Heaven's Light is Our Guide

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

Bioactivity Prediction from Target Proteins using Machine Learning Models

Author

Tasneem Sakif Ibne Alam

Roll No.: 1603094

Department of Computer Science & Engineering

Rajshahi University of Engineering & Technology

Supervised by

Prof. Dr. Md. Rabiul Islam

Department of Computer Science & Engineering

Rajshahi University of Engineering & Technology

ACKNOWLEDGEMENT

At first, we would like to thank the Almighty Allah for giving us the opportunity and enthusiasm along the way for the completion of our thesis work.

I would like to express my sincere appreciation, gratitude, and respect to my supervisor Prof. Dr. Md. Rabiul Islam, Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi. Throughout the time he has not only given me technical guidelines, advice and necessary documents to complete the work, he has also given me continuous encouragement, advice, helps and sympathetic co-operation whenever he deemed necessary. His continuous support was the most successful tool that helped me to achieve my result. Whenever I was stuck in any complex problems or situation he was there for me at any time of the day. Without his sincere care, this work not has been materialized in the final form that it is now at the present.

I am also grateful to all the respective teachers of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi for good & valuable suggestions and inspirations from time to time.

Finally, I convey my thanks to my parents, friends, and well-wishers for their constant inspirations and many helpful aids throughout this work.

10 October, 2022 Tasneem Sakif Ibne Alam

RUET, Rajshahi

Heaven's Light is Our Guide

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

CERTIFICATE

This is to certify that the thesis report entitled ***"Bioactivity Prediction from Target Proteins using Machine Learning Models"*** *submitted by* ***Tasneem Sakif Ibne Alam, Roll no: 1603094*** *in partial fulfillment of the requirement for the award of degree of Bachelor of Science in Computer Science & Engineering of Rajshahi University of Engineering & Technology, Bangladesh is a record of the candidate's own work carried out by him under my supervision. This thesis has not been submitted for the award of any other degree.*

Supervisor

Prof. Dr. Md. Rabiul Islam
Department of Computer Science
& Engineering
Rajshahi University of Engineering
& Technology
Rajshahi-6204

External Examiner

Shyla Afroge
Assistant Professor
Department of Computer Science
& Engineering
Rajshahi University of Engineering
& Technology
Rajshahi-6204

ABSTRACT

With the emergence of new diseases in rapid scale and the dangerous and sometimes life threatening effects of those diseases in human body, acceleration in drug discovery or bioactivity prediction has become a necessity nowadays. Computational approaches have replaced manual approaches for cost and time efficiency. Different machine learning and deep learning models have already been tested out in the field of drug discovery and have shown satisfactory results. Nevertheless, there is always scope for further research in the comparatively new field of bioactivity prediction because of the availability of large scale datasets. Some of the previous researches showed machine learning models like Support Vector Machine perform on par with deep learning models. Our research work is to implement different machine learning models in different datasets of target proteins and find out if there is a best model for bioactivity prediction. We implemented Random Forest, Support Vector Machine and Feed Forward Neural Network models on two different datasets: Plasmodium Falciparum and Leukocyte Elastase where the former consists ten times more data sample than the latter. We chose the datasets from ChEMBL website, preprocessed those to make the data more uniform for implementing the proposed models. After data preprocessing, features or molecular descriptors were extracted using PaDEL descriptor in the form of PubChem fingerprints. Then, the models were implemented on the data. Feed Forward Neural Network outperformed the other models for both datasets producing 99% accuracy for Leukocyte Elastase and 97% accuracy for Plasmodium Falciparum dataset where each of the other two models gave around 90% accuracy for Leukocyte Elastase and around 80% accuracy for Plasmodium Falciparum dataset. So we concluded Feed Forward Neural Network as the best model for bioactivity prediction.

CONTENTS

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Introduction

Human body is a complex yet highly structured network composed of living cells. The cells inside of the human body work closely with one another. The cells perform different activities which are necessary for sustaining life and ensuring well-being of the body.

Proteins are basic building blocks of the human cell. According to scientists a single cell is made up of 42 million protein molecules [1]. So when proteins function properly human cells also function properly. But when the structure of protein changes inside a cell, the human body can become dysfunctional which is known as being infected with disease. A single protein is at the root of causing disease. The single protein then transmits to several other proteins in multiple pathways to spread the disease. The one who is infected with disease is identified as a patient.

To alter a patient's dysfunctional state to the normal or functional state, firstly, the relevant protein or organism associated with the disease and their role in causing the disease, needs to be identified. The protein which is the cause of the disease is known as target protein.

