not been developed yet. A lot of computational algorithms have been used for finding mutations associated with drug resistance as well as for prediction of HIV resistance.

Antiretroviral drugs are a very effective therapy against HIV infection. However, the high mutation rate of HIV permits the emergence of variants that can be resistant to the drug treatment.

Predicting drug resistance to previously unobserved variants is therefore very important for an optimum medical treatment. In this paper, we propose the use of ANFIS designer (MATLAB tool) functions to predict drug resistance from virus sequence data.

These tool functions are very simple to implement and are able to take into account HIV data particularities, such as membership function, rules and to predict the different drug resistance scores of each mutation pattern, as it is known that not all positions contribute equally to the resistance.

To train a fuzzy system using neuro-adaptive methods, you must collect input/output training data using experiments or simulations of the system you want to model. In general, ANFIS training works well if the training data is fully representative of the features of the data that the trained FIS is intended to model.

To specify your training data, you can create an array in the MATLAB workspace. Each row contains a data point, with the final column containing the output value and the remaining columns containing input values. You can then pass this data to the training Data input argument of the ANFIS function or load it into the Neuro-Fuzzy Designer app.

Load the data from file. Each line of the file contains a data point with values separated by white space. The final value on each line is the output, and the remaining values are the inputs.

When using the ANFIS function, create or load the input data and pass it to the training Data input argument. When using Neuro-Fuzzy Designer, in the Load data section, select Training, and then to load data from a file, select file. To load data from the MATLAB workspace, select workspace.

**1. Mystery of HIV Drug Resistance: A Machine Learning Perspective
(Arumugam Mohanapriya, Nirmaladevi Ponnusamy, Sajitha Lulu
Sudhakaran, Vino Sundararajan, Pandjassaram Kanguene, 23
November 2019)**

Human immunodeficiency virus (HIV) is one of the fastest developing pathogens known. HIV/AIDS is an incurable disease which causes severe damage to the immune system. The recommended treatment for HIV/AIDS is a combination of three antiretroviral (ARV) drugs from two or more different drug groups and is known as highly active antiretroviral therapy (HAART). Drug resistance is a major impediment experienced by therapist in treating HIV infected patients. Theoretically, drug resistance can be predicted from the presence of specific mutations in the viral genome. With the current disease burden and lack of resources in developing countries, phenotypic tests are not viable. Developing a computational prediction of drug resistance phenotype will enable efficient and timely selection of the best treatment regimens. Nevertheless, the very large range of possible drug combinations and of viral mutational patterns leads to an extremely complex scenario, making prediction of in vivo treatment response extremely challenging. To deal with such complexity, machine learning methods are being increasingly explored. Clinical and technological developments has generated and stored large volumes of data in public databases which facilitates the use of machine learning methods for predicting drug resistance. Quite a lot of machine learning approaches such as neural networks, support vector machine, Bayesian networks, decision trees and linear regression have been proposed for the prediction of phenotypic drug resistance. Therefore, conducting resistance testing is certainly significant in order to administer appropriate antiviral drugs to HIV infected patients.

2. Data Mining: Practical Machine Learning Tools and Techniques (Ian Witten & Eibe Frank., January 2005)

Over time it has gained great importance and, as in other areas, there is an exponential growth of data that overload us and that we cannot control. As result, a gap between the production of data and our understandability of it emerges. To acquire knowledge from data the use of tools is required to allow the discovery of hidden information in databases being Knowledge Discovery in Databases (KDD) the provider of it consisting in the process of discovering new, valid, useful and perceptible patterns in the data. Consequently, Machine Learning (ML) is a type of approach to the discovery of knowledge in the data. Focus on the constuction of computational algorithms that can learn through data (from the past) to make predictions and when associated with DM tools can help to perceive complex phenomena as well as solve various problems from many areas. This step consists in transforming existing data and in dimensionality reduction. Witten, Frank, and Hall describe it as "engineering the input data into a form suitable for the learning scheme chosen and engineering the output to make it more effective". First, to increase the model's performance, the different scales of variables are scaled by applying the min-max method, and variables are rescaled to a range between 0 and 1.

3. Two-steps Learning of Fuzzy Cognitive Maps for Prediction and Knowledge Discovery on the HIV-1 Drug Resistance (Gonzalo Nápoles, Isel Grau, Rafael Bello & Ricardo Grau., February 2014)

The Human Immunodeficiency Virus (HIV) is a complex retrovirus that progressively deteriorates the immune system of infected patients, eventually causing death. Although antiviral drugs are not able to eradicate the HIV, they are designed to inhibit the function of three essential proteins in the virus replication process: protease, reverse transcriptase and integrase. Nevertheless, due to a high mutation rate, this virus is capable to develop resistance to existing drugs causing the treatment failure. Several machine

learning techniques have been proposed for predicting HIV drugs resistance, but most of them are difficult to interpret. Actually, in last years the protein modelling of this virus has become, from diverse points of view, an open problem for researchers. In this paper we propose a novel Fuzzy Cognitive Maps (FCM) based model for analysing the behaviour of the HIV-1 protease protein. With this goal in mind, a two-steps learning algorithm using Swarm Intelligence is also introduced for optimizing the modelling parameters. The first step is capable to estimate, through a continuous search method, the biological causality among amino acids describing the sequence. Resulting adjusted maps are combined into a single one through an aggregation procedure for obtaining an initial prototype map. The second step optimize the prototype by finding those amino acids directly associated with resistance using a discrete meta-heuristic. At the end, a fully optimized prototype map is obtained allowing predicting HIV-1 drug resistance and also discovering relevant knowledge on causal in sequences directly associated with resistance, for seven well-known protease inhibitors.

4. Computational Drug Design: A Guide for Computational and Medicinal Chemists (David Young., February 2009)

Helps you choose the right computational tools and techniques to meet your drug design goals Computational Drug Design covers all of the major computational drug design techniques in use today, focusing on the process that pharmaceutical chemists employ to design a new drug molecule. The discussions of which computational tools to use and when and how to use them are all based on typical pharmaceutical industry drug design processes. Following an introduction, the book is divided into three parts: Part One, The Drug Design Process, sets forth a variety of design processes suitable for a number of different drug development scenarios and drug targets. The author demonstrates how computational techniques are typically used during the design process, helping readers choose the best computational tools to meet their goals. Part Two, Computational Tools and Techniques, offers a series of chapters, each one dedicated to a single computational technique. Readers discover the strengths and weaknesses of each technique.

Moreover, the book tabulates comparative accuracy studies, giving readers an unbiased comparison of all the available techniques. Part Three, Related Topics, addresses new, emerging, and complementary technologies, including bioinformatics, simulations at the cellular and organ level, synthesis route prediction, proteomics, and prodrug approaches. The book's accompanying CD-ROM, a special feature, offers graphics of the molecular structures and dynamic reactions discussed in the book as well as demos from computational drug design software companies. Computational Drug Design is ideal for both students and professionals in drug design, helping them choose and take full advantage of the best computational tools available.

5. Predicting human immunodeficiency virus inhibitors using multi-dimensional Bayesian network classifiers (Hanen Borchani, Concha Bielza, Carlos Toro, & Pedro Larranaga., January 2013)

Our aim is to use multi-dimensional Bayesian network classifiers in order to predict the human immunodeficiency virus type 1 (HIV-1) reverse transcriptase and protease inhibitors given an input set of respective resistance mutations that an HIV patient carries. Materials and methods: Multi-dimensional Bayesian network classifiers (MBCs) are probabilistic graphical models especially designed to solve multi-dimensional classification problems, where each input instance in the data set has to be assigned simultaneously to multiple output class variables that are not necessarily binary. In this paper, we introduce a new method, named MB-MBC, for learning MBCs from data by determining the Markov blanket around each class variable using the HITON algorithm. Our method is applied to both reverse transcriptase and protease data sets obtained from the Stanford HIV-1 database. Regarding the prediction of antiretroviral combination therapies, the experimental study shows promising results in terms of classification accuracy compared with state-of-the-art MBC learning algorithms. For reverse transcriptase inhibitors, we get 71% and 11% in mean and global accuracy, respectively; while for protease inhibitors, we get more than 84% and 31% in mean and global accuracy, respectively. In addition, the analysis

of MBC graphical structures lets us gain insight into both known and novel interactions between reverse transcriptase and protease inhibitors and their respective resistance mutations. MB-MBC algorithm is a valuable tool to analyse the HIV-1 reverse transcriptase and protease inhibitors prediction problem and to discover interactions within and between these two classes of inhibitors.

6. A Rough Set-Based Model of HIV-1 Reverse Transcriptase Resistome (Marcin Kierczak, Krzysztof Ginalski, Michał Dramiński, & Jacek Koronacki., October 2009)

Reverse transcriptase (RT) is a viral enzyme crucial for HIV-1 replication. Currently, 12 drugs are targeted against the RT. The low fidelity of the RT-mediated transcription leads to the quick accumulation of drug-resistance mutations. The sequence resistance relationship remains only partially understood. Using publicly available data collected from over 15 years of HIV proteome research, we have created a general and predictive rule-based model of HIV-1 resistance to eight RT inhibitors. Our rough set-based model considers changes in the physicochemical properties of a mutated sequence as compared to the wild-type strain. Thanks to the application of the Monte Carlo feature selection method, the model takes into account only the properties that significantly contribute to the resistance phenomenon. The obtained results show that drug-resistance is determined in more complex way than believed. We confirmed the importance of many resistance-associated sites, found some sites to be less relevant than formerly postulated and—more importantly—identified several previously neglected sites as potentially relevant. By mapping some of the newly discovered sites on the 3D structure of the RT, we were able to suggest possible molecular-mechanisms of drug-resistance. Importantly, our model has the ability to generalize predictions to the previously unseen cases. The study is an example of how computational biology methods can increase our understanding of the HIV-1 resistome.

CHAPTER-3

AIM AND OBJECTIVES

The primary aim for the study is to determine and predict the resistance in drugs towards HIV-AIDS.

Listed are the prominent objectives of the study:

- To study all the HIV-AIDS drug resistance database online.
- To study HIV-AIDS drug resistance dataset – international as well as national.
- To create a worksheet based on the HIV drug resistance database.
- To work-on predicting drug resistance score using machine learning technique.
- To work-on ANFIS algorithm to simulate the given dataset.
- To analyze the statistics of HIV drug resistance dataset.
- To co-construct the rule editor and interpret the data accordingly in rule viewer.
- To interpret the three-dimensional graphical representation in surface viewer.

CHAPTER-4

METHODOLOGY

4.1. HIV DRUG RESISTANCE DATABASES USED

4.1.1. STANFORD HIV DRUG RESISTANCE DATABASE (HIVDB):

It is an essential resource for public health officials monitoring ADR and TDR, for scientists developing new ARV drugs, and for HIV care providers managing patients with HIVDR. Researchers studying HIVDR also use HIVDB to place their findings in the context of previous studies and to perform meta-analyses that require data from large numbers of studies, making it possible to discover new knowledge that cannot be discovered from a single study. Figure – 4.1 shows the source of the dataset.

4.1.2. HIV DRUG RESISTANCE DATABASE

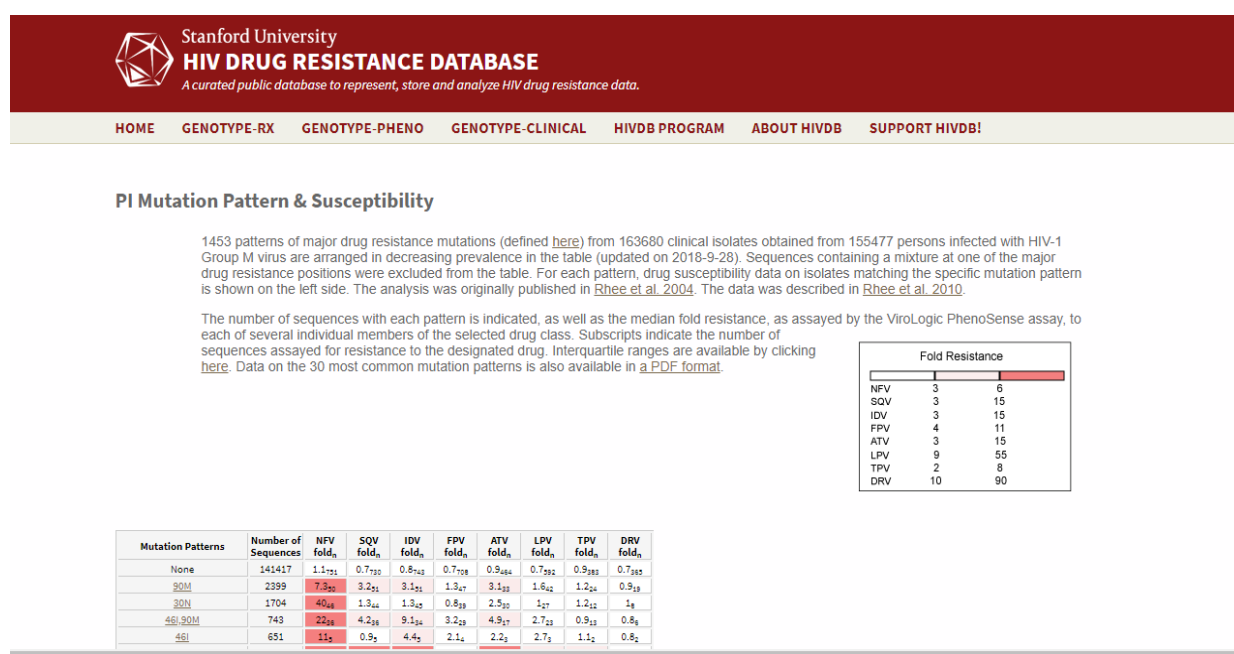


Figure – 4.1: Source of the dataset

This collect, annotate, and analyse genotypic drug resistance data from previously untreated (ARV-naïve) and previously treated (ARV-experienced) persons and systematically aggregate data from many published studies to analyse temporal global trends in the drug-resistance mutations responsible for ADR and TDR, thus creating a resource for clinical and molecular epidemiologists. Our work is made

challenging by the extensive genetic sequence heterogeneity of HIV, including the many different HIV subtypes and recombinant forms. Nonetheless, by collecting well-characterized HIV sequences from more than 150,000 individuals, we have been able to distinguish naturally occurring HIV mutations from those that are selected by ARV therapy and contribute to HIVDR.

4.1.3. WHO HIV Drug Resistance database:

This collaborate with researchers from around the world to track global trends in ADR and TDR and to characterize the genetic mechanisms of HIVDR to the ARVs for which more data are needed including those belonging to the nucleoside RT inhibitor (NRTI), non-nucleoside RT inhibitor (NNRTI), protease inhibitor (PI), and integrate strand transfer inhibitor (INSTI) class of drugs.

4.2. MACHINE LEARNING APPROACH:

With the rise in big data, machine learning has become a key technique for solving problems.

Machine learning is a data analytics technique that teaches computers to do what comes naturally to humans and animals: learn from experience. Machine learning algorithms use computational methods to “learn” information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases.

4.2.1. Fuzzy Logic Toolbox:

Fuzzy Logic Toolbox software provides a command-line function (anfis) and an interactive app (Neuro-Fuzzy Designer) for training an adaptive neuro-fuzzy inference system (ANFIS).

The architecture and learning procedure underlying ANFIS (adaptive-network-based fuzzy inference system) is presented, which is a fuzzy inference system

implemented in the framework of adaptive networks. By using a hybrid learning procedure, the proposed ANFIS can construct an input-output mapping based on both human knowledge (in the form of fuzzy if-then rules) and stipulated input-output data pairs.

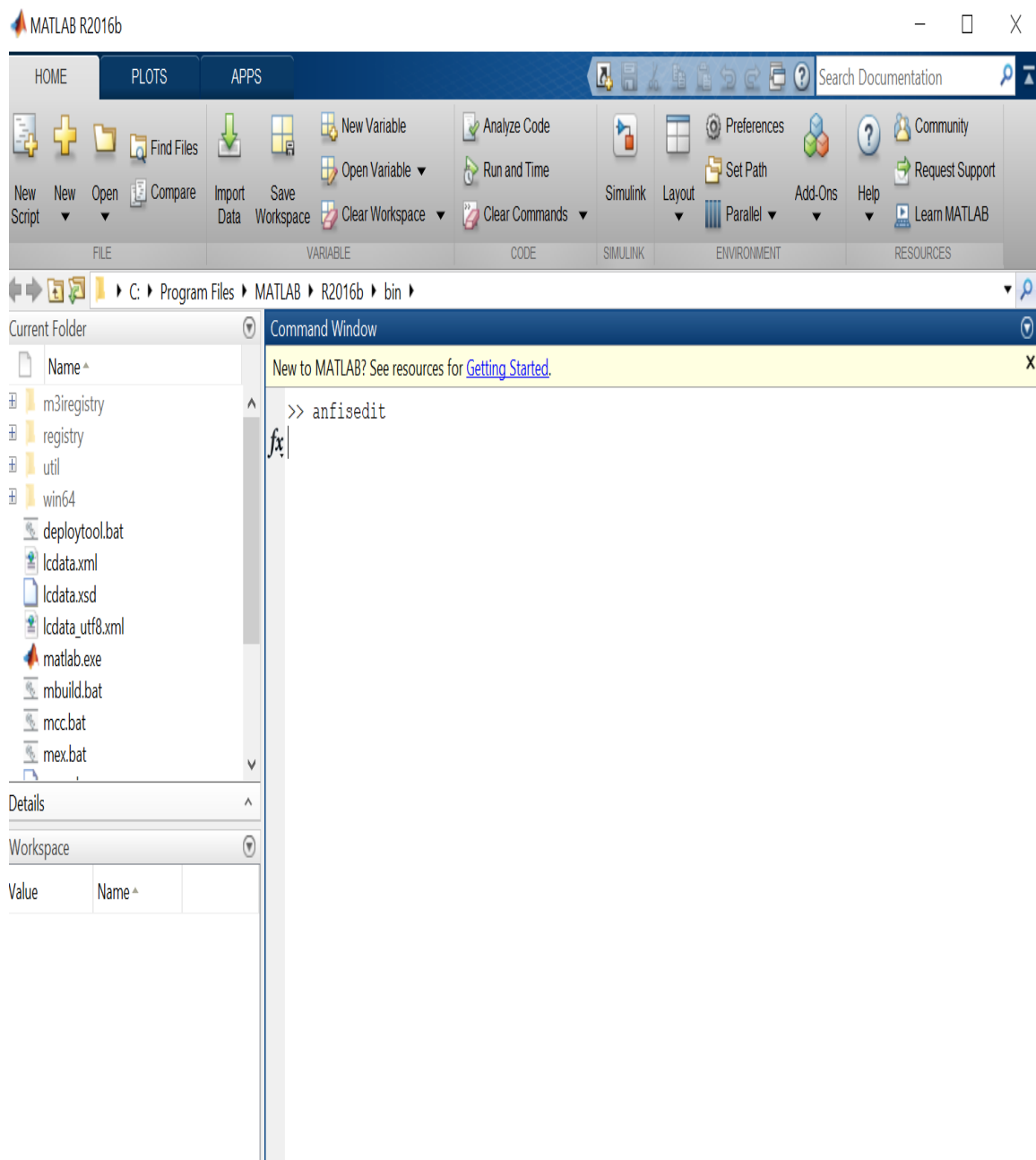


Figure – 4.2: MATLAB Command Window

The toolbox make the model complex system behaviours using simple logic rules, and then implement these rules in a ANFIS System interface.