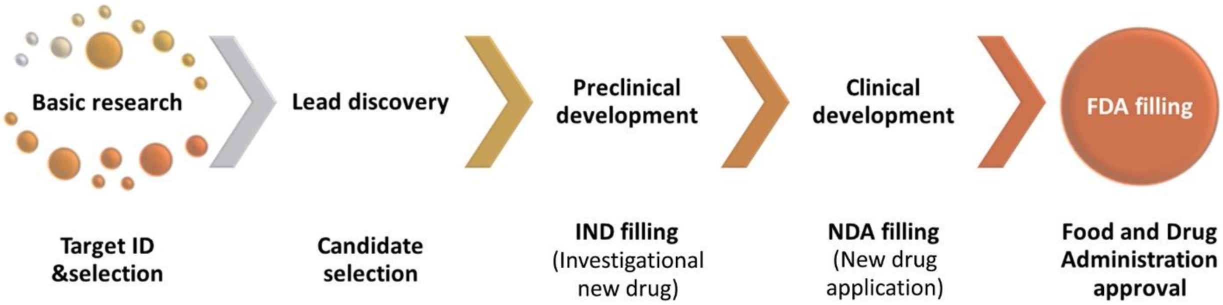


Figure 1.1: Drug discovery Process [2]

A drug is a biological entity (antibodies) or chemical entity (small molecules) that can control the course of a disease state by interacting with its target protein. Identifying the best target for treating and preventing disease is the first step of drug discovery. Knowing how protein pathways, by which a particular disease spreads, work help to identify appropriate target protein.

A drug can have a neutral or beneficial or adverse effect on target protein. Biological activity or bioactivity describes these effects of a drug on living matter. The goal of drug discovery is to find out the compounds that modulate the target in a way which is expected to alter the disease. These compounds are known as hits. After some processing a hit can be considered as a drug candidate.

There are different approaches for drug discovery ranging from manual to computational approaches. And in computational approaches there are multiple subdivisions including machine learning methods and deep learning methods.

1.2 Motivation

Health is arguably the finest blessing of a human being's life. Good health indicates all the bodily functions are working properly which results in peace of mind. A disease directly affects a well functioned body which hampers the natural state of the body and the mind. According to a large scale study around 95% of the world population suffer from some sort of diseases [3].

And the variety of diseases are also increasing in recent years. Some of these diseases are fatal, resulting in many deaths. Some of the diseases are highly infectious. One recent example of an infectious disease would be COVID-19 Coronavirus, which resulted in over 6.5 million deaths across the world in less than three year time [4].

With the rapid advent of new diseases every year, finding the right cure for the diseases in a short time has become a necessity. Before technological advancement drug discovery for a disease took a long amount of time. Now with the technological advancement and discovery of many powerful computational methods together with the huge amount of dataset of targets that are available, rapid advancement in drug discovery is possible. All we need to do is to implement the right model in the right way.

There has been various methods which are already implemented to find drug discovery. With the availability of new large data, not all methods have been tested for all datasets.

To connect the dots between the dataset and the computational methods, namely machine learning models in drug discovery, has motivated us to compare different machine learning models to accelerate the process of drug discovery.

1.3 Problem Statement

ChEMBL database can provide us with large dataset of different target proteins which can be used for drug discovery. The data however needs a lot of preprocessing before the proposed machine learning and deep learning models can be used on them.

The raw data that are found in ChEMBL are not uniform. The proposed models will not give correct results if those are implemented on the raw data. So Uniform data needs to be selected. Also the duplicate data needs to be balanced out for finding the proper result.

Whether a protein is active or inactive which is known as the bioactivity of the protein needs to be calculated based on the standard value.

To extract the molecular descriptors, which are the features for our proposed models, from the target protein necessary steps need to be followed.

Finally, we will implement machine learning models in uniquely prepared uniform dataset to analyze and compare their performances and to find out if one model significantly outperforms other models.

1.4 Challenges

While performing bioactivity prediction different challenges arise about which we need to be wary of. Some of the challenges are:

- Imbalanced dataset.
- Duplicate data.

- Extracting important features from the data.
- To select the parameters of the machine learning models correctly.
- To select the correct parameters for FNN.
- To select the hidden layers of FNN.
- To select the best evaluation method
- To handle overfitting if occurs.
- To handle underfitting if occurs.

These are just some of the challenges. Other challenges of our research are discussed throughout the paper. The methods to overcome the challenges are also discussed at the same time.

1.5 Research Objectives

The main objective of our research is to compare machine learning models by implementing the models on a large uniform dataset.

Besides this some other objectives of our research are:

- To learn to prepare a uniform dataset by data preprocessing.
- To Learning about target proteins and how they propagate disease.
- To learn and prepare molecular descriptors for our dataset using padel descriptor.
- To learn the working procedures of different machine learning models of classification like Random Forest Classifier, Support Vector Machine etc. and implementing them on our prepared dataset.
- To learn the working procedures Feed Forward Neural Network architecture models and implementing it on our prepared dataset.
- To learn whether our proposed models perform differently on a smaller dataset compare to the large dataset by implementing the models on two different datasets.