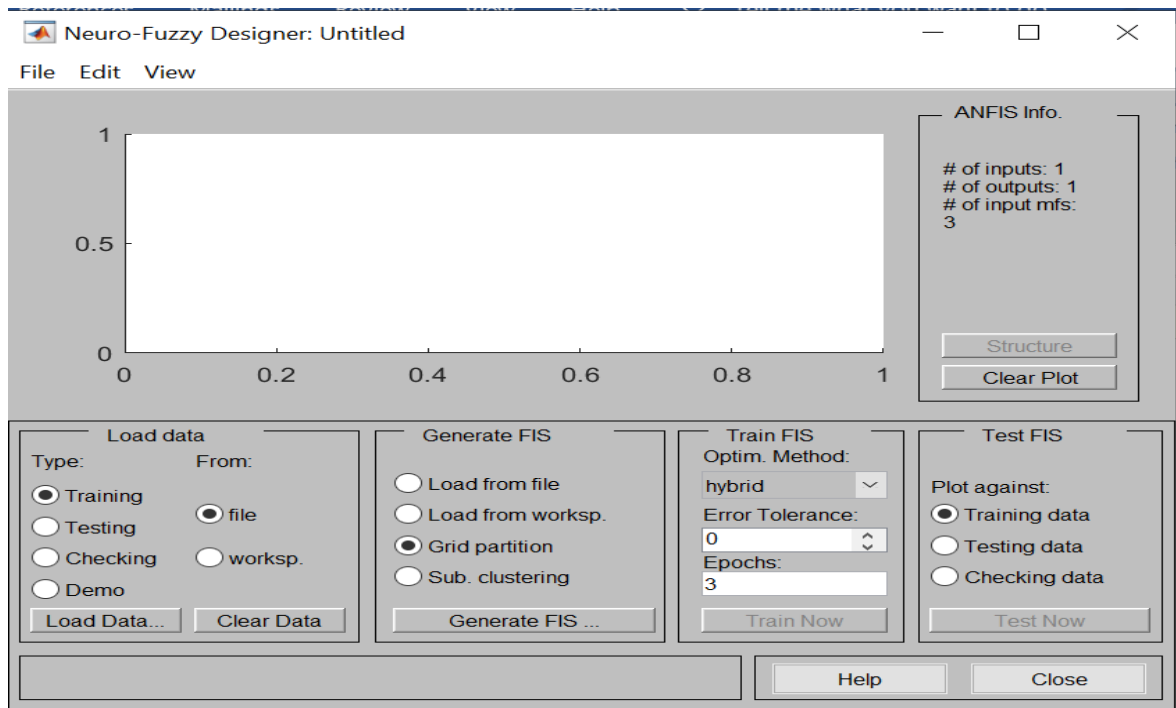


Figure – 4.3: ‘anfisedit’ command opens ANFIS Designer window.

4.2.1.1. TRAINING DATA:

To train a fuzzy system using neuro-adaptive methods, need to collect input/output training data using experiments or simulations of the system we want to model. In general, ANFIS training works well if the training data is fully representative of the features of the data that the trained FIS is intended to model.

- Load the data from a file. Each line of the file contains a data point with values separated by white space. The final value on each line is the output, and the remaining values are the inputs.

When using the `anfis` function, create or load the input data and pass it to the training Data input argument. When using Neuro-Fuzzy Designer, in the Load data section, select Training, and then load data from a file.

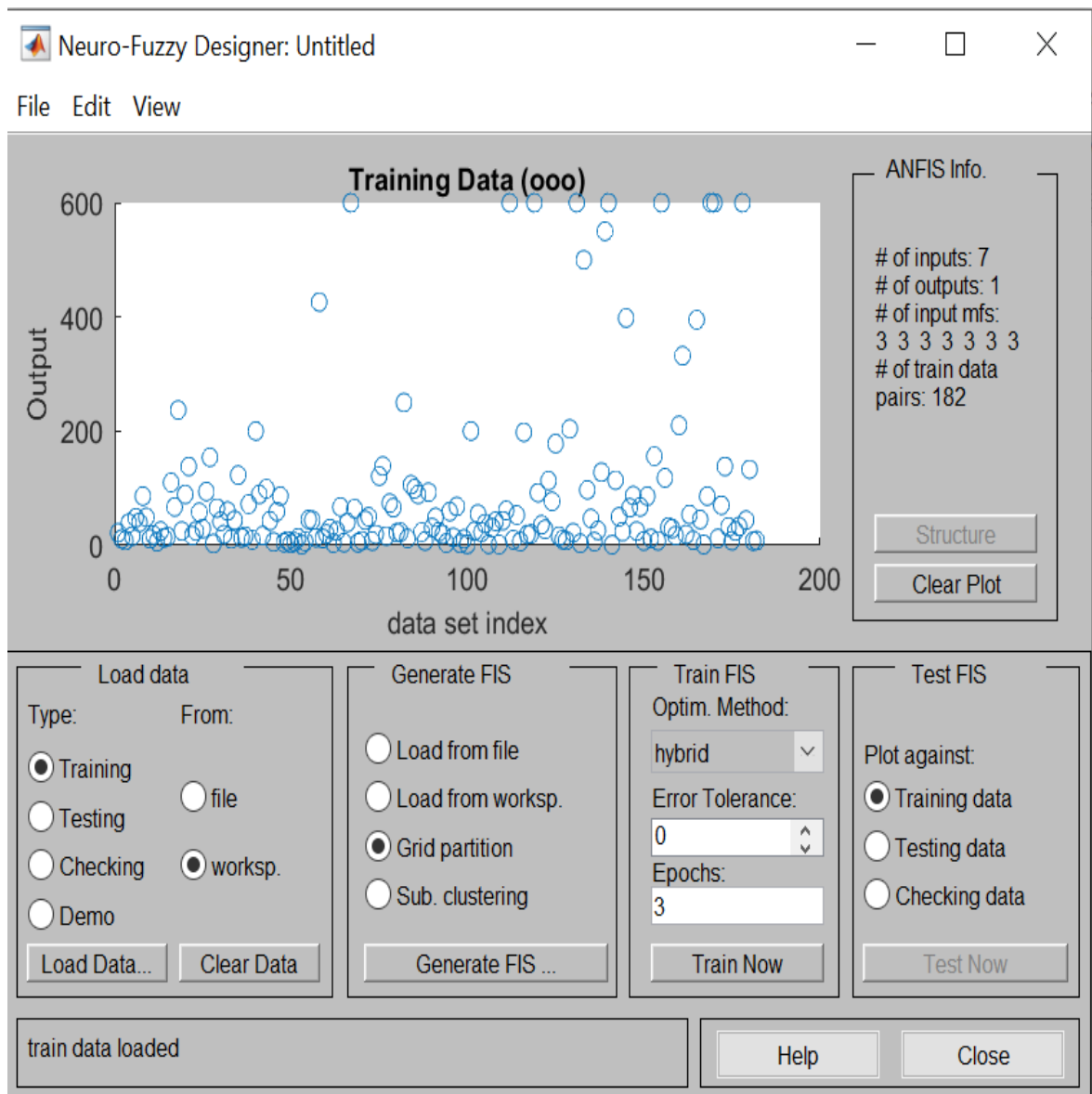


Figure – 4.5: Data loaded in Neuro-Fuzzy Designer

4.2.1.2. GENERATE OR LOAD FIS STRUCTURE:

Before start the FIS training, you must specify an initial FIS model structure. To specify the model structure, perform one of the following tasks:

- Load a previously saved FIS object from a file or the MATLAB workspace.
- Generate the initial FIS model using grid partitioning.

Click Generate FIS. In the Add Membership Functions dialog box:

- In the Input section, in Number of MFs, specify the number of input membership functions. For this example, use 3 3 3 3 3 membership functions for all input variables.
- In MF Type, select gbellmf as the input membership function type.

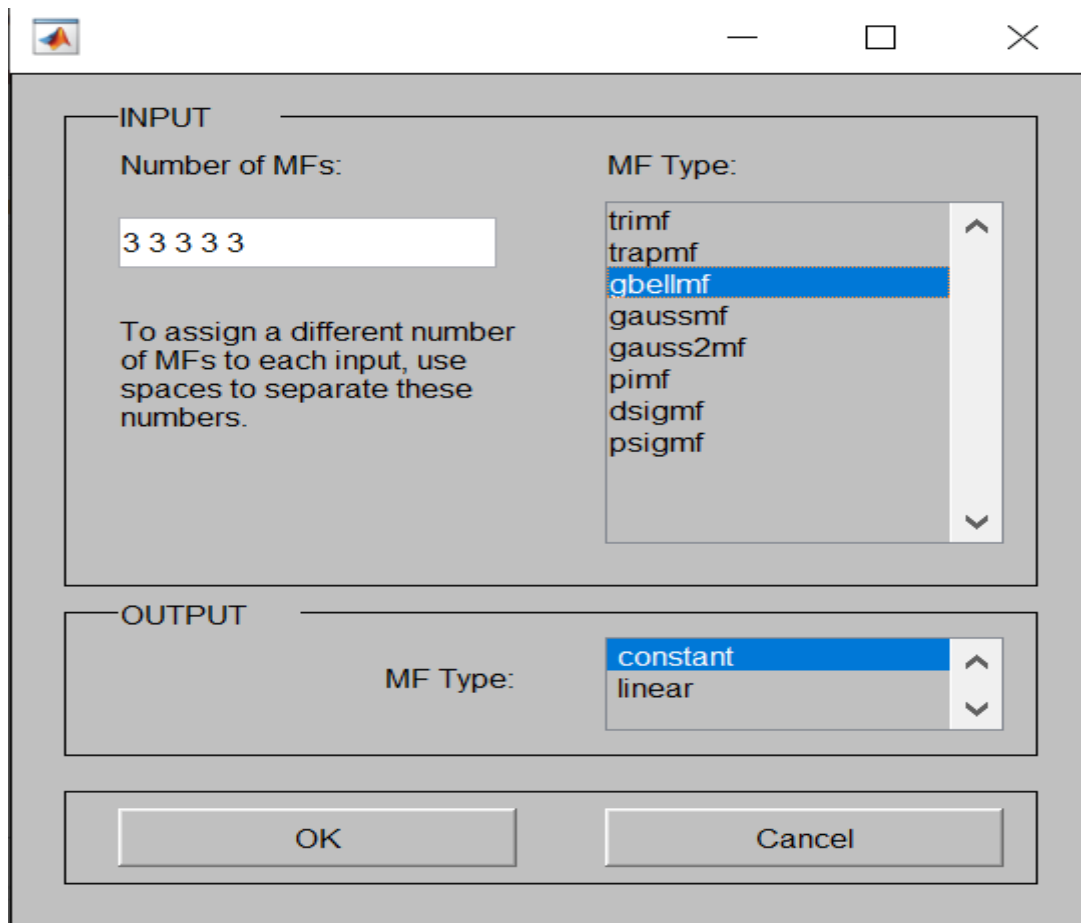


Figure – 4.6: Loading FIS Structure

4.2.1.4. VIEW FIS STRUCTURE:

After you load or generate the FIS, you can view the model structure. To do so, in Neuro-Fuzzy Designer, click structure.

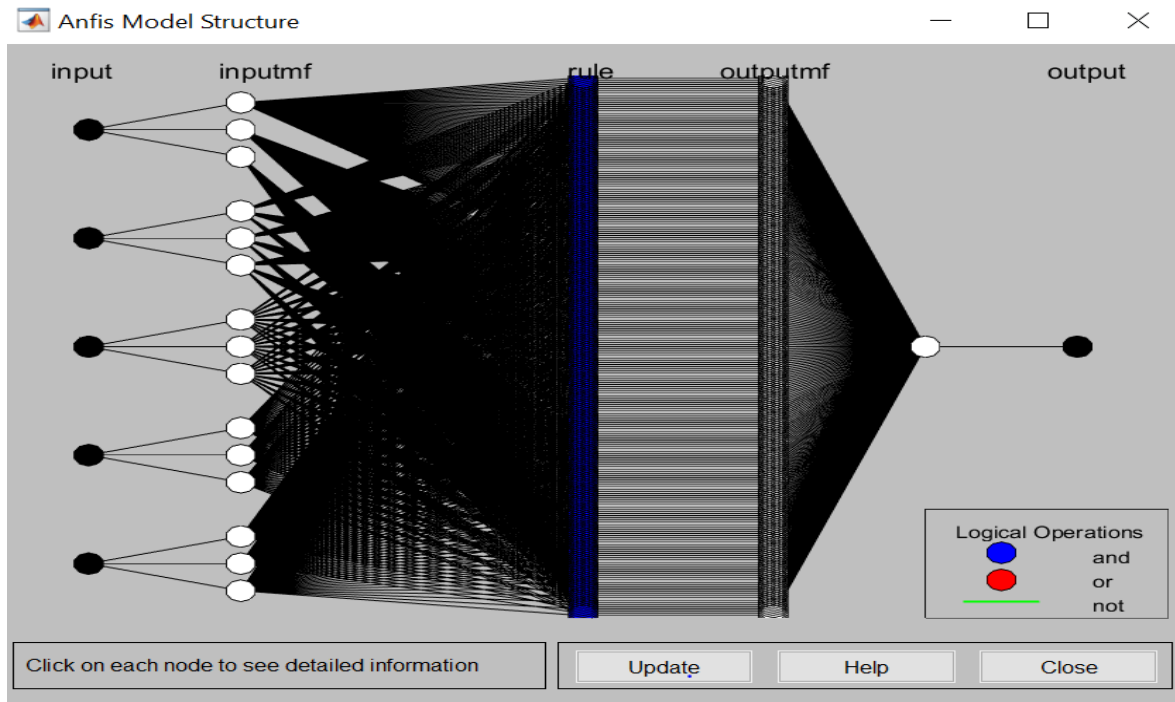


Figure – 4.7: View of model structure

The branches in this graph are color coded. Color coding of branches characterize the rules and indicate whether or not AND, NOT, or OR are used in the rules. The input is represented by the left-most node and the output by the right-most node. The node represents a normalization factor for the rules.

4.2.1.5. TRAIN FIS:

After loading the training data and generating the initial FIS structure, you can train your FIS. To do so, in Neuro-Fuzzy Designer, in the Train FIS section, specify the following parameters.

- **Optim.Method** — Optimization method. For this example, select the hybrid method, which uses a combination of backpropagation and least-squares regression to tune the FIS parameters.
- **Epochs** — Number of training epochs. For this example, specify 40 epochs.
- **Error Tolerance** — Error tolerance stopping condition. For this example, specify a value of 0, which indicates that the training will stop when the number of training epochs is reached.

To train the FIS, click Train Now.

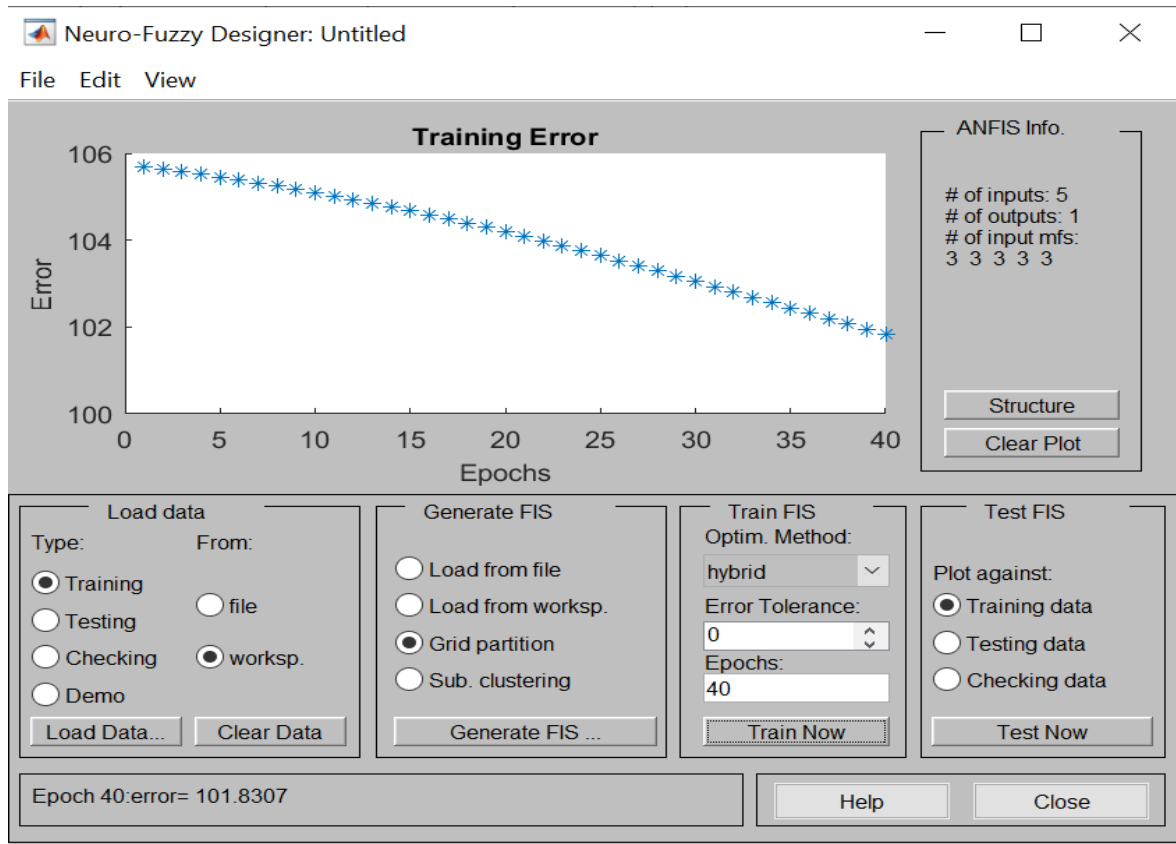


Figure – 4.8: Trained FIS

The app trains the FIS and plots the training error (as stars) and checking error for each training epoch. The checking error decreases up to a certain point in the training, and then it increases. This increase occurs at the point where the training starts overfitting the training data. The app selects the FIS associated with this overfitting point as the trained ANFIS model.

4.2.1.6. SURFACE VIEWER:

Surface Viewer use to view the output surface for your fuzzy system. To view the output surface, you must specify the input and output variables of your FIS, their corresponding membership functions, and the fuzzy rules for your system.

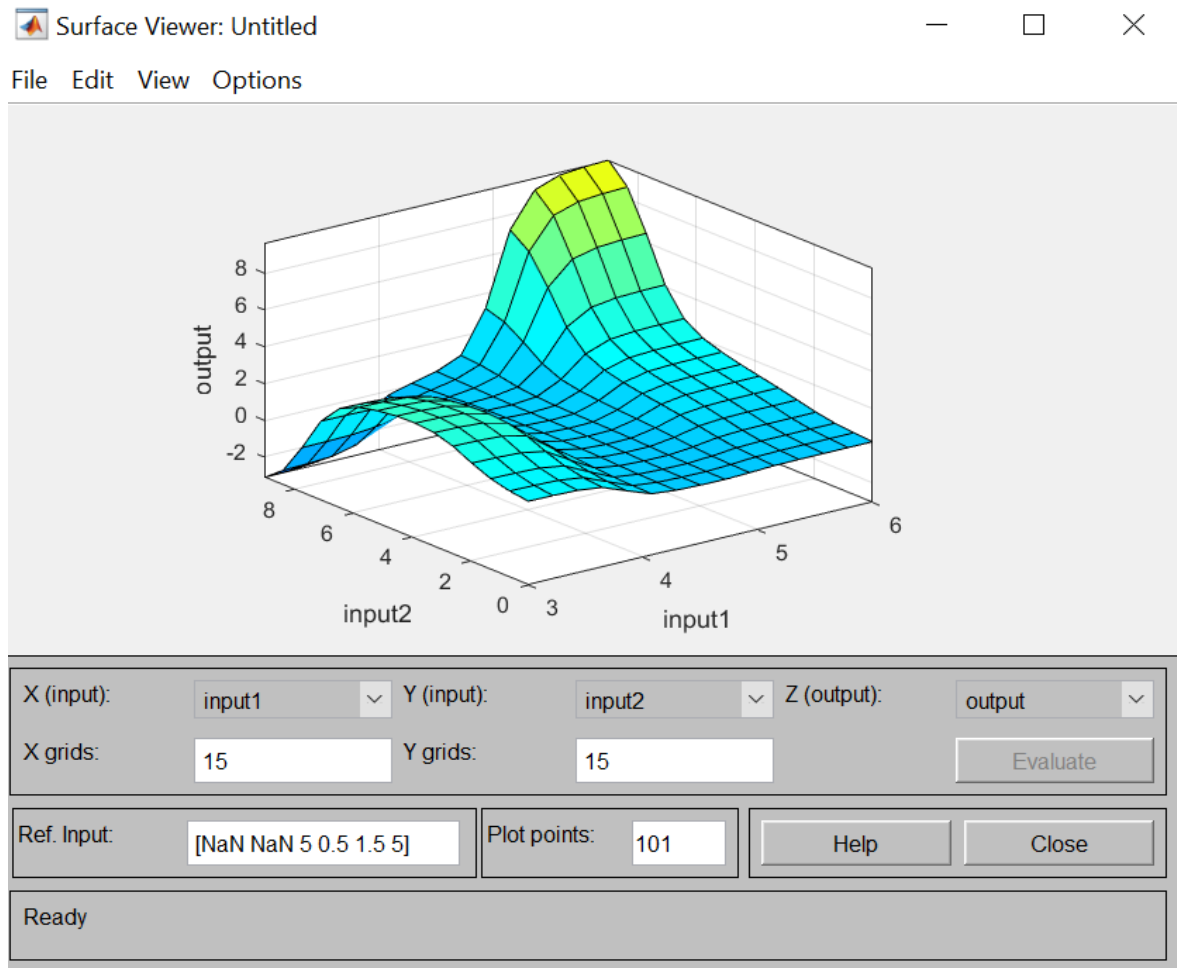


Figure – 4.9: Surface viewer of NFV

4.2.1.7. RULE VIEWER:

Rule Viewer use to view the inference process for your fuzzy system. You can adjust the input values and view the corresponding output of each fuzzy rule, the aggregated output fuzzy set, and the defuzzified output value. To view the inference process, you must specify the input and output variables of your FIS, their corresponding membership functions, and the fuzzy rules for your system.

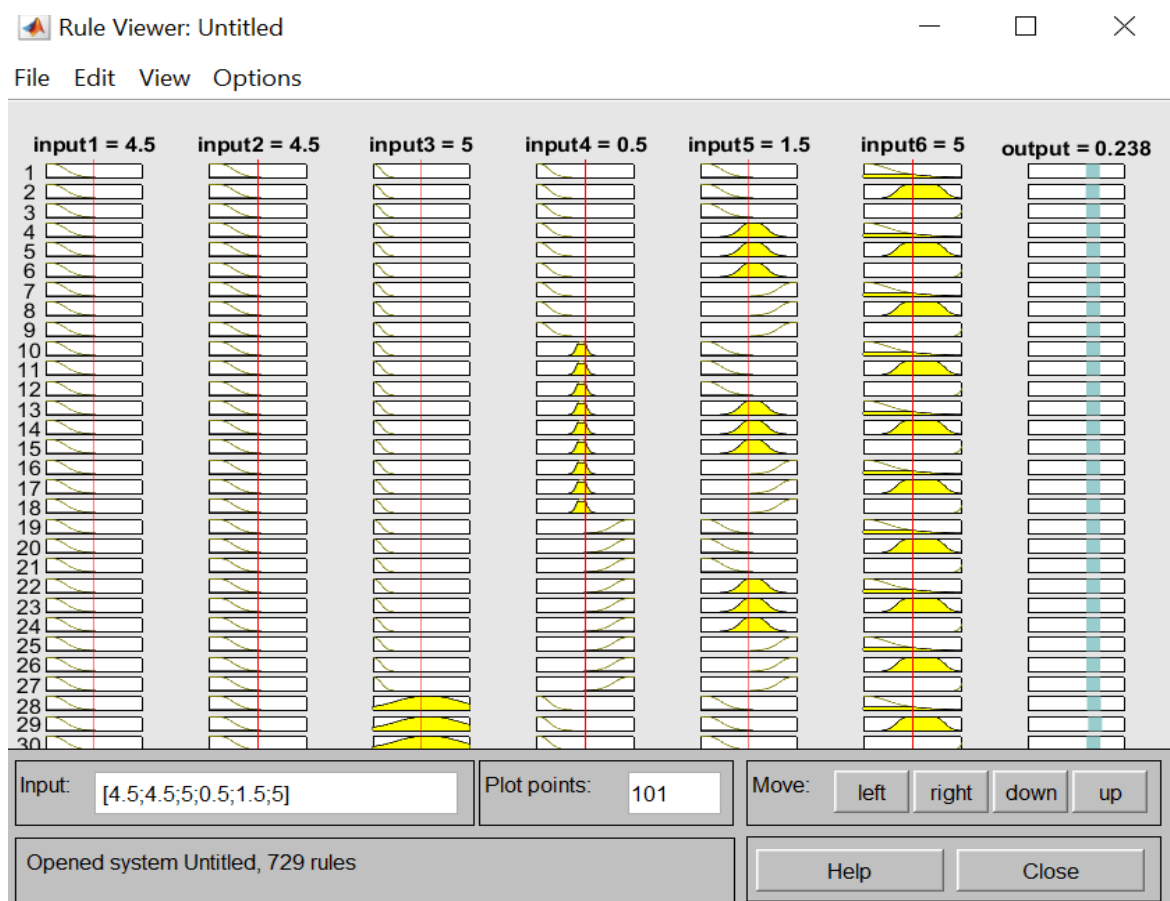


Figure – 4.10: Rule viewer represents the drug resistance scores

CHAPTER-5

RESULT AND DISCUSSION

5.1 EXCEL SHEET OF PROTEASE TABLE:

From the HIV drug resistance database we have made excel sheet for protease table.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	MUTATION PATTERNS	30	32	46	47	48	50	54	76	82	84	88	90				NFV	SQV	IDV	FPV	ATV	LPV	TPV	DRV
2	90M	0	0	0	0	0	0	0	0	0	0	0	1		M-1		7.3	3.2	3.1	1.3	3.1	1.6	1.2	0.9
3	30N	2	0	0	0	0	0	0	0	0	0	0	0		N-2		40	1.3	1.3	0.8	2.5	1	1.2	1
4	46I,90M	0	0	3	0	0	0	0	0	0	0	0	1		I-3		22	4.2	9.1	3.2	4.9	2.7	0.9	0.8
5	46I	0	0	3	0	0	0	0	0	0	0	0	0		A-4		11	0.9	4.4	2.1	2.2	2.7	1.1	0.8
6	54V,82A,90M	0	0	0	0	0	0	5	0	4	0	0	1		V-5		57	27	24	3.6	36	39	2.4	1.6
7	54V,82A	0	0	0	0	0	0	5	0	4	0	0	0		L-6		9.4	1.4	8	1.5	2.3	11	1.6	0.8
8	46L	0	0	6	0	0	0	0	0	0	0	0	0		S-7		8.2	1.5	2.6	1.1	2.4	1.6		
9	46L,54V,82A	0	0	6	0	0	0	5	0	4	0	0	0		F-8		39	6.2	19	4.6	31	37	1.1	1.7
10	46I,54V,82A	0	0	3	0	0	0	5	0	4	0	0	0		T-9		16	1.6	15	2.5	7.5	27	1.6	0.8
11	46I,84V,90M	0	0	3	0	0	0	0	0	0	5	0	1		C-10		48	43	28	7.7	22	17	2	4
12	46L,54V,82A,90M	0	0	6	0	0	0	5	0	4	0	0	1				40	11	22	8.2	17	50	4.4	4.1
13	84V,90M	0	0	0	0	0	0	0	0	0	5	0	1				19	25	11	4.5	15	6.4	2	2.1
14	46I,54V,84V,90M	0	0	3	0	0	0	5	0	0	5	0	1				86	161	53	20	171	96	4.7	11
15	46I,54V,82A,90M	0	0	3	0	0	0	5	0	4	0	0	1				49	22	29	6.4	24	58	2.6	1.8
16	82A	0	0	0	0	0	0	0	0	4	0	0	0				2.1	0.8	2.1	1	1.9	2.2	0.4	
17	88S	0	0	0	0	0	0	0	0	0	0	7	0				7.7	1.2	2.5	0.1	7.2	1.4	0.6	0.3
18	46I,54V,76V,82A	0	0	3	0	0	0	5	5	4	0	0	0				11	1.6	37	6.5	1.4	82	0.5	
19	54V,84V,90M	0	0	0	0	0	0	5	0	0	5	0	1				144	773	50	13	155	43	17	132
20	54V,90M	0	0	0	0	0	0	5	0	0	0	0	1				34	60	5	1.3	72	28	23	2.1
21	30N,90M	2	0	0	0	0	0	0	0	0	0	0	1				74	6	2.8	1.3	4.8	1.8	0.5	0.4
22	50L	0	0	0	0	0	6	0	0	0	0	0	0				0.3	0.3	0.3	0.3	6	0.3		
23	84V	0	0	0	0	0	0	0	0	0	5	0	0				2.5	2.8	1.9	3.9	2.3	2.6	3.2	2.6
24	46I,88S	0	0	3	0	0	0	0	0	0	0	7	0				18	1.3	5.8	0.2	6.1	0.8		
25	46L,82A	0	0	6	0	0	0	0	0	4	0	0	0				6.1	1.3	7.4	5.8	2.5	7		
26	54V	0	0	0	0	0	0	5	0	0	0	0	0				18	8.1	6.2	3	8.6	5.4	2.7	0.6
27	30N,46I	2	0	3	0	0	0	0	0	0	0	0	0				25	0.65	1.25	0.85	1.53	0.84	0.42	0.51
28	82A,90M	0	0	0	0	0	0	0	0	4	0	0	1				6.1	1.6	7.1	1.6	5.5	2.7		
29	46I,84V	0	0	3	0	0	0	0	0	0	5	0	0				13	11	5.6	6.7	14	14	11	
30	46L,90M	0	0	6	0	0	0	0	0	0	0	0	1				15	3.4	7.5	1.7	5.4	2.8	0.6	0.7
31	48V,54V,82A	0	0	0	0	5	0	5	0	4	0	0	0				29	152	23	2.5	27	33	1	0.8
32	46I,54V,82F,90M	0	0	3	0	0	0	5	0	8	0	0	1				110	6.5	92	25	67	87	4.3	11
33	54L,84V,90M	0	0	0	0	0	0	6	0	0	5	0	1				50	197	20	30	32	16	1.5	19
34	46I,82A	0	0	3	0	0	0	0	0	4	0	0	0				67	22	41	7.8	43	45		
35	46I,47V,54V,84V,90M	0	0	3	5	0	0	5	0	0	5	0	1				237		148	54	184	170	8.7	13
36	32I,46I,82A,90M	0	3	3	0	0	0	0	0	4	0	0	1				25	4.5	21	12	20	15	1.5	46
37	46L,54V,84V,90M	0	0	6	0	0	0	5	0	0	5	0	1				89	257	69	32	57	103	5.5	7.5
38	32I,46I,47V,54M,82A,90M	0	3	3	5	0	0	1	0	4	0	0	1				138	27	144	300	71	350	9.8	346
39	46L,54V,82A,84V	0	0	6	0	0	0	5	0	4	5	0	0				17	79	12	13	16	35	12	2.1

Figure - 5.1: Excel sheet of protease table with 8 drugs.

Then we filtered the each and every column. Filter is nothing but a removing blanks and zero from the column. Here we select the 46 mutation because it has more combination and less blanks and zero compare to other mutation value. Then we compare all the eight drugs with 46 mutation and make a graph based on this.

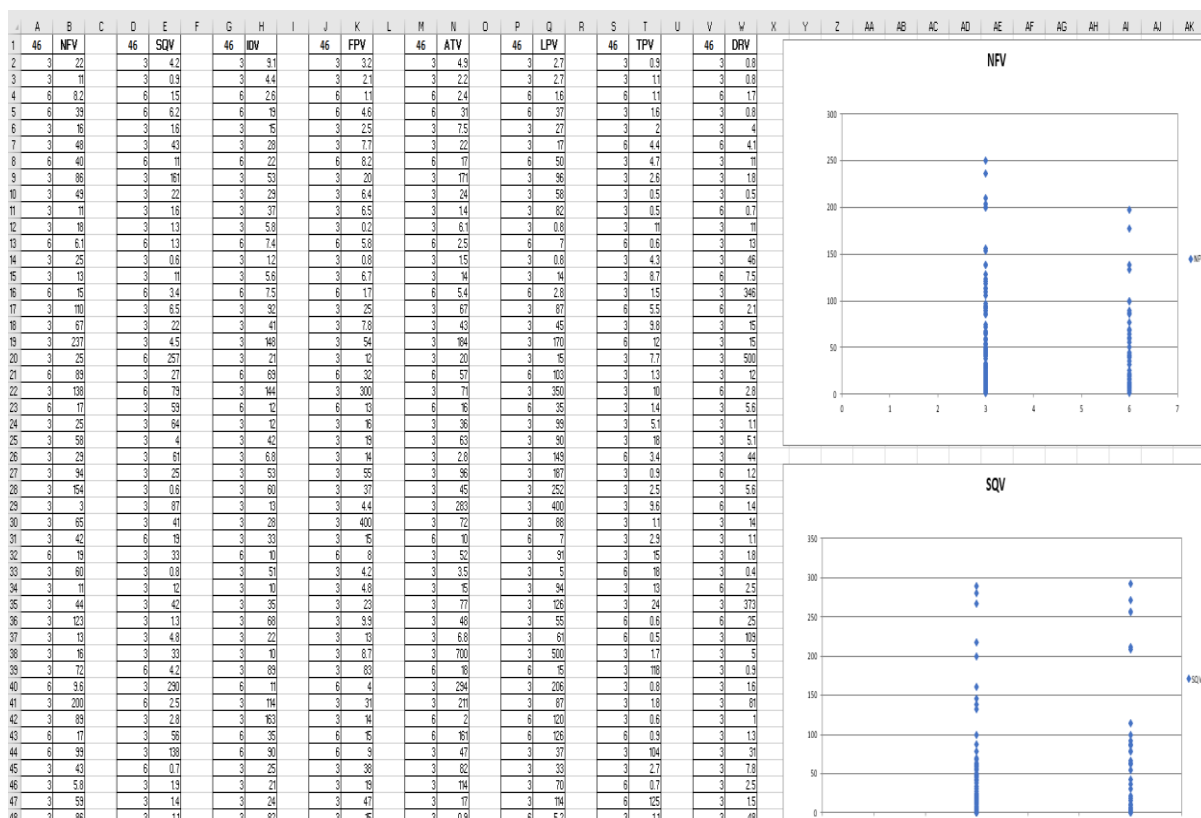


Figure - 5.2: Excel sheet with 46 mutation and 8 drugs.

After that we made the final excel sheet based on the grouping of amino acids to perform the machine learning analysis.

	A	B	C	D	E	F	G	H	I
1	3	0	0	1	0	0	22		
2	3	0	0	0	0	0	11		
3	6	0	0	0	0	0	8.2		
4	6	5	0	0	0	4	39		
5	3	5	0	0	0	4	16		
6	3	0	5	1	0	0	48		
7	6	5	0	1	0	4	40		
8	3	5	5	1	0	0	86		
9	3	5	0	1	0	4	49		
10	3	5	0	0	0	4	11		
11	3	0	0	0	0	0	18		
12	6	0	0	0	0	4	6.1		
13	3	0	0	0	2	0	25		
14	3	0	5	0	0	0	13		
15	6	0	0	1	0	0	15		
16	3	5	0	1	0	8	110		
17	3	0	0	0	0	4	67		
18	3	5	5	1	0	0	237		
19	3	0	0	1	0	4	25		
20	6	5	5	1	0	0	89		
21	3	1	0	1	0	4	138		
22	6	5	5	0	0	4	17		
23	3	5	5	0	0	4	25		
24	3	5	5	0	0	0	58		
25	3	5	0	1	0	4	29		
26	3	5	5	1	0	4	94		
27	3	0	0	1	0	4	154		
28	3	0	0	0	0	0	3		
29	3	6	5	1	0	0	65		

Figure - 5.3: Final excel sheet for analysis.

5.1.1 RULE VIEWER OF STANFORD HIV DRUG RESISTANCE DATASET:

As datasheet of 1453 patterns of major drug resistance mutations with 287 rules is created in rule editor. So, rule viewer interpreted the median fold drug resistance score for mutation positions as well as drugs.

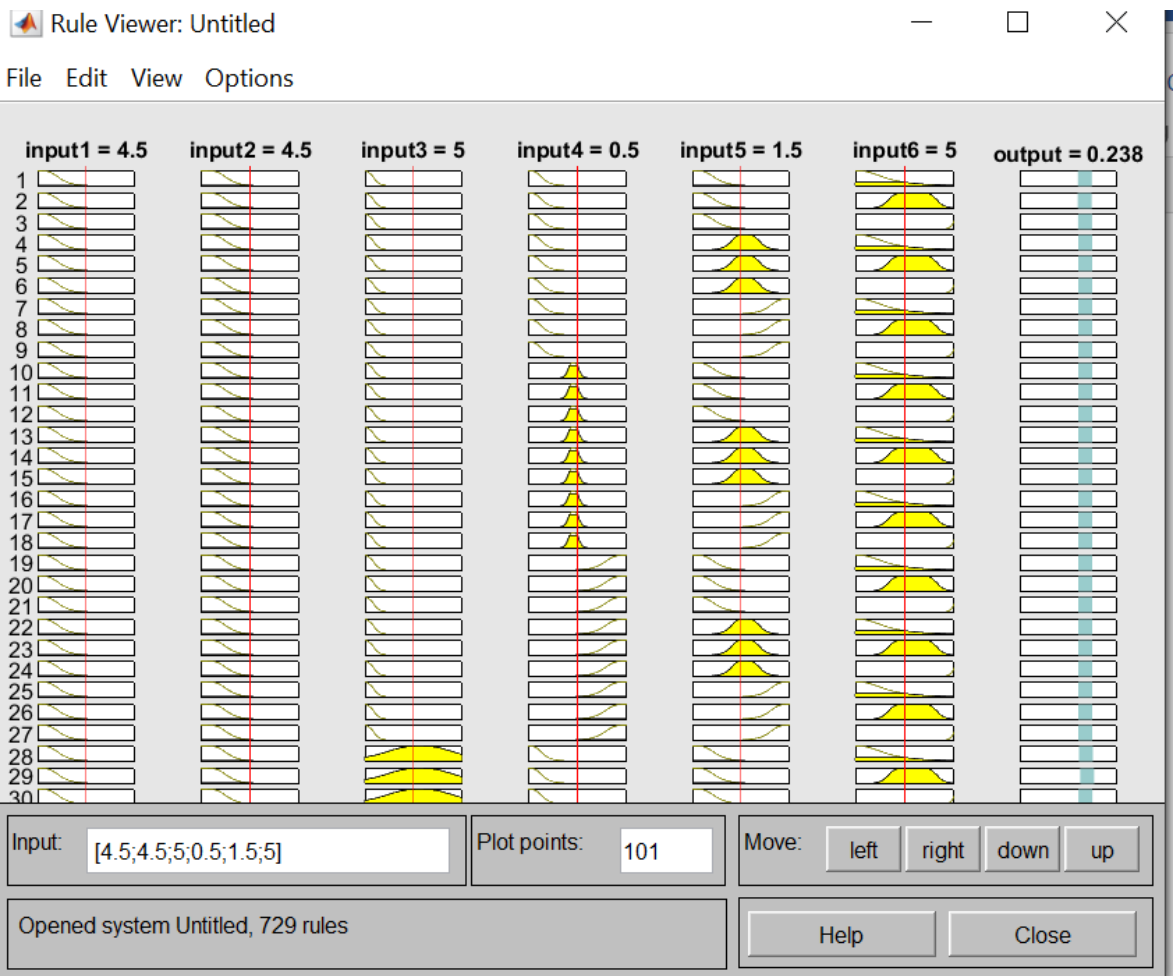


Figure - 5.4: Rule viewer of different dataset with NVF

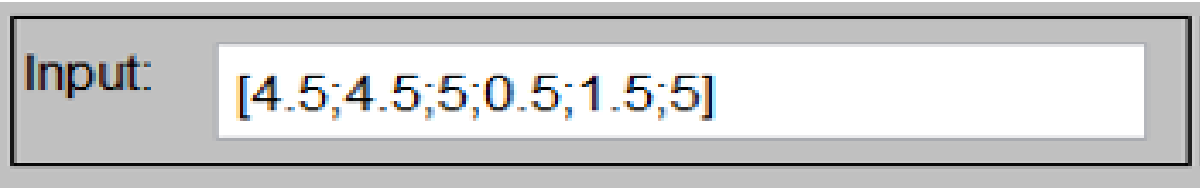


Figure – 5.5: Input datas

In figure - 5.4 rule viewer represented the median fold drug resistance score and in figure - 5.5 where input values are given for mutation position [54 = 4.5; 84 = 4.5; 90 = 5; 30 = 0.5; 82 = 1.5; 48 = 5]. Plot points of input values is in figure – 5.6.