- To learn and work with PubChem fingerprint features.
- To compare our dataset with other dataset of other researches.
- To compare our results with other results from other papers.
- To data find the best model.

1.6 Thesis Organization

The rest of our thesis is organized in the following order:

CHAPTER 2

Literature Review

This chapter will be a discussion on some of the previous works conducted in the field of drug discovery using various machine learning and deep learning methods.

CHAPTER 3

Background Study

In this chapter, we will further elaborate on the background study that we have conducted for our research which will include study regarding the target proteins and discussion about the different machine learning models and deep learning model.

CHAPTER 4

Methodology

In this chapter, we will discuss about the criteria for model selection for our research. We will also discuss about the models that we used for our research which are Random Forest, Support Vector Machine and Feed Forward Neural Network.

CHAPTER 5

Implementation

This chapter will discuss the implementation of our models.

CHAPTER 6

Experimental Results and Performance analysis

In this chapter we will analyze and compare the results of our models.

CHAPTER 7

Conclusion and Future Works

This chapter will close out our research work mentioning our research summary, limitations and future works.

1.7 Conclusion

In this chapter, we introduced our research work. We discussed about what motivated us to do this research. We also discussed about research challenges and objectives. We also gave a brief overview on our thesis organization. Much elaborated discussion about our research will be in the following chapters

CHAPTER 2

Literature Review

2.1 Introduction

This chapter is a discussion on some of the previous works conducted in the field of drug discovery using various machine learning and deep learning methods.

Ranging from simple to complex various neural network models have been used in drug discovery [5-13]. Among the deep learning models FNN, LSTM [9], CNN [13], Graph neural network [7] have been implemented on different datasets for drug discovery.

For working with continuous variable different machine learning regression models like Logistic regression [14] and Naïve Bayes regression [15] are also used in bioactivity prediction. Among the machine learning classification models Random Forest [16], Decision Tree [17] have also been tested. In a Study Support Vector Machine (SVM) models performed even better than deep learning models in bioactivity prediction [18]. In another large scale study, it was concluded that Feed Forward Neural Network (FNN) model outperformed all other models [19].

In the rest of chapter, we will discuss some of these different models, their performance along with their limitations.

2.2 Literature Review

In a paper titled "Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction" by M. Robinson, R. Glen and A. Lee, published in Journal of Computer-Aided Molecular Design, in 2020 [18], they concluded that machine learning models like Support Vector Machine (SVM) performs on par with deep learning models in bioactivity prediction.

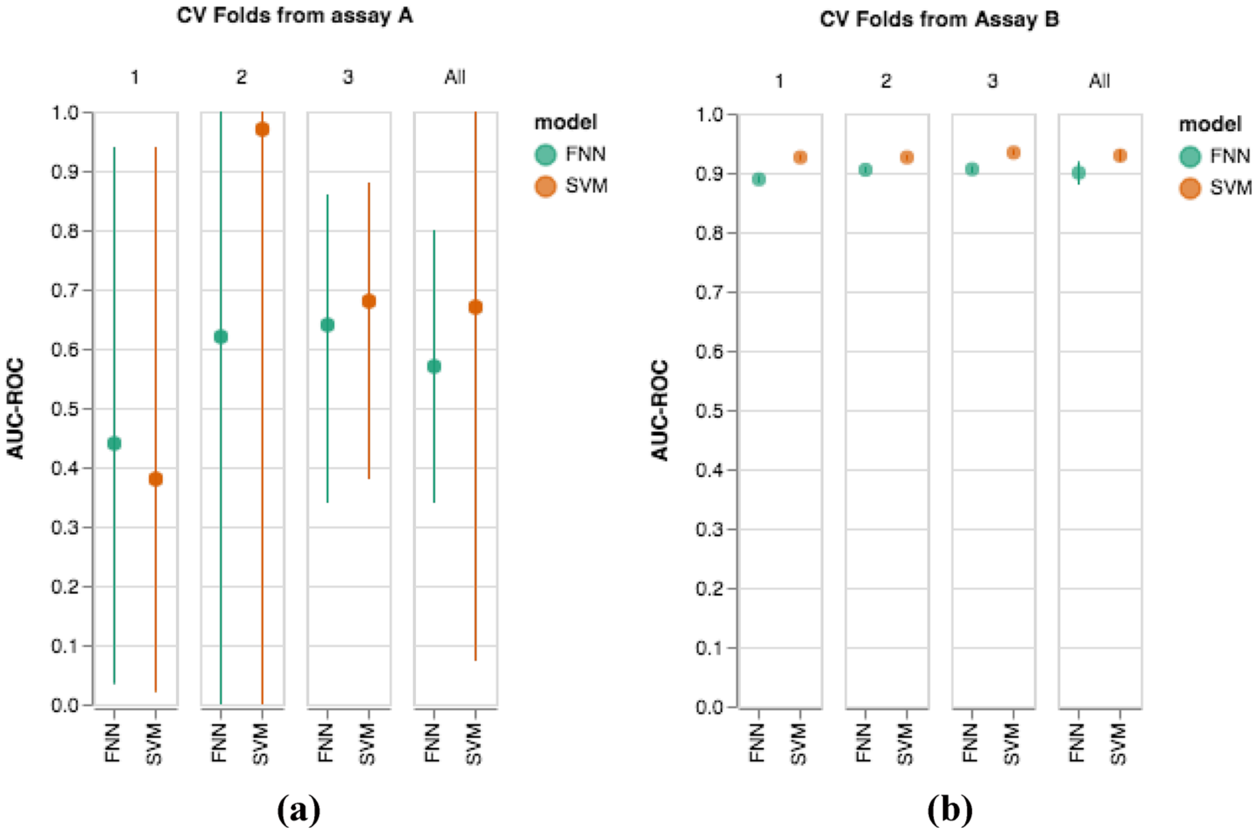
Contribution:

They experimented with two datasets, one with small sample size and another with large sample size. For the large sample data, Support Vector Machine (SVM) and Feed Forward Neural Network (FNN) showed the two best results among all other models.

Table 2.1: Robinson’s experimental results using 3 Fold cross validation [18]

	Fold 1	Fold	Fold	Mean	SEM
A: ChEMBL 1964055					
FNN AUC–ROC (95% CI)	0.44 (0.035, 0.94)	0.62 (0.0, 1.0)	0.64 (0.34, 0.86)	0.57	0.05
SVM AUC–ROC (95% CI)	0.38 (0.02, 0.94)	0.97 (0.0, 1.0)	0.68 (0.38, 0.88)	0.67	0.14
Test set size (actives/ inactives)	35 (32/3)	30 (29/1)	35 (29/6)		
B: ChEMBL 1794580					
FNN AUC–ROC (95% CI)	0.889 (0.883, 0.895)	0.905 (0.900, 0.910)	0.906 (0.900, 0.911)	0.900	0.005

SVM AUC–ROC (95% CI)	0.936 (0.921, 0.931)	0.926 (0.921, 0.930)	0.934 (0.930, 0.939)	0.929	0.002
Test set size (actives/ inactives)	19388 (5553/13885)	25165 (6918/18247)	19363 (5491/13872)		



(a)

(b)

Figure 2.1: Information contained in table 2.1 in graphical format [18].

Limitations:

There are some limitations of their result:

- Main issue is that their dataset is not uniform. They did not separate and choose only one standard from IC50, EC50, potency etc. standards in a same dataset.
- How they classified actives and inactives is not mentioned clearly.

In another paper, titled "Large-scale comparison of machine learning methods for drug target prediction on ChEMBL", by A. Mayr and co, published in 2018, in Chemical Science journal, the researchers concluded that deep learning method performed best among the other models.

Contribution:

The experiment was conducted on large scale dataset, probably the largest till date. Deep learning model's accuracy on various assays is given in the following table:

Table 2.2: Mayr's experimental results (partial) using 2 Fold cross validation [19]

Assay	Surrgate Assay	Target	Surrogate Assay Accuracy	Deep learning accuracy
CHEMBL1909134	CHEMBL1613777	CYP450-2C19	0.54 [0.4136, 0.653]	0.95 [0.9198, 0.9658]
CHEMBL1909200	CHEMBL1614521	ERK	0.56 [0.4012, 0.7005]	0.98 [0.9615, 0.9912]
CHEMBL1909136	CHEMBL1614110	CYP450-2D6	0.51 [0.3923, 0.6197]	0.91 [0.8734, 0.9319]

CHEMBL1909135	CHEMBL1614027	CYP450-2C9	0.68 [0.5567, 0.7853]	0.95 [0.9278, 0.9713]
CHEMBL1909138	CHEMBL1614108	CYP450-3A4	0.86 [0.8041, 0.9071]	0.74 [0.657, 0.8105]
CHEMBL1963940	CHEMBL1794352	Luciferase	1.00 [0.8076, 1]	0.87 [0.775, 0.9344]
CHEMBL1741321	CHEMBL1614110	CYP450-2D6	0.99 [0.9889, 0.9956]	0.83 [0.8184, 0.8352]
CHEMBL1741325	CHEMBL1614027	CYP450-2C9	0.99 [0.9839, 0.993]	0.75 [0.7428, 0.762]
CHEMBL1741323	CHEMBL1614027	CYP450-2C19	0.99 [0.9822, 0.9911]	0.77 [0.7602, 0.7789]

Limitations:

Although, the experiment was a large scale experiment on various types of biochemical assays, it has some limitations. Some of the limitations are:

- Data was not preprocessed properly. That’s why for some dataset deep learning model gave 97% accuracy but the same model gave only 75% accuracy for different dataset.
- Imbalance dataset.

Other papers that we have studied also had limitations. Main limitation for most of the researches is regarding the dataset like missing data, duplicate data, imbalanced data etc. To overcome these limitations we have implemented machine learning models in our processed dataset which will be discussed throughout the rest of the paper.

2.3 Conclusion

In this chapter, we discussed about the literature that we reviewed before conducting our research. Different machine learning and deep learning models were implemented on different datasets in the previous researches. We discussed some of the contribution and limitation regarding other researches in the field of drug discovery.

CHAPTER 3

Background Study

3.1 Introduction

In the previous chapter we discussed about our reviewed literatures. In this chapter, we will further elaborate on the background study that we have conducted for our research. The background study will include study regarding the target proteins of the two datasets that we have used, one is Plasmodium Falciparum and another is Leukocyte Elastase. We will, also discuss about the different machine learning models and Feed Forward Neural Network (FNN) deep learning model.

3.2 Plasmodium Falciparum

Plasmodium falciparum, one species of parasite that causes malaria in humans, is transmitted through the bites of female *Anopheles* mosquitoes. Once a pregnant mother is found to be infected, chloroquine and sulfadoxine-pyrimethamine can be used to help prevent maternal–fetal transmission. Plasmodium falciparum infects pregnant women more often and more severely than nonpregnant women, partially due to pregnancy-induced cell-mediated immunity depression. The severity of the infection is affected by many factors, including the number of previous pregnancies, previous malarial infections and immunity, trimester of pregnancy, and comorbidity [20].

It is responsible for around 50% of all malaria cases. P. falciparum is therefore regarded as the deadliest parasite in humans. It is also associated with the development of blood cancer and is classified as a Group 2A carcinogen.

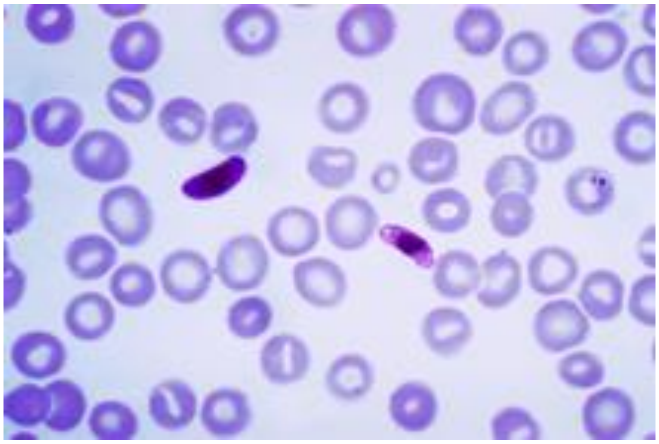


Figure 3.1: Plasmodium Falciparum [21]

3.3 Leukocyte Elastase

Human leukocyte elastase (HLE) is a serine protease found in the azurophilic granules of the neutrophil. It is also known as human neutrophil elastase, and has been assigned a unique number by the Enzyme Commission of the International Union of Biochemistry, based on its activity. Its potential substrates include almost all components of the extracellular matrix, as well as proteins as diverse as clotting factors, complement, immunoglobulins, and cytokines. Interest in HLE was engendered in part by the observation that in an inherited disease, α1-antitrypsin deficiency, unopposed action of HLE because of lack of a protease inhibitor predisposed to a premature and sometimes severe form of emphysema.

The requirement for neutrophils to migrate out of the vasculature and through the basement membrane, as well as the potent proteolytic repertoire of HLE have led to the supposition that HLE might be involved in the pathogenesis of inflammatory tissue injury such as occurs in acute lung injury and ARDS. Nonetheless, the role of HLE in acute lung injury remains far from clear. Indeed, HLE is only one of a myriad of proteases synthesized by leukocytes and other cells in the lung, and some of these (e.g., matrix metalloproteinases from macrophages) have also been implicated in acute lung injury [22].

3.4 Machine Learning Models

Different machine learning and deep learning models were considered before conducting our research. Here the theoretical studies about the models we used is discussed.

Random Forest (RF) Classification Model:

Random Forest (RF) classifier is a classification model that is made of a number of decision trees on various subsets of the given dataset. Using the decision trees Random Forest (RF) classifier and finds out the average of the result and uses it to improve the accuracy of the dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. Generally, the greater number of trees in the Random Forest model, the higher the accuracy of the results. Multiple decision trees also prevents the problem of overfitting.