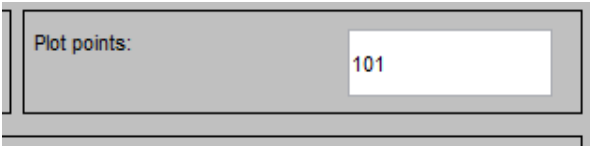


Figure – 5.6: Plot points

We input one set of values in figure 5.5. But what if we change input values or replace with another values, so in figure 5.7 input values are replaced by another values and drug’s median fold resistance scores are interpreted accordingly.

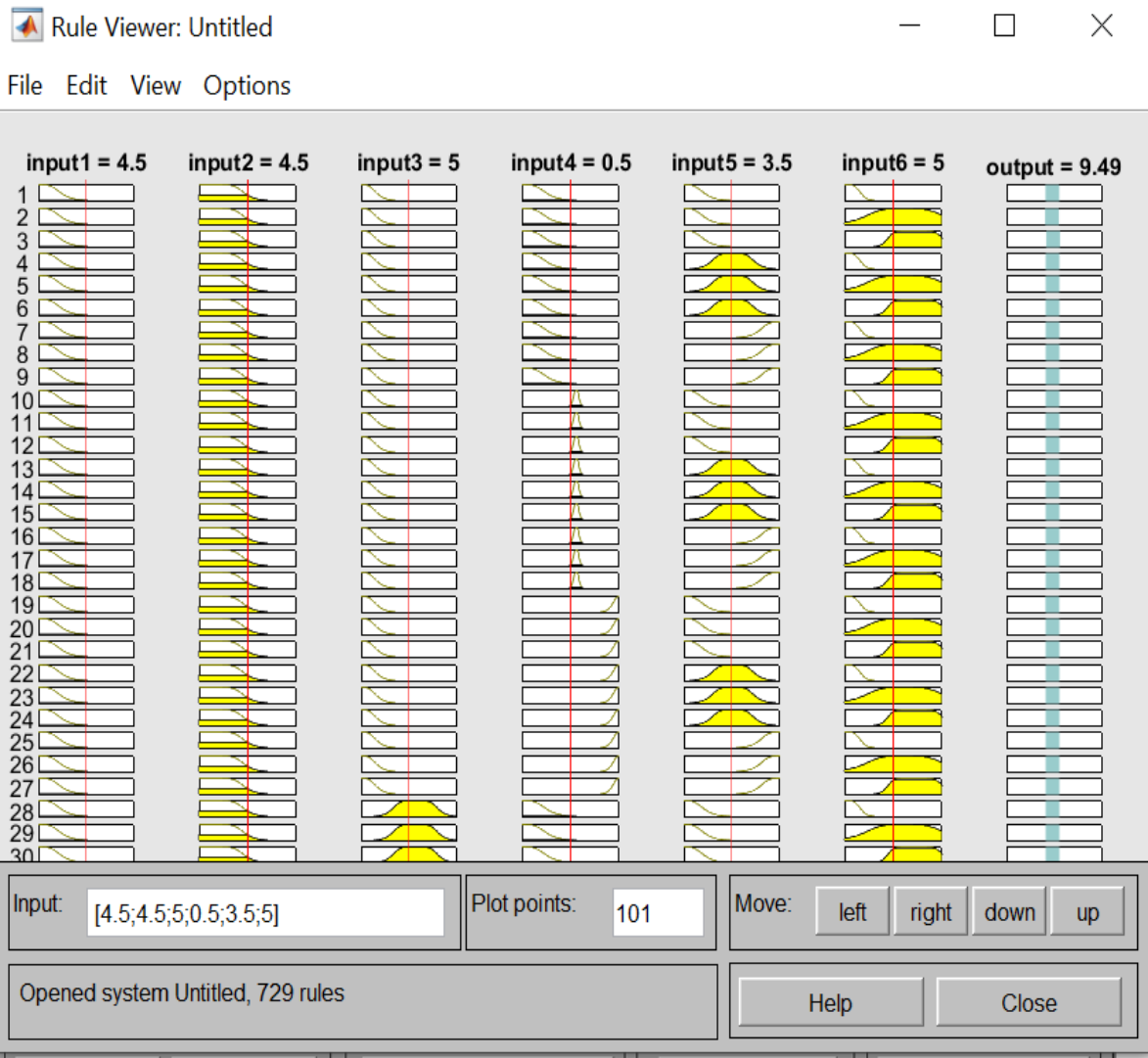


Figure - 5.7: Rule viewer of different dataset with SQV

Input: [4.5;4.5;5;0.5;3.5;5]

Figure – 5.8: Different input values

In figure - 5.7 rule viewer represented the median fold drug resistance score and in figure - 5.8 where input values are given for mutation position [54 = 4.5; 82 = 4.5; 90 = 5; 48 = 0.5; 84 = 3.5; 50 = 5;]. Plot points selected for this set of values.

Plot points: 101

Figure – 5.9: Plot points

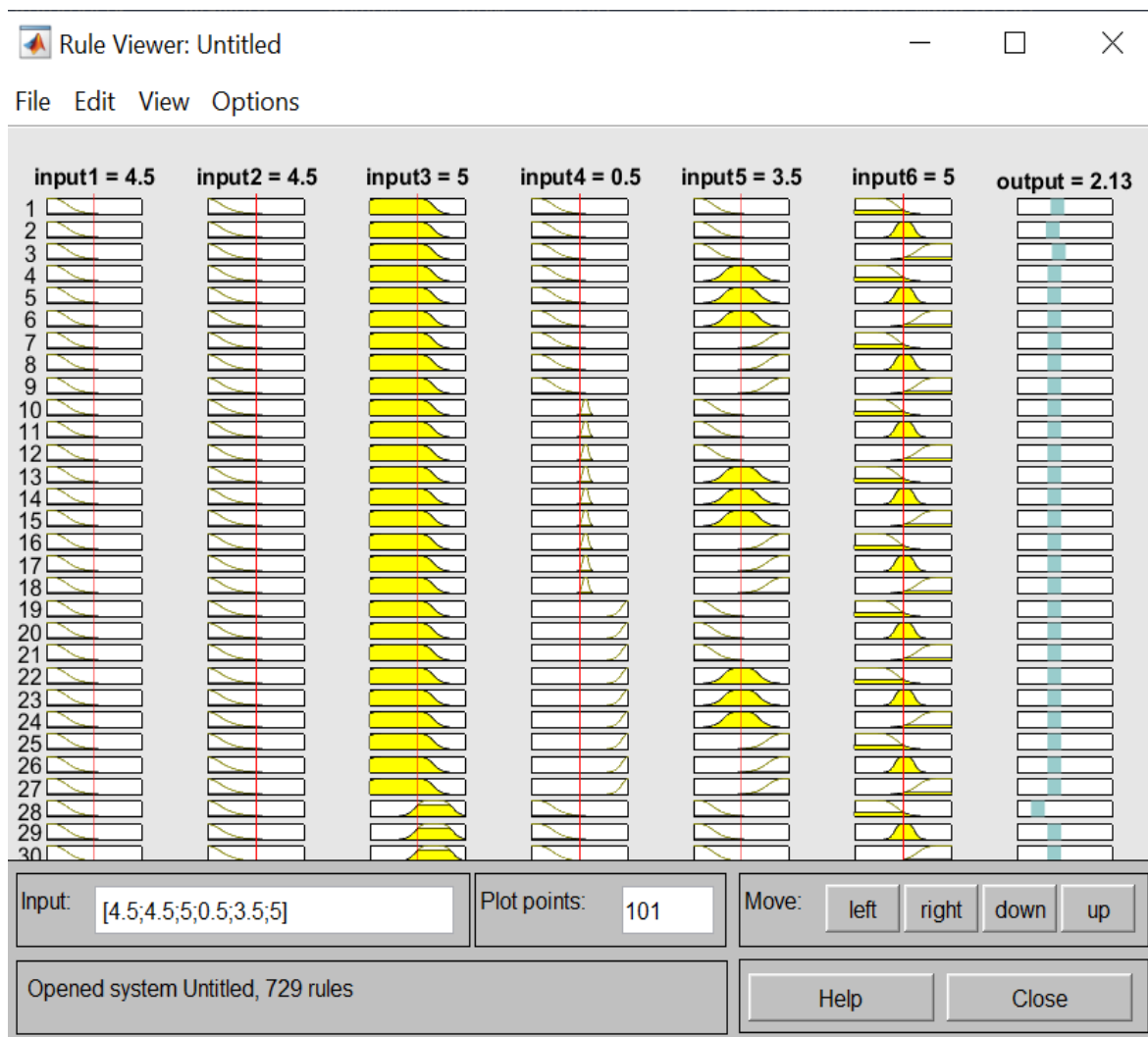


Figure – 5.10: Rule viewer of different dataset with IDV

Input:

[4.5;4.5;5;0.5;3.5;5]

Figure – 5.11: Different input values

In figure - 5.10 rule viewer represented the median fold drug resistance score and in figure - 5.11 where input values are given for mutation position [54 = 4.5; 82 = 4.5; 90 = 5; 48 = 0.5; 84 = 3.5; 76 = 5;]. Plot points selected for this set of values.

Plot points:

101

Figure – 5.12: Plot points

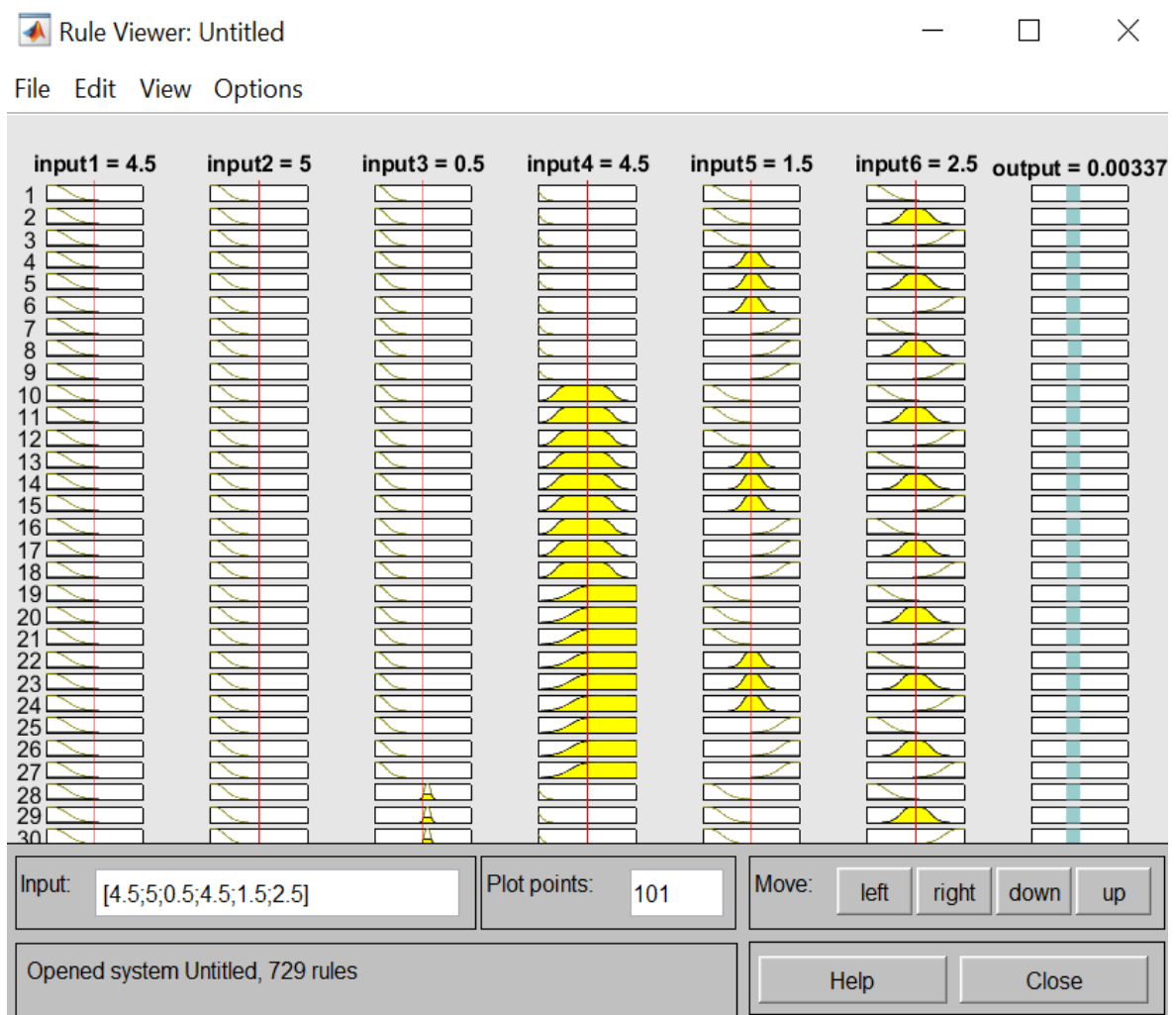


Figure – 5.13: Rule viewer of different dataset with FPV

Input:

[4.5;5;0.5;4.5;1.5;2.5]

Figure – 5.14: Different input values

In figure - 5.12 rule viewer represented the median fold drug resistance score and in figure - 5.13 where input values are given for mutation position [84 = 4.5; 90 =5; 54 = 0.5; 32 = 4.5; 47 = 1.5; 76 = 2.5;]. Plot points selected for this set of values.

Plot points:

101

Figure – 5.15: Plot points



Figure – 5.16: Rule viewer of different dataset with ATV

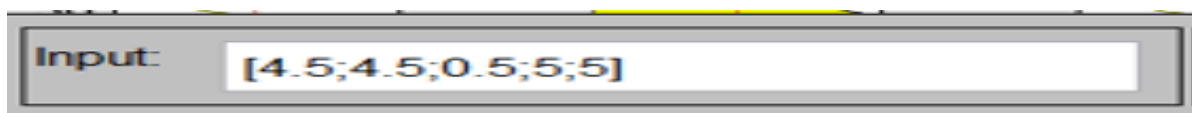


Figure – 5.17: Different input values

In figure - 5.16 rule viewer represented the median fold drug resistance score and in figure - 5.17 where input values are given for mutation position [54 = 4.5; 90 =4.5; 84 = 0.5; 82 = 5.5]. Plot points selected for this set of values.

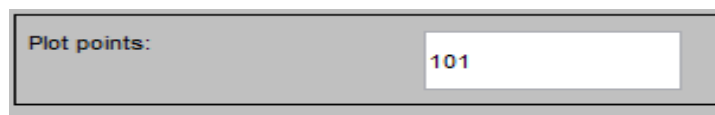


Figure – 5.18: Plot points

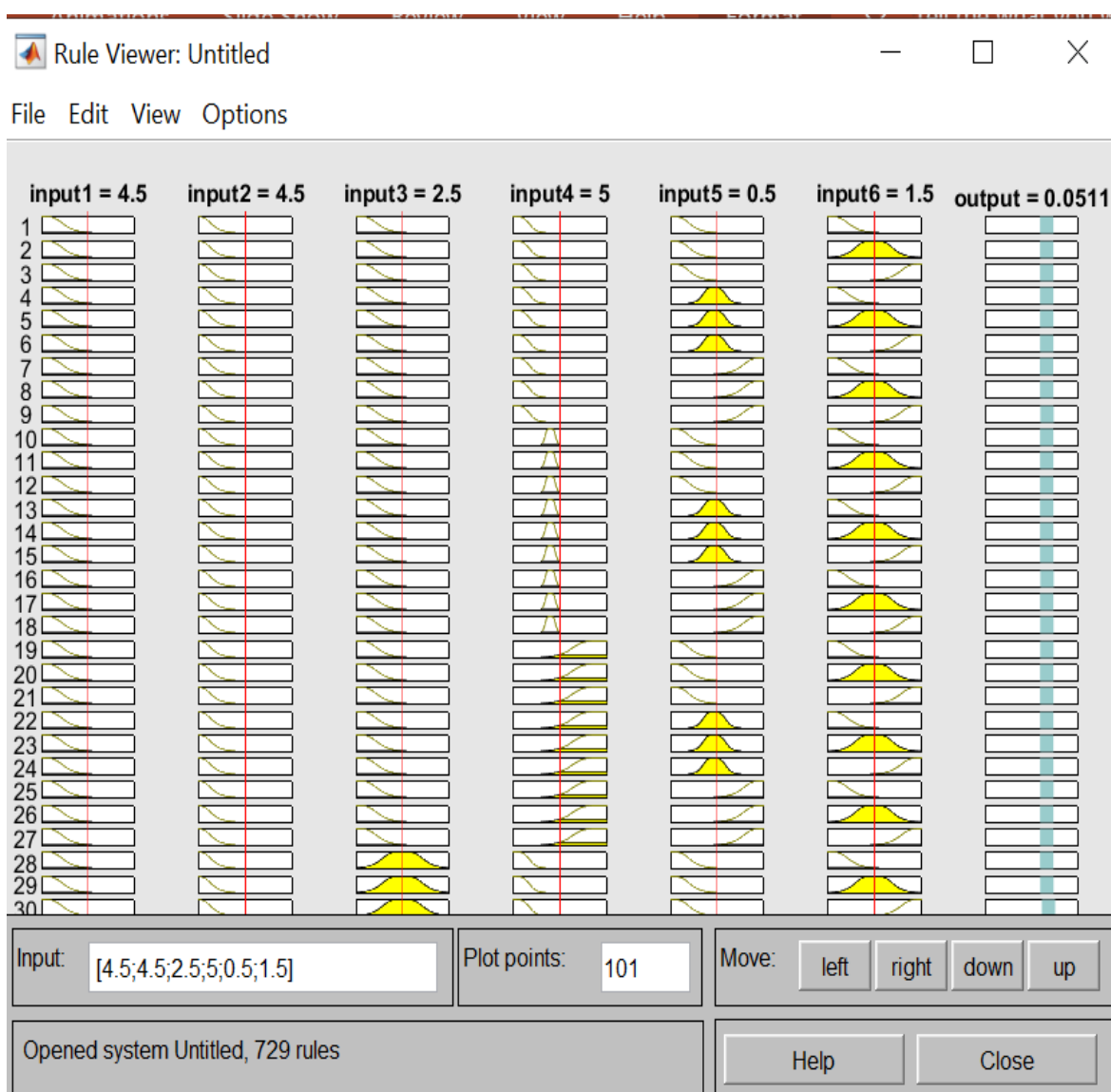


Figure – 5.19: Rule viewer of different dataset with LPV

Input:

Figure – 5.20: Different input values

In figure - 5.19 rule viewer represented the median fold drug resistance score and in figure - 5.20 where input values are given for mutation position [54 = 4.5; 47 =4.5; 82 = 2.5; 90 = 5; 32 = 0.5; 48= 1.5]. Plot points selected for this set of values.