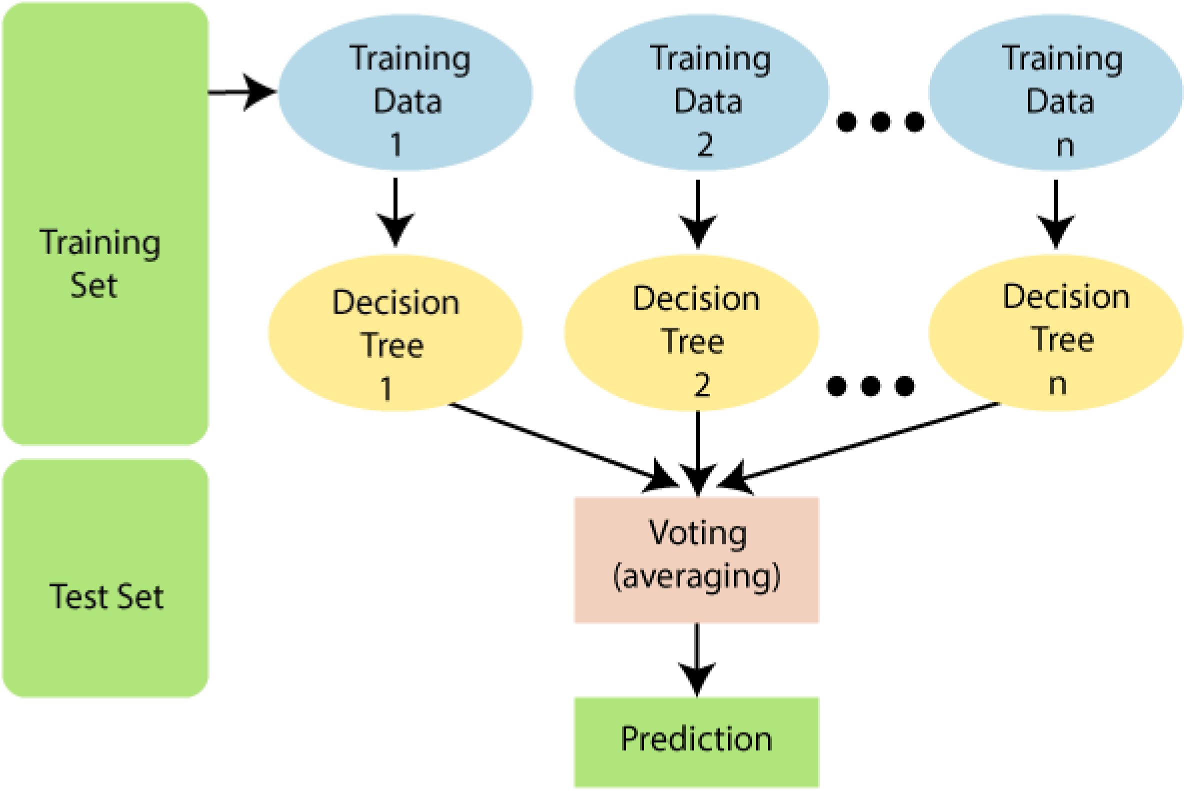


Figure 3.2: Random Forest Classification model [23]

Support Vector Machine (SVM) Classification Model:

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

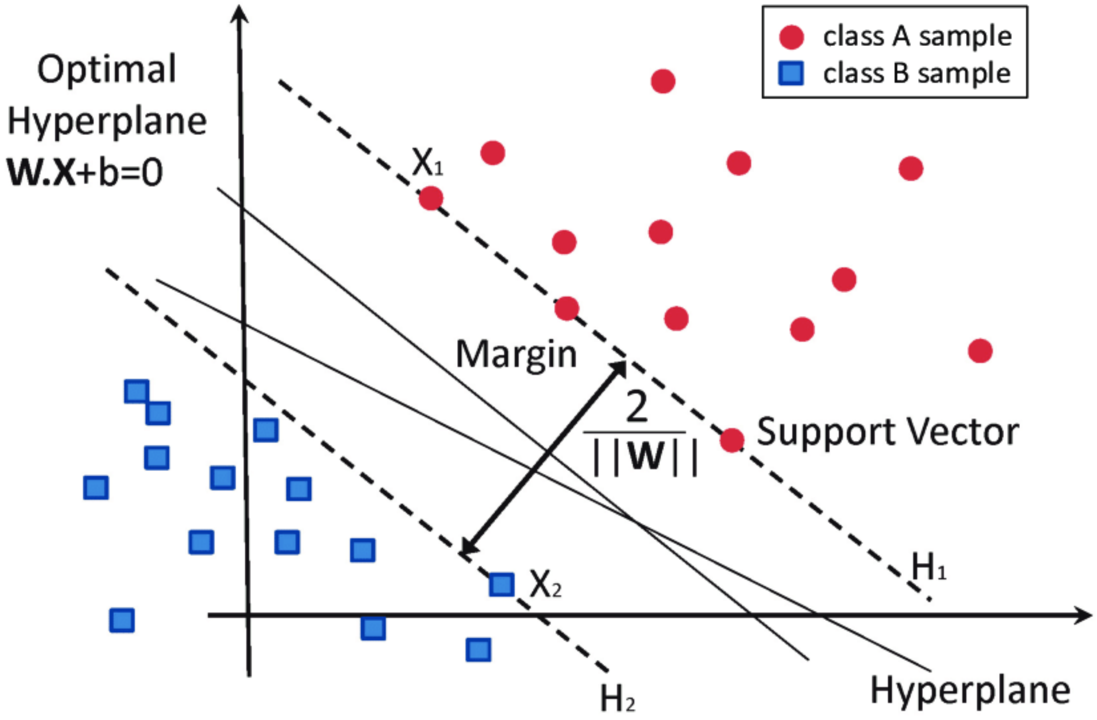


Figure 3.3: Support Vector Machine (SVM) classification model [24]

Feed Forward Neural Network (FNN) Model:

This Neural Network or Artificial Neural Network has multiple hidden layers that make it a multilayer neural Network and it is feed-forward because it is a network that follows a top-down approach to train the network. In this network there are the following layers:

1. **Input Layer:** It is starting layer of the network that has a weight associated with the signals.
2. **Hidden Layer:** This layer lies after the input layer and contains multiple neurons that perform all computations and pass the result to the output unit.

3. **Output Layer:** It is a layer that contains output units or neurons and receives processed data from the hidden layer, if there are further hidden layers connected to it then it passes the weighted unit to the connected hidden layer for further processing to get the desired result.

The input and hidden layers use sigmoid and linear activation functions whereas the output layer uses a Heaviside step activation function at nodes because it is a two-step activation function that helps in predicting results as per requirements. All units also known as neurons have weights and calculation at the hidden layer is the summation of the dot product of all weights and their signals and finally the sigmoid function of the calculated sum. Multiple hidden and output layer increases the accuracy of the output [25].

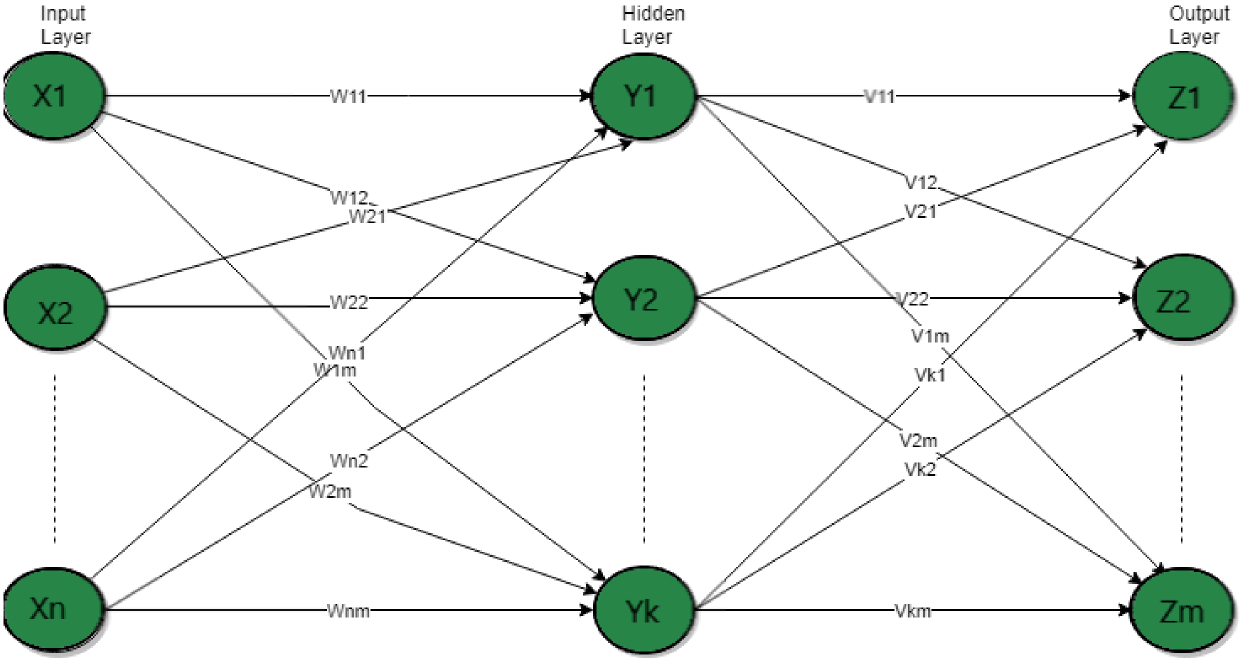


Figure 3.4: Feed Forward Neural Network (FNN) model [25]

3.5 Conclusion

In this chapter we discussed about Plasmodium Falciparum and Leukocyte Elastase. We also discussed the theoretical study about machine learning models. In the next chapter, we will discuss about the methodology of our work.