Plot points:

Figure – 5.21: Plot points

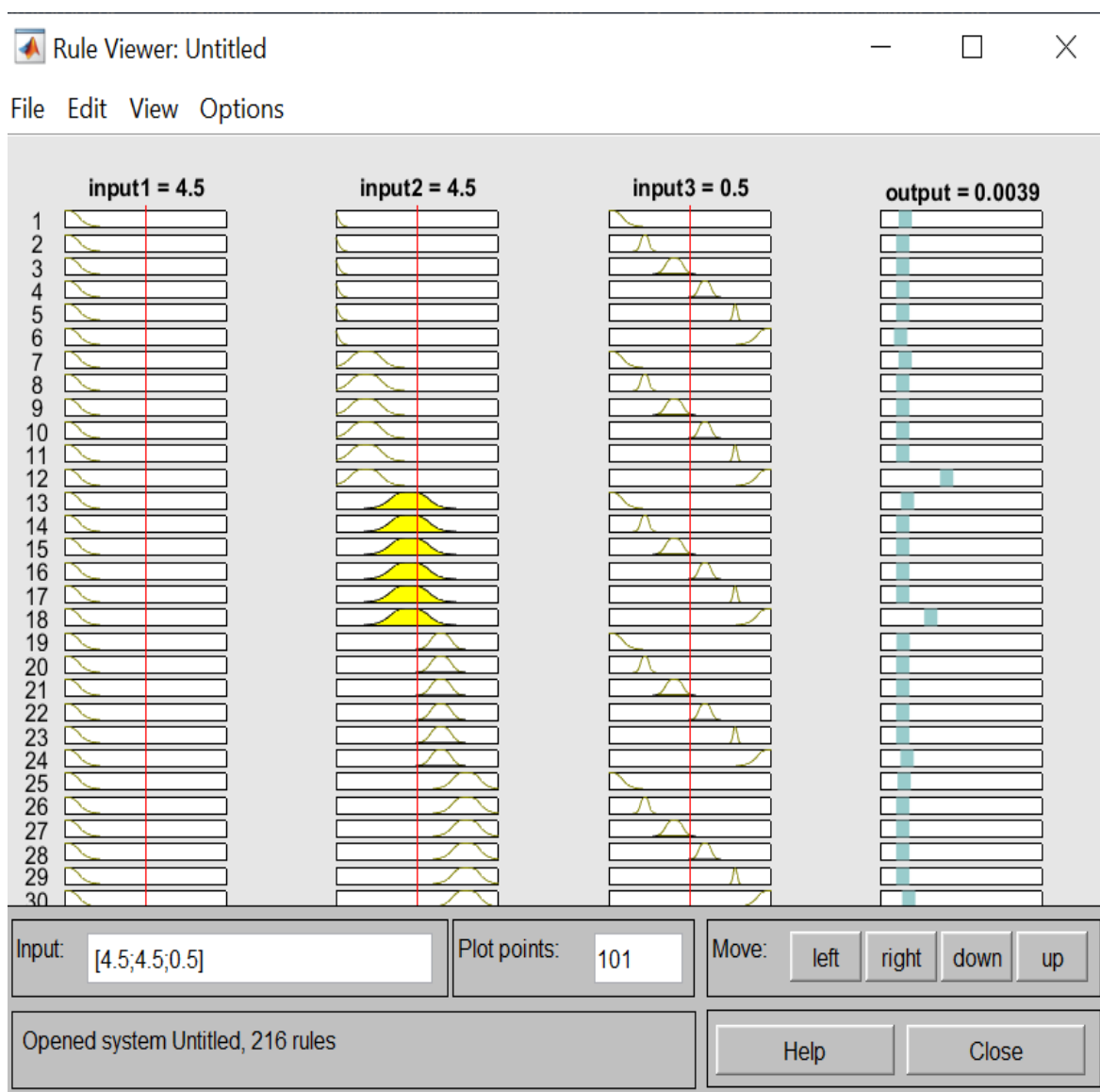


Figure – 5.22: Rule viewer of different dataset with TPV

Input:

[4.5;4.5;0.5]

Figure – 5.23: Different input values

In figure - 5.22 rule viewer represented the median fold drug resistance score and in figure - 5.23 where input values are given for mutation position [54 = 4.5; 90 =4.5; 32 = 5; 47 = 0.5]. Plot points selected for this set of values.

Plot points:

101

Figure – 5.24: Plot points

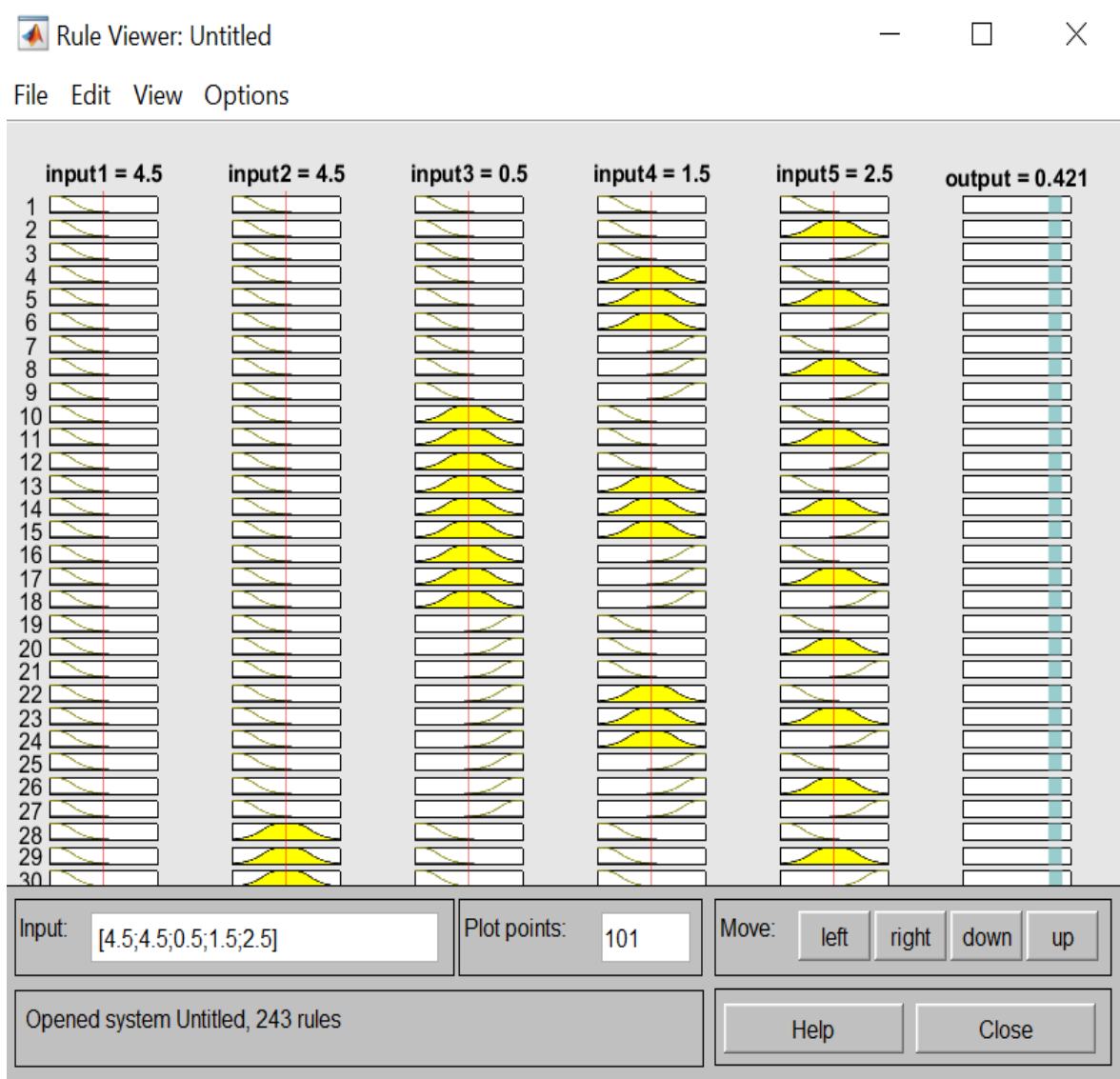
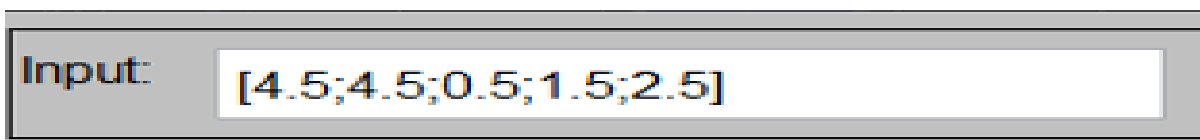


Figure – 5.25: Rule viewer of different dataset with DRV

A screenshot of a software interface showing an 'Input:' label followed by a text box containing the array [4.5;4.5;0.5;1.5;2.5].

Input: [4.5;4.5;0.5;1.5;2.5]

Figure – 5.26: Different input values

In figure - 5.25 rule viewer represented the median fold drug resistance score and in figure - 5.26 where input values are given for mutation position [54 = 4.5; 90 =4.5; 32 = 0.5; 84 = 1.5; 47= 2.5]. Plot points selected for this set of values.

A screenshot of a software interface showing a 'Plot points:' label followed by a text box containing the number 101.

Plot points: 101

Figure – 5.27: Plot points

5.1.2. SURFACE VIEW OF STANFORD HIV DRUG RESISTANCE DATASET:

Surface viewer tool provides graphical representation as in three -dimensional structure provided median fold drug resistance score and mutation positions. The figures (5.28-5.35) given below represents surface structure. According to the scores of x, y, z axis, peaks are formed and peaks represent higher median fold drug resistance.

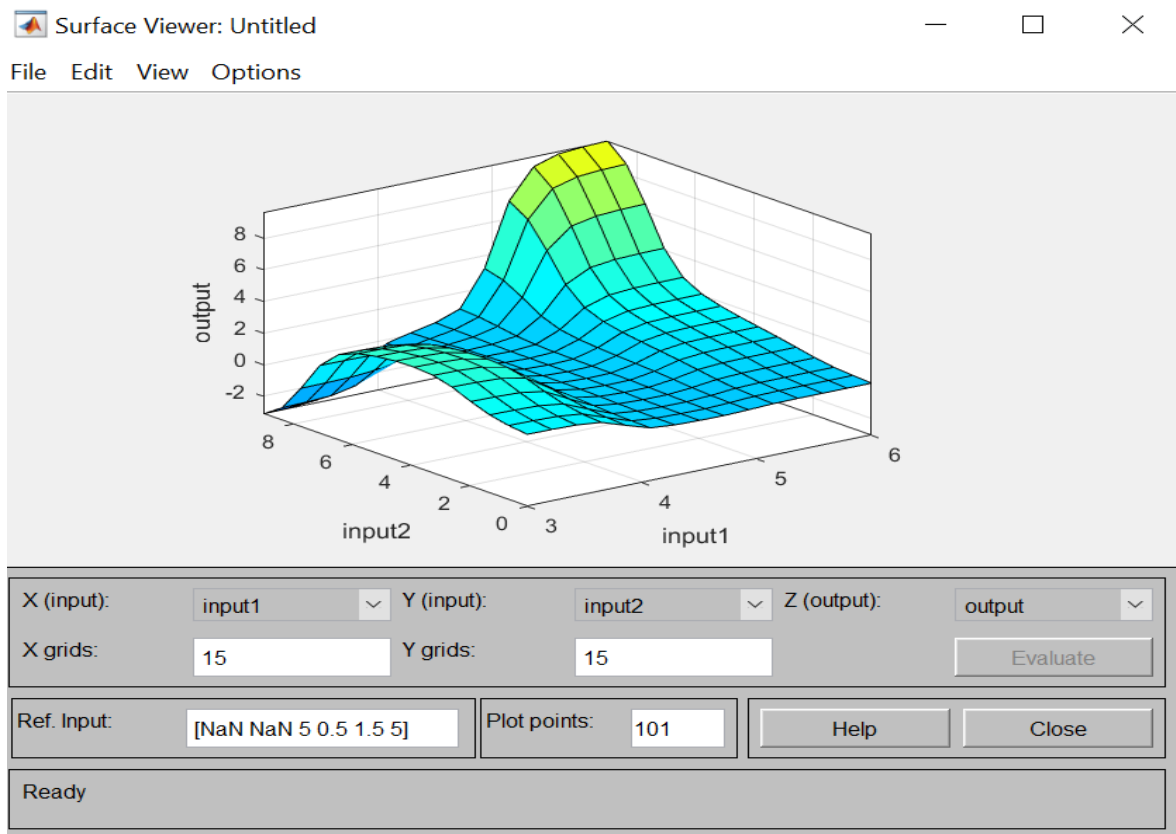


Figure – 5.28: Surface plot of NFV

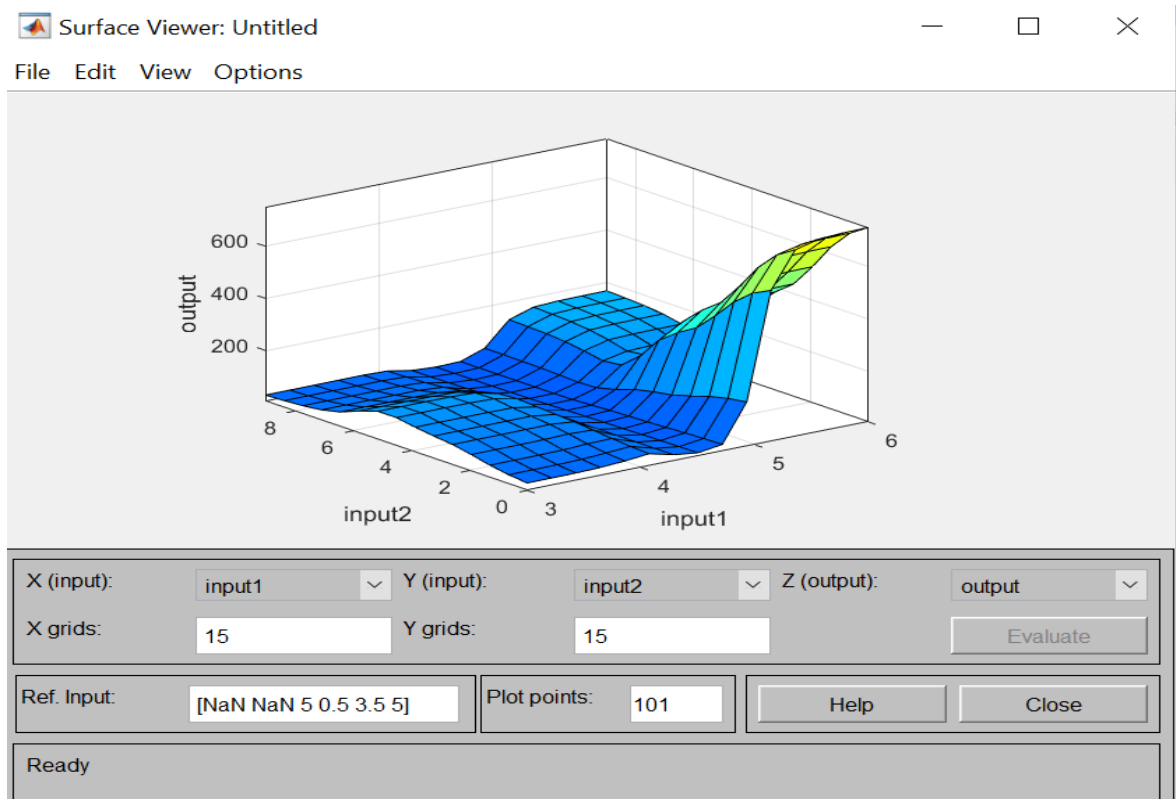


Figure – 5.29: Surface plot of SQV

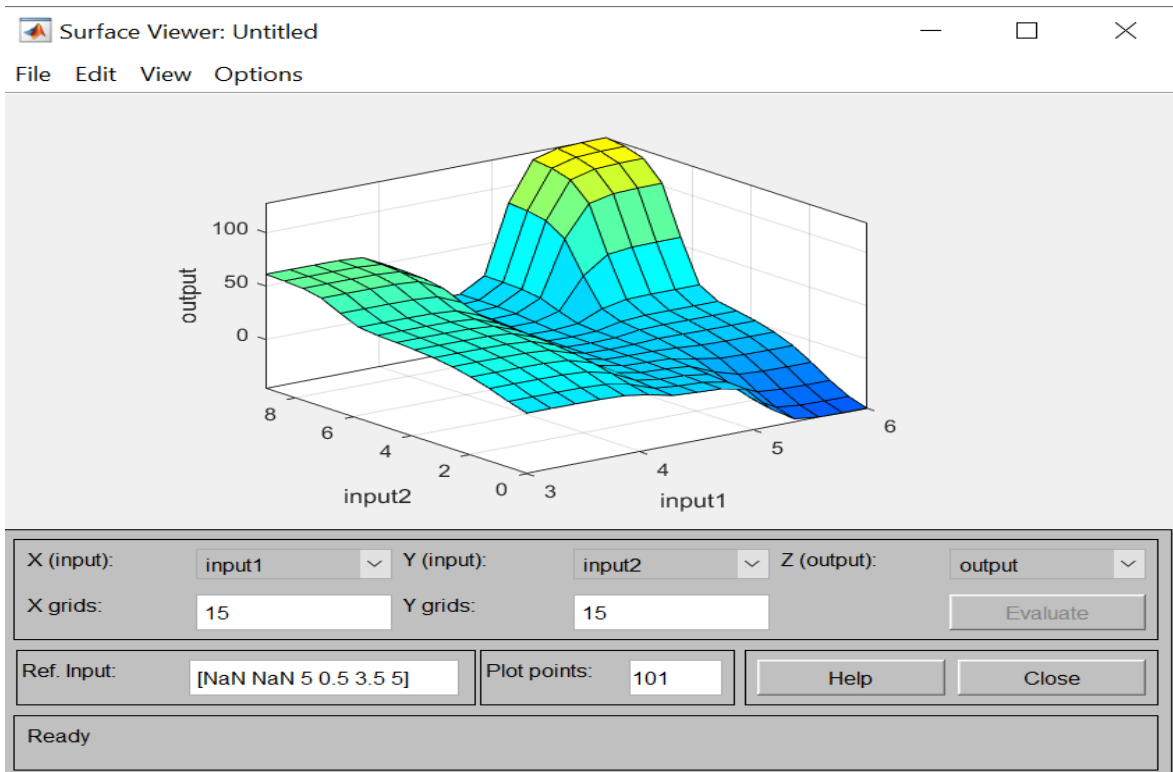


Figure – 5.30: Surface plot of IDV

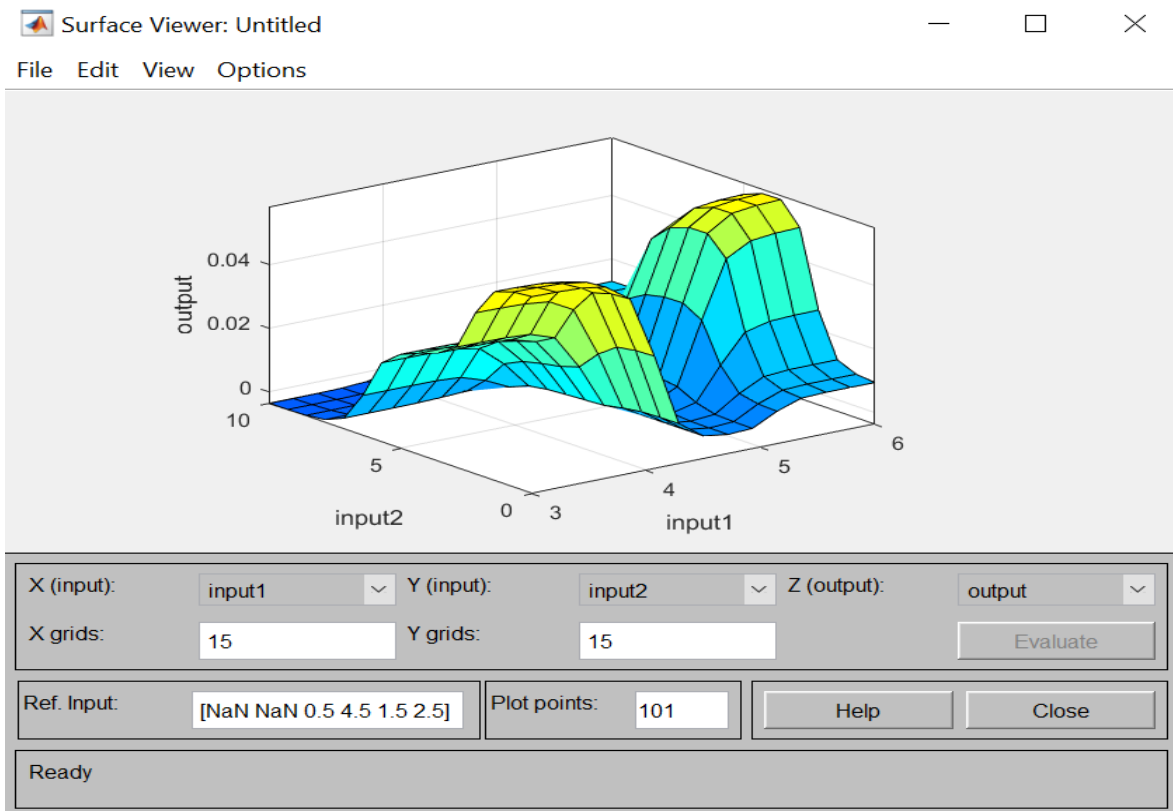


Figure – 5.31: Surface plot of FPV

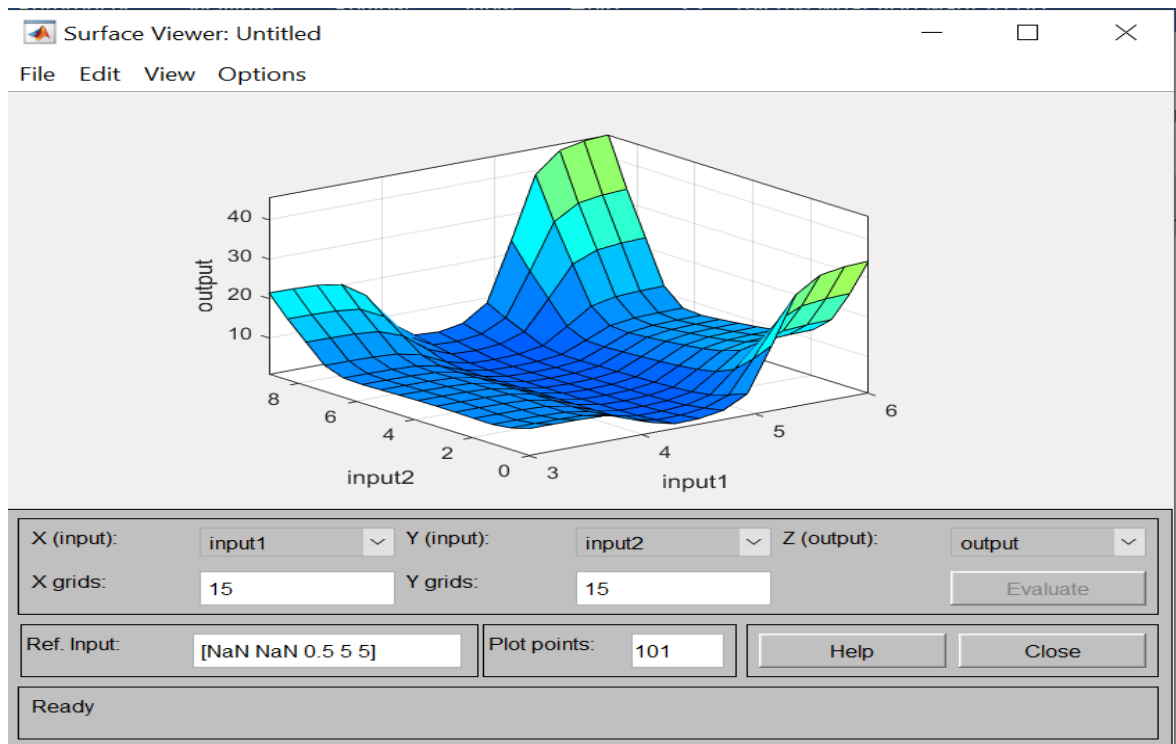


Figure – 5.32: Surface plot of ATV

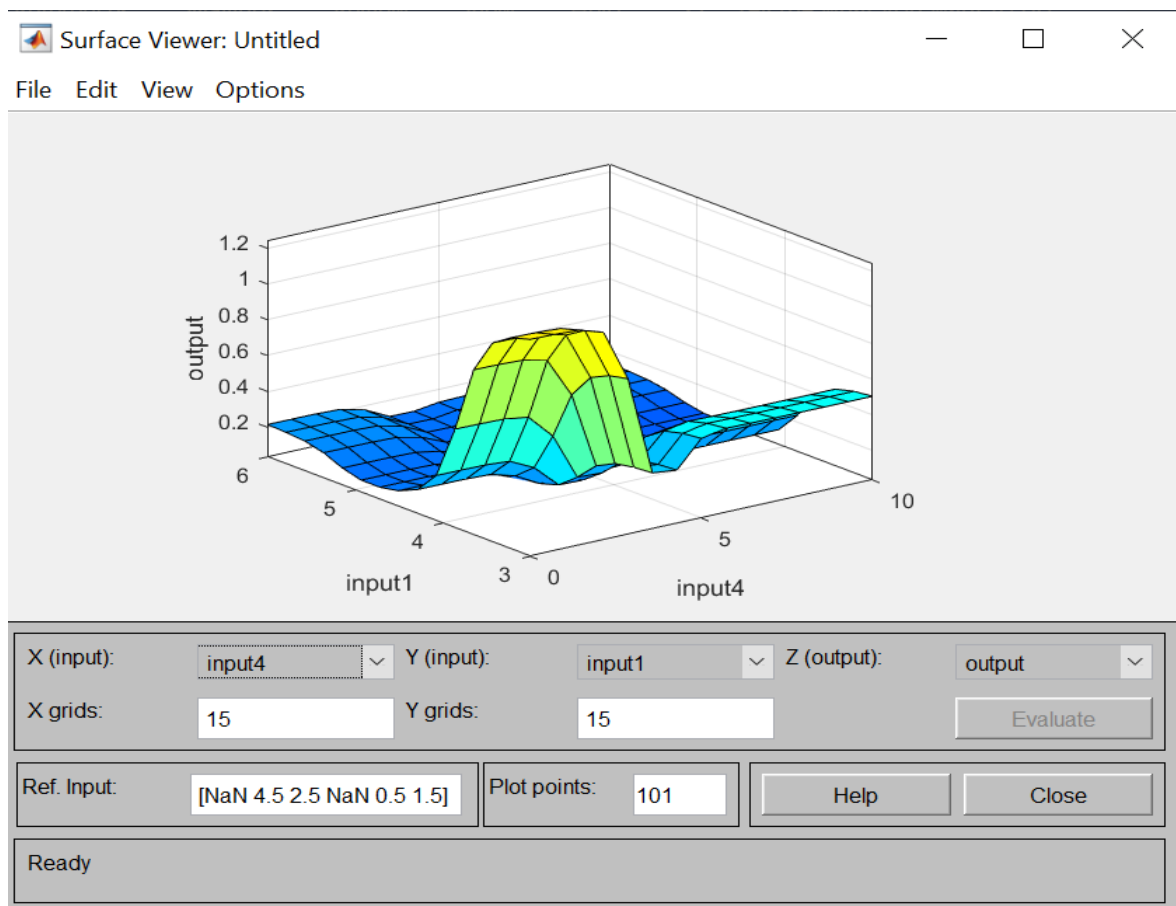


Figure – 5.33: Surface plot of LPV

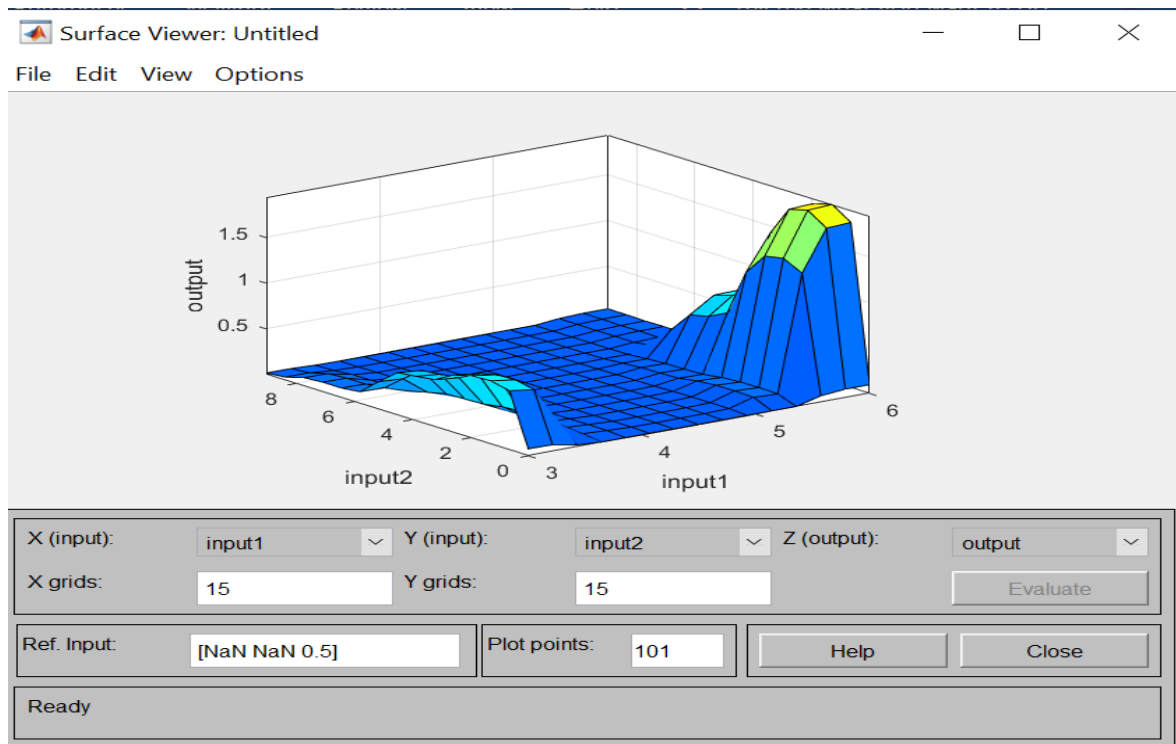


Figure – 5.34: Surface plot of TPV

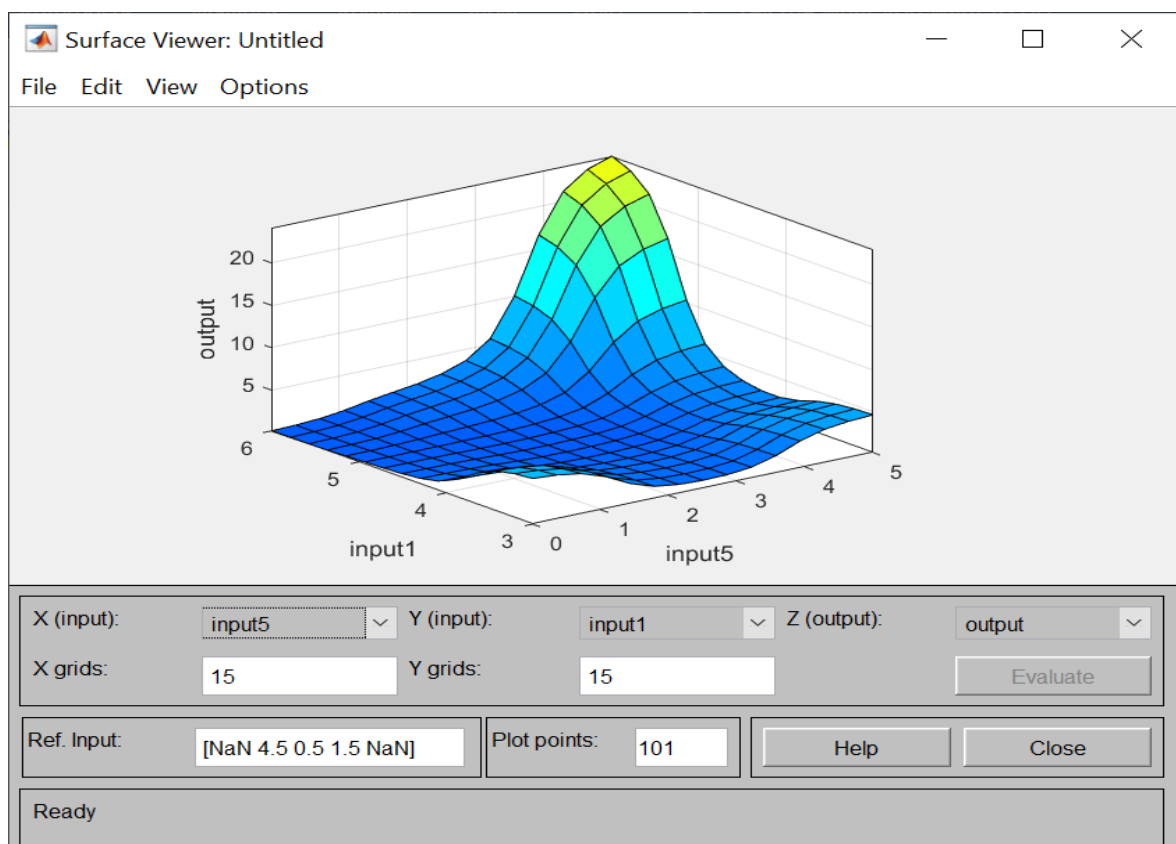


Figure – 5.35: Surface plot of DRV

5.2 EXCEL SHEET OF PROTEASE TABLE AFTER GROUPING:

Previously we have made the excel sheet of the Raw Data from the Stanford Drug Resistance Database in a normal manner of amino acid mutational numbers but we have now separate the amino acids and grouped them on the basis of Hydrophobicity – Hydrophilicity, Charged – Non Charged, Polar – Non Polar Amino Acids group and from that we have got 5 Different amino – acids mutational number from which we have made a different excel sheet based on Amino – Acid Classification.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	MUTATION PATTERNS	46	54	84	90	NFV	SQV	IDV	FPV				
2	46I,90M	5	0	0	1	22	4.2	9.1	3.2				
3	46I	5	0	0	0	11	0.9	4.4	2.1				
4	46L	3	0	0	0	8.2	1.5	2.6	1.1				
5	46L,54V,82A	3	2	0	0	39	6.2	19	4.6				
6	46I,54V,82A	3	2	0	0	16	1.6	15	2.5				
7	46I,84V,90M	5	0	2	1	48	43	28	7.7				
8	46L,54V,82A,90M	3	2	0	1	40	11	22	8.2				
9	46I,54V,84V,90M	5	2	2	1	86	161	53	20				
10	46I,54V,82A,90M	5	2	0	1	49	22	29	6.4				
11	46I,54V,76V,82A	5	2	0	0	11	1.6	37	6.5				
12	46I,88S	5	0	0	0	18	1.3	5.8	0.2				
13	46L,82A	3	0	0	0	6.1	1.3	7.4	5.8				
14	30N,46I	1	0	0	0	25	0.6	1.2	0.8	HYDROPHOBIC, NON - CHARGED, POLAR	M, N	1	
15	46I,84V	5	0	2	0	13	11	5.6	6.7				
16	46L,90M	3	0	0	1	15	3.4	7.5	1.7	HYDROPHOBIC, NON - CHARGED, NON-POLAR	A, V	2	
17	46I,54V,82F,90M	5	2	0	1	110	6.5	92	25				
18	46I,82A	5	0	0	0	67	22	41	7.8	HYDROPHOBIC, CHARGED, NON- POLAR	L	3	
19	46I,47V,54V,84V,90M	5	2	2	1	237		148	54				
20	32I,46I,82A,90M	5	0	0	1	25	4.5	21	12	HYDROPHOBIC, NEUTRALLY CHARGED, NON-POLAR	F	4	
21	46L,54V,84V,90M	3	2	2	1	89	257	69	32				
22	32I,46I,47V,54M,82A,90M	5	1	0	1	138	27	144	300	HYDROPHILIC, NON - CHARGED, POLAR	I, T, C, S	5	
23	46L,54V,82A,84V	3	2	2	0	17	79	12	13				
24	46I,54V,82A,84V	5	2	2	0	25	59	12	16				
25	46I,54V,84V	5	2	2	0	58	64	42	19				
26	46I,54V,76V,82A,90M	5	2	0	1	29	4	6.8	14				
27	46I,54V,82A,84V,90M	5	2	2	1	94	61	53	55				

Figure – 5.36: Excel sheet of protease table with 4 drugs after grouping

After that we made the final excel sheet based on the grouping of amino acids to perform the machine learning analysis.

	A	B	C	D	E
1	0	1	0	0	0.8
2	0	0	0	0	0.8
3	5	0	0	0	1.7
4	5	0	0	0	0.8
5	0	1	0	5	4
6	5	1	0	0	4.1
7	5	1	0	5	11
8	5	1	0	0	1.8
9	0	0	0	0	0.5
10	0	1	0	0	0.7
11	5	1	0	0	11
12	5	1	0	5	13
13	0	1	3	0	46
14	5	1	0	5	7.5
15	1	1	3	0	346
16	5	0	0	5	2.1
17	5	0	0	5	15
18	0	1	0	0	15
19	6	1	3	5	500
20	5	1	0	5	12
21	0	1	0	5	2.8
22	5	1	3	0	5.6
23	5	1	0	0	1.1

Figure – 5.37: Final excel sheet for analysis

This is the final excel sheet after the grouping and variation of excel sheet with most occurring mutational numbers for each and every drugs of the Protease Table of the Stanford Database.

5.2.1. RULE VIEWER OF STANFORD HIV DRUG RESISTANCE DATASET:

As datasheet of 1453 patterns of major drug resistance mutations with 287 rules is created in rule editor; so, rule viewer interpreted the median fold drug resistance score for mutation positions as well as drugs.

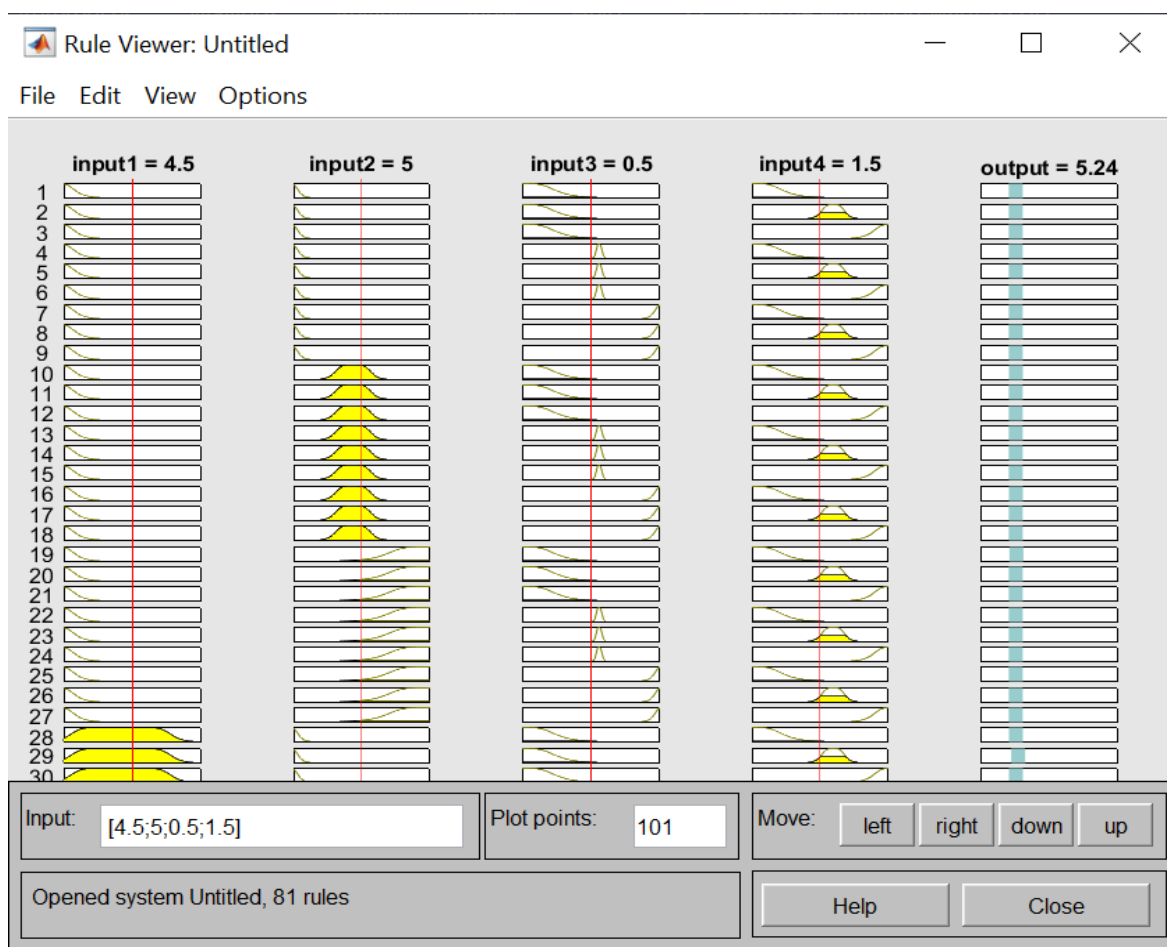


Figure – 5.38: Rule viewer of different dataset with NVF

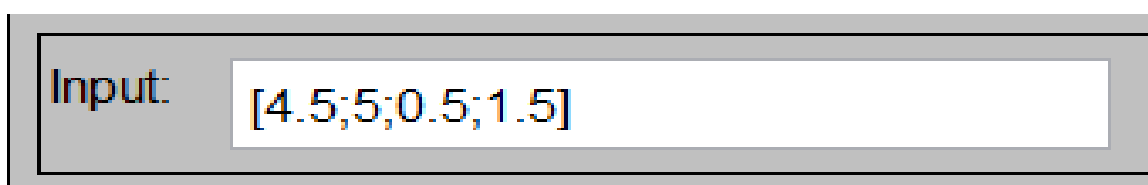


Figure – 5.39: Different Input Values.

In figure - 5.38 rule viewer represented the median fold drug resistance score and in figure - 5.39 where input values are given for mutation position [54 = 4.5; 84 = 5; 30 = 0.5; 82 = 1.5;]. Plot points of input values are in figure – 5.40.

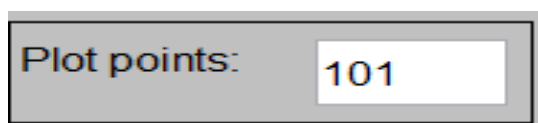


Figure – 5.40: Plot Points.

We input one set of values in figure 5.39. But what if we change input values or replace with another values, so in figure 5.43 input values are replaced by another values and drug's median fold resistance scores are interpreted accordingly.

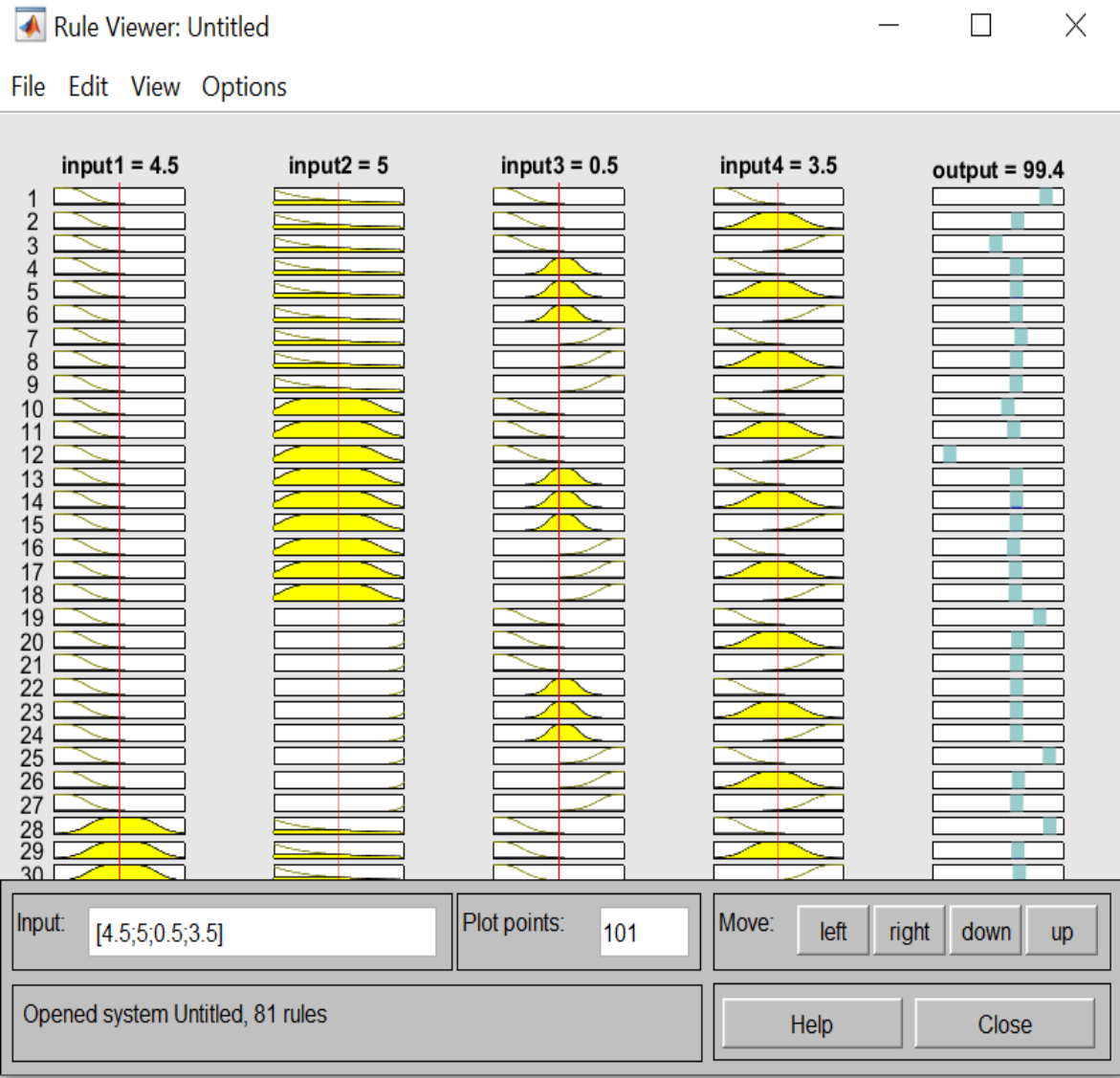


Figure – 5.41: Rule viewer of different dataset with SQV

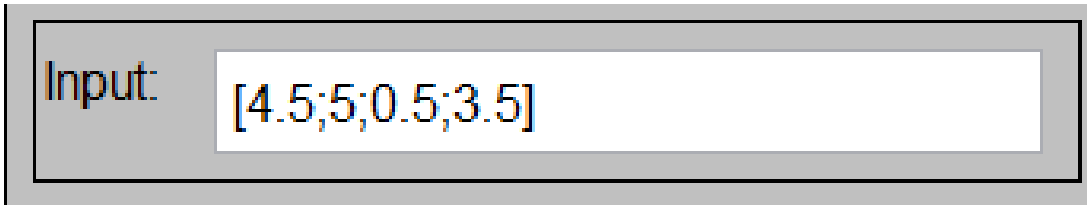


Figure – 5.42: Different Input Values.

In figure - 5.41 rule viewers represented the median fold drug resistance score and in figure - 5.42 where input values are given for mutation position [54 = 4.5; 90 = 5; 48 = 0.5; 84 = 3.5;]. Plot points selected for this set of values.

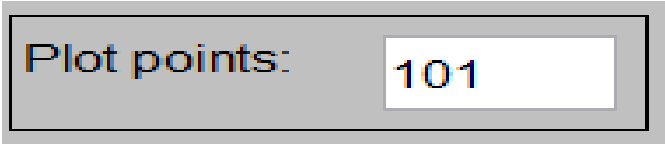


Figure – 5.43: Plot Points

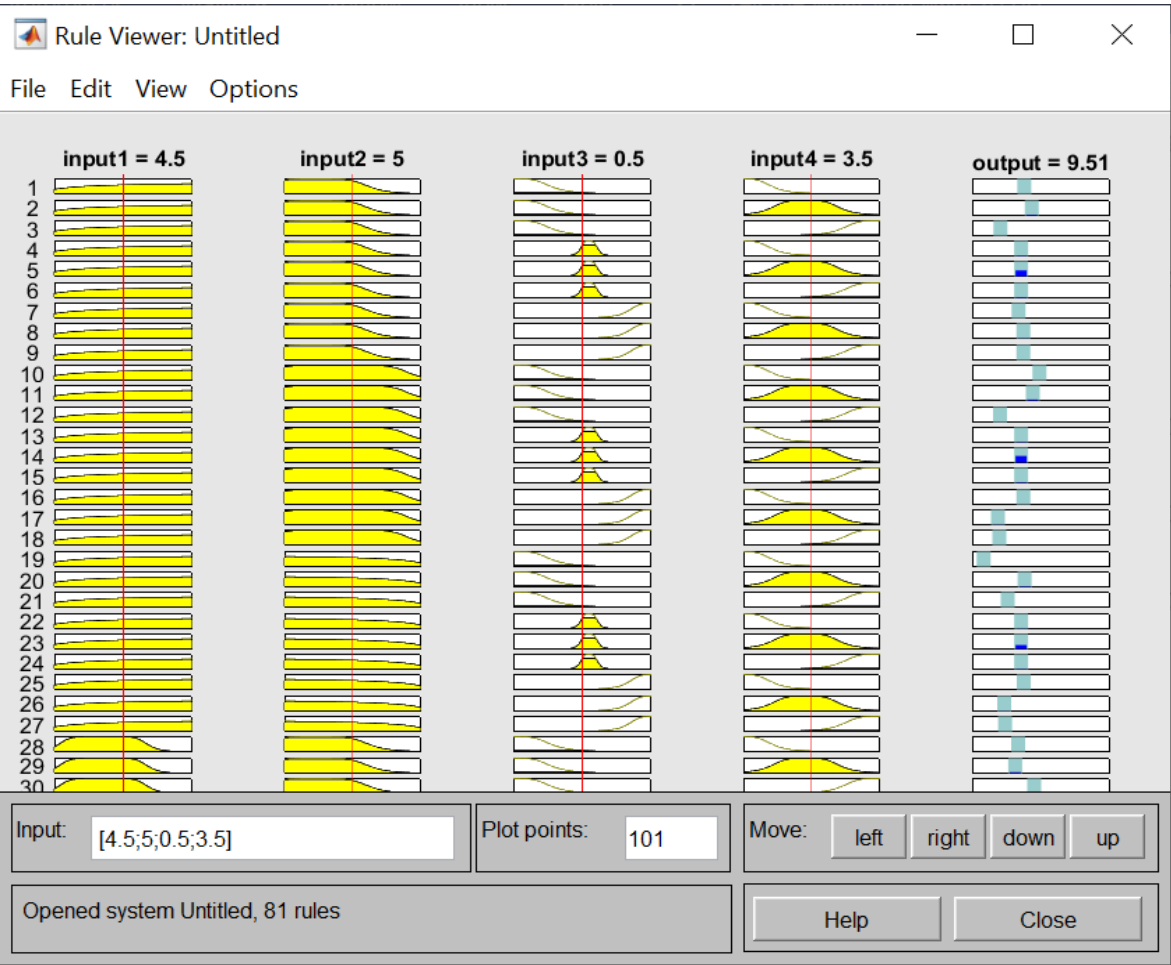


Figure – 5.44: Rule viewer of different dataset with IDV

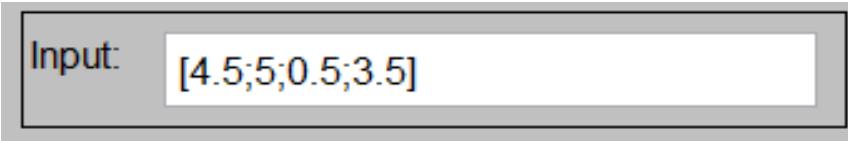


Figure – 5.45: Different Input Values.

In figure - 5.44 rule viewer represented the median fold drug resistance score and in figure - 5.45 where input values are given for mutation position [54 = 4.5; 90 = 5; 48 = 0.5; 84 = 3.5;]. Plot points selected for this set of values.

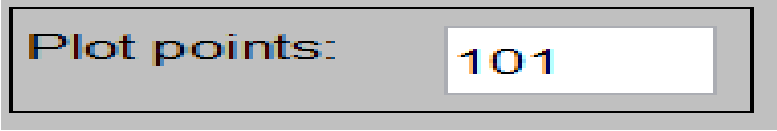


Figure – 5.46: Plot Points

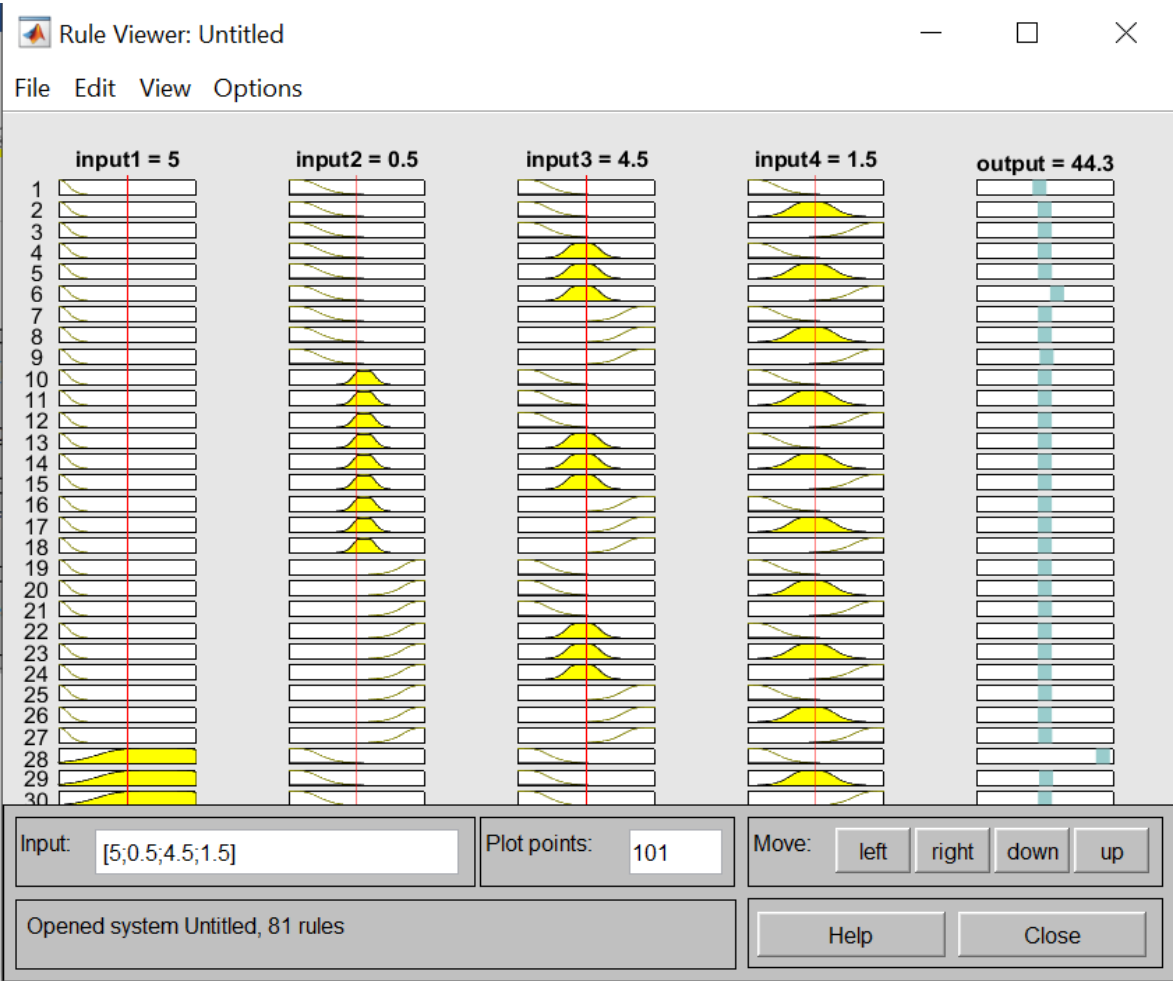


Figure – 5.47: Rule viewer of different dataset with FPV

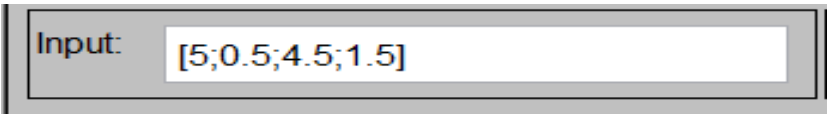


Figure – 5.48: Different Input Values.

In figure - 5.47 rule viewers represented the median fold drug resistance score and in figure - 5.48 where input values are given for mutation position [90 =5; 54 = 0.5; 32 = 4.5; 47 = 1.5;]. Plot points selected for this set of values.

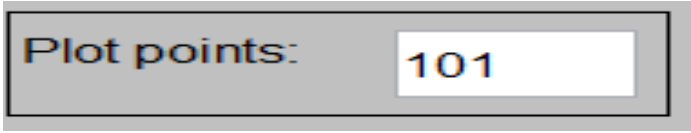


Figure – 5.49: Plot Points

5.2.2. SURFACE VIEWOF STANFORD HIV DRUG RESISTANCE DATASET:

Surface viewer tool provides graphical representation as in three-dimensional structure provided median fold drug resistance score and mutation positions. The figures (5.50-5.53) given below represents surface structure. According to the scores of x, y, z axis, peaks are formed and peaks represent higher median fold drug resistance.

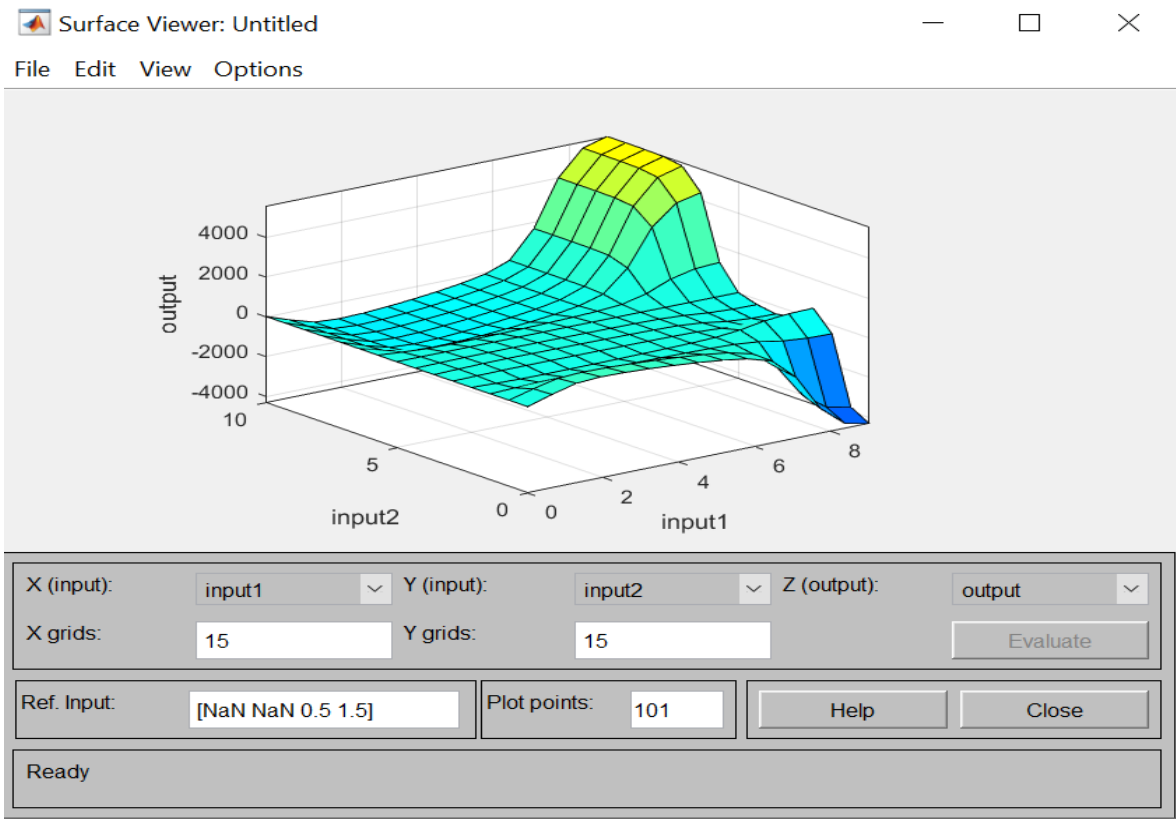


Figure – 5.50: Surface plot of NFV

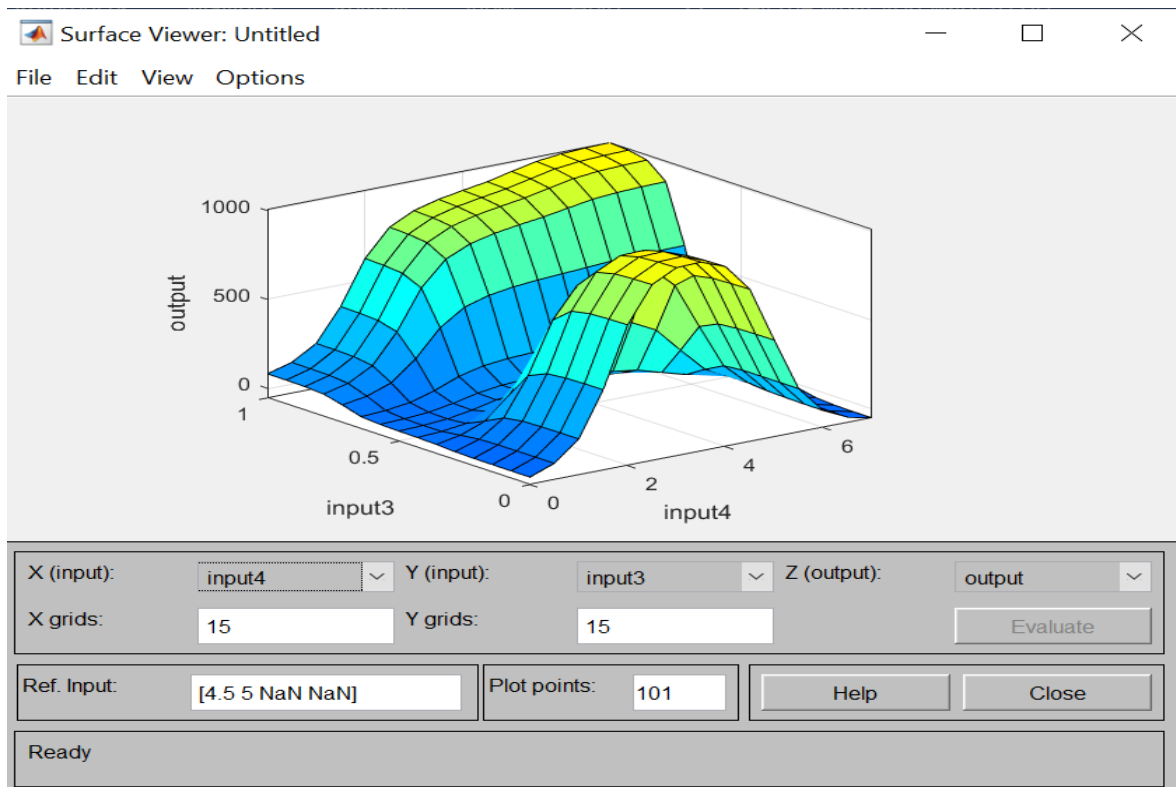


Figure – 5.51: Surface plot of SQV

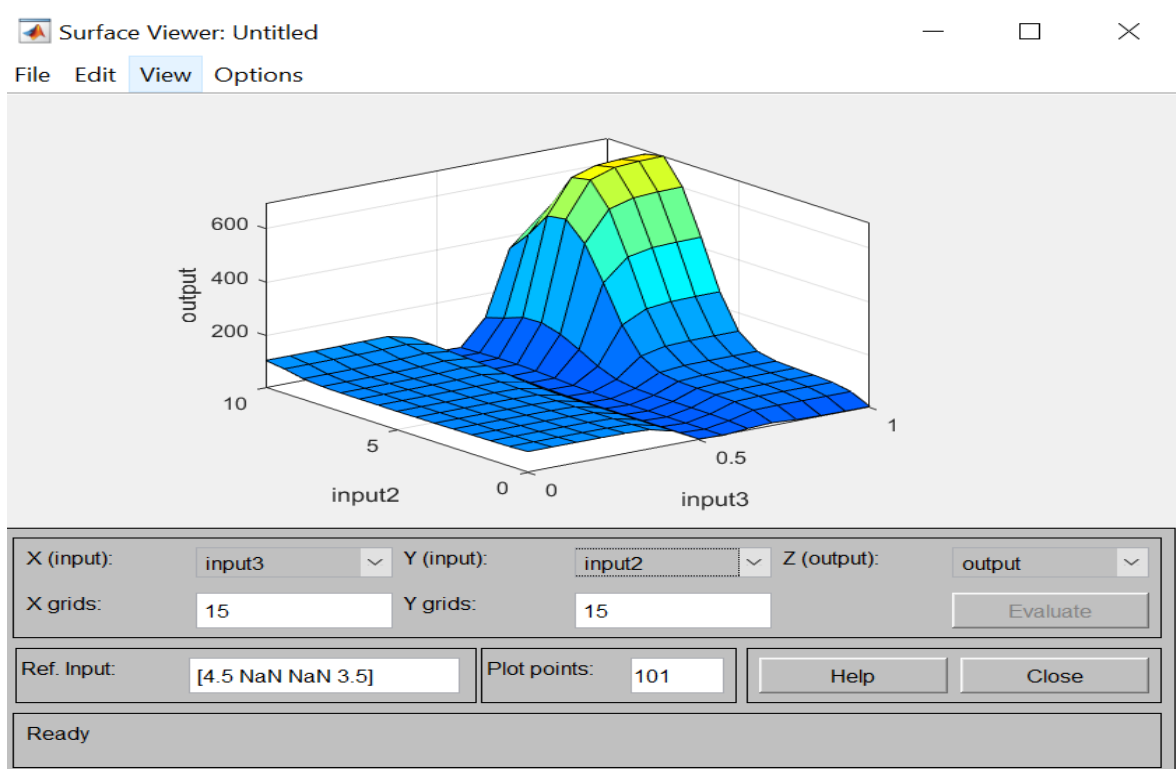


Figure – 5.52: Surface plot of IDV

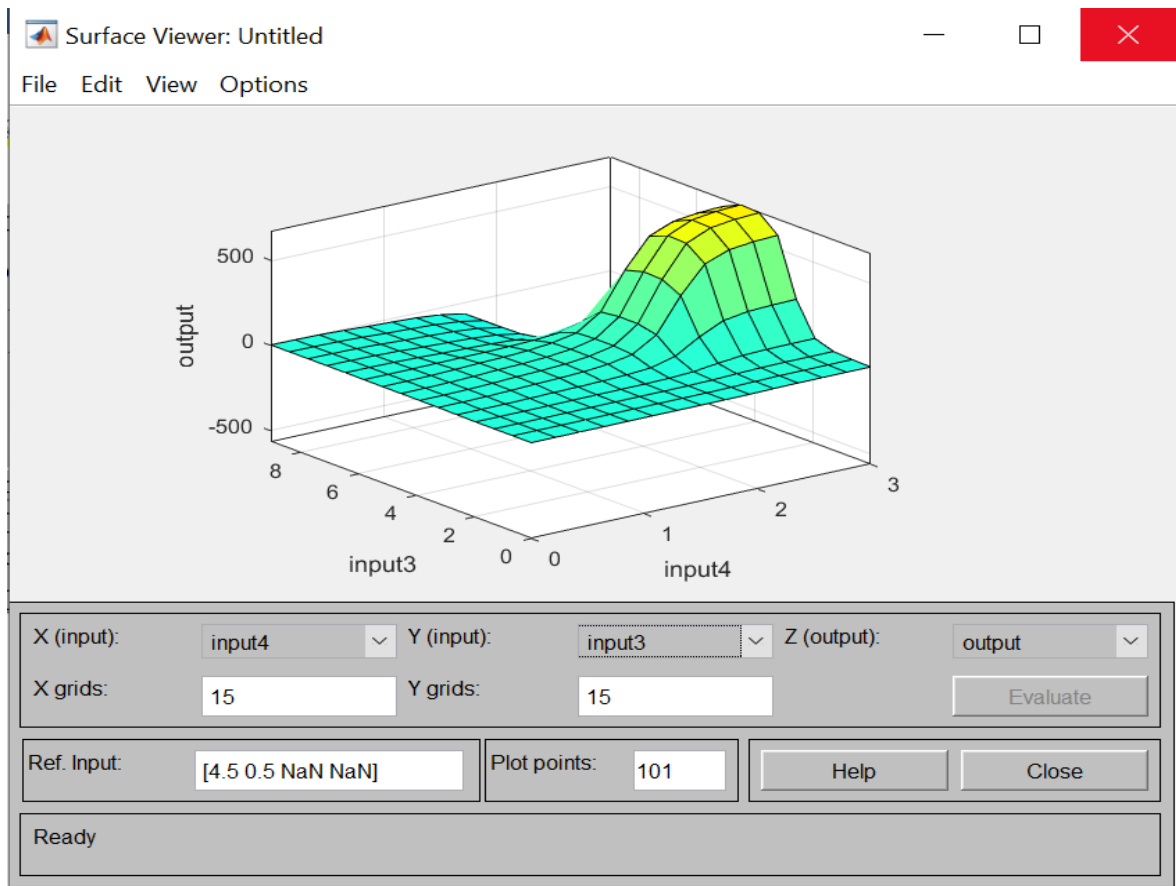


Figure – 5.53: Surface plot of FPV

CHAPTER-6

SUMMARY AND CONCLUSION

6.1. SUMMARY:

- The summary for this was twelve dataset of mutation position and eight dataset of median fold drug resistance score.
- The Raw data which we are going to grouped, filtered and analyse for our project has been taken from Stanford HIV Drug Resistance Database.
- We select protease table as our main source for the raw data from the database.
- After that we convert the data of protease table into a meaningful mutational data in the Microsoft Excel sheet then we have checked the drug resistance scores from the table.
- Then we filtered the each and every raw data (filter is a nothing but removing 0 or blanks spaces from the column).
- After filtering each and every raw we took “46 mutation” as our main column because it has more no of variations and less no of blanks and 0 compared to the other columns.
- Then we remove the outlier of all the eight drugs with filtered 46 mutation. (Removing outlier is nothing but removing blank spaces of drugs based on 46 Mutation). Then we have compared each and every drugs with 46 Mutation.
- Based on this comparison we make a graph for every drugs along with 46 Mutation.
- Based on the observation of graphs we took the outlier for each and every drug.
 - ❖ [Point – Outlier is nothing but a certain higher mutational value after which the other mutational values does not be either higher or does not occurs much times in the table].
 - ❖ Here we observe that the outlier for each and every drug is different during the process of removing outliers.
- The outlier of NFV drug is 300, above 300 we have marked as red colour in the table and same procedure has been repeated for each and every drugs.

- Then we make final excel sheet for the analysis in “Matlab” based on the previous table of mutational pattern.
- In Matlab we need to import the final table in matlab and select as numeric matrix and import it in the workspace and then in command we need to type “anfisedit”.
- Later one new interface will flash on the screen & there we have to select the workspace and load the data.
- After loading the data we have to click on Generate FIS button in the screen and in the Option “MFTYPE” we need to check which option could be better for the analysis. We have used “gbelmf” because it was the best for our analysis to getting best results.
- In Epochs, we gave 1500 and train the data for every drugs.
- After completing the training wave click on view & we have to check the Rule and surface viewer for the given drugs which give us the proper explanation about the given drugs.
- After this we get our final results.
- After completing the above analysis we have done another new analysis.
- Previously we have made the excel sheet of the Raw data in a normal manner of amino acid mutation but we have now separate the amino acids and grouped them on the basis of Hydrophobicity, Hydrophilicity, Charged, Non Charged, Polar, and Non Polar Amino acids group.
- Based on the classification of Amino acid we have made a different excel sheet and followed the same procedure which is mentioned above for the first 4 drugs and observed the result.
- Mostly we need to focus on our input and output functional data to get more accurate and good drug prediction results. We have to observe the rule viewer and surface viewer data for each and every drug which is also provided Drug resistance scores and different mutational patterns. The drugs which have the lesser drug resistance score is good to be used.

6.2. CONCLUSION:

- A mutation which is involved the 46th position have the most influence drug resistance which can be observed in the regression model of Matlab through response surface methodology. The concept of mutual amino acid residue sites could be seen here. Here the V is mutated in the 46th mutation and 47th Mutation will not be mutated with L. No other mutations were observed in the 30, 48, 50, 76, and 88th positions respectively. In the 82nd mutational position, a mutation only with A was observed and no other mutations were present. For the residues T, L, S, F no mutations were observed.
- There were no significance outliers in the model and we could observe interactions between various NFV specific mutations in the same sequence responding to the drug resistance.
- This model does not follow the linear or nonlinear regression model and here we performed the modelling of the data of Protease Inhibitor drug resistance and its susceptibility.
- Our dataset is trained with the grid partition algorithm including all the mutations. The dataset was set to train for 1500 epochs. The initial Root Mean Square Error (RMSE) was 9 which is reduced to 8.4 within the first 30 epochs and no further reduction in the errors as per our observation in results.
- For each and every drug resistance prediction the maximal number of mutation pattern was analysed, given variation in mutation position with mutual accuracy of drug resistance prediction. For a better and accurate prediction of various drug resistance, there is provided an increase in membership function for the input and output data. The median fold drug resistance score with each source data was fully analysed; thoroughly observed rule view and surface view, provided mutation pattern with different drug resistance. A variation in drug resistance fold is observed, that is interpreted and based on the resultants; the resistance of HIV drug is predicted and called out for lesser median fold drug resistance score. The drug with lesser resistance score is more to be used, in general.

REFERENCES

1. Alessandra Lumini , Loris Nanni et al, “Machine learning for HIV-1 protease cleavage site prediction”, Oct 2006.
2. Antonio Carvajal-Rodríguez et al, “The Importance of Bio-Computational Tools for Predicting HIV Drug Resistance”, Feb 2007.
3. ChenHsiang Shen, Xiaxia Yu, Robert W. Harrison & Irene T. Weber et al, “Automated prediction of HIV drug resistance from genotype data”, 31 August 2016.
4. Harish Bhaskar , David Hoyle , Sameer Singh et al, “Machine learning in bioinformatics: A brief survey and recommendations for practitioners” , Nov 2006.
5. Ian Witten, Eibe Frank et al, “Data Mining: Practical Machine Learning Tools and Techniques”, Jan 2005.
6. Ludmila Kuncheva et al, “Combining Pattern Classifiers: Methods and Algorithms”, Jan 2004.
7. M.J. Kozal et al, “Drug-resistant human immunodeficiency virus”, February 2009.
8. Mattia Prosperi, Andrea De Luca et al, “Computational models for prediction of response to antiretroviral therapies” , Apr 2012.
9. Mattia Prosperi, Giovanni Ulivi et al, “Evolutionary Fuzzy Modelling for Drug Resistant HIV-1 Treatment Optimization”, 2008.
10. Moses E.Ekpenyong Philip I. Etebong Tenderwealth C. Jackson et al, “Fuzzy-multidimensional deep learning for efficient prediction of patient response to antiretroviral therapy” , July 2019.
11. Seth Inzaule and Silvia Bertagnolio (WHO, HIV Department), under the coordination of Meg Doherty (WHO, HIV Department) et al, “HIV Drug Resistance Report”, 2019.
12. Sorin Draghicir, Brian Potter et al, “Predicting HIV drug resistance with neural networks” , Feb 2003.
13. UNAIDS, “Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)”, Jan 2012.