CHAPTER 4

Methodology

4.1 Introduction

Lots of factors need to be considered before selecting models for a particular research work [26]. We have primarily worked with two machine learning models Random Forest (RF) and Support Vector Machine (SVM) and we have proposed a Feed Forward Neural Network model for our research. Our research was conducted on two datasets collected from the ChEMBL [27] database. One was on Plasmodium Falciparum and the other was on Leukocyte Elastase dataset. The reasoning behind the selected models and the details of the models will be briefly discussed on this chapter.

4.2 Classification Fundamentals

A problem can be categorized in two ways. One is a classification problem and another is a regression problem. Classification is concerned with predicting a label, whereas, regression is concerned with predicting a quantity. Thus, classification works with discrete values and regression works with continuous values.

Classification is the method which identifies the target class for each data object or observation of a dataset. Classification problems can be further divided into two types based on class labels. One is a binary classification problems and another is multiclass classification problems. In binary classification problems the number of unique class labels is two. And in multiclass classification problems the number of unique class labels is more than two. Classification problems can be further divided into two types based on the data, supervised classification problem and unsupervised classification problem.

Supervised classification problem deals with labeled input and output data, whereas unsupervised classification problem deals with unlabeled input and output data. Using the information about the known class labels of some of the data, supervised classification problem finds out the class labels of one or more unknown data objects. On the other hand, unsupervised classification does not have this advantage.

4.3 Dataset Type

Our work is based on the dataset of two classes, active and inactive. The class labels of the samples of the dataset were also identified before the implementation. Hence, the classification problem of our research work is a supervised binary classification problem.

The data type of our research is tabular data. So this was taken into consideration before model selection. Because different data types require different machine learning and deep learning models to produce the best solutions [28, 29]. For example, a model which may give high accuracy for image data, may not give the similar kind of result for a tabular data and vice versa.

4.4 Model Selection

There are various machine learning and deep learning models available for research work in different. For research in specific fields, specific machine learning and deep learning models may work better than other models. And for specific research in specific fields may require specific models to produce the best results. To select models which are likely to give optimal results for a particular research is not an easy task. Various criteria need to be taken into consideration before model selection [28, 29].

For our research, we also tried our best to identify the criteria and to select model according. Our research is on bioactivity prediction from target proteins. Bioactivity prediction is of a binary classification problem. The goal of our research is to find active and inactive classes for the target.

Machine learning models have a solid history in bioactivity prediction producing very good results [30]. Among the machine learning models Support Vector Machine (SVM) produced the best result in Robinson’s research [18]. Based, on that result we have selected Support Vector Machine (SVM) as one of the classification models for our research.

Random Forest (RF) classifier also tends to do very well in classification models. It has shown steady performance among various results conducted in drug discovery over the years [30]. So we have selected Random Forest (RF) as one of the classification models for our research.

Among the deep learning models, various models have already been implemented in drug discovery [18, 19]. These includes, Feed Forward Neural Network (FNN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Graph Neural Network etc.

Convolutional Neural Network (CNN) is one of the most powerful and also popular models that have been implemented in various research areas today. This neural network computational model uses a variation of multilayer perceptrons and contains one or more convolutional layers that can be either entirely connected or pooled. These convolutional layers create feature maps that record a region of image which is ultimately broken into rectangles and sent out for nonlinear processing [31]. Convolutional Neural Network (CNN) is a very powerful model for image classification. But it lags behind when classification one dimensional tabular data like that which we are using for our research. As the data needs to be reshaped for implanting CNN, research will become much complex. And the result of CNN was not the best in previous researches conducted in the field of drug discovery [18, 19].

Recurrent Neural Network (RNN) works best for sequence prediction problems. RNN models creates a loop by saving the output of processing nodes and feeding the result back into the model. This is the way the RNN model learns to predict the outcome of a layer. Each node in the RNN model acts as a memory cell, continuing the computation and implementation of operations. If the network's prediction is incorrect, then the system self-learns and continues working towards the correct prediction during backpropagation. Although, RNN has many advantages, it is not best for tabular data classification problems [32]. LSTM suffers from the same problems.

Feed Forward Neural Network (FNN) is the simplest form of neural network architecture yet a very powerful one. It does not have the complexity of Recurrent Neural Network (RNN) as it does not use backpropagation unlike RNN. Backpropagation does not play a major role in drug discovery. Plus Feed Forward Neural Network (FNN) works very well with tabular data. Historically, this models have also produced best results in drug discovery as shown in Robinson and his team's research [18]. So this is why, we have selected Feed Forward Neural Network (FNN) as one of the models for our research.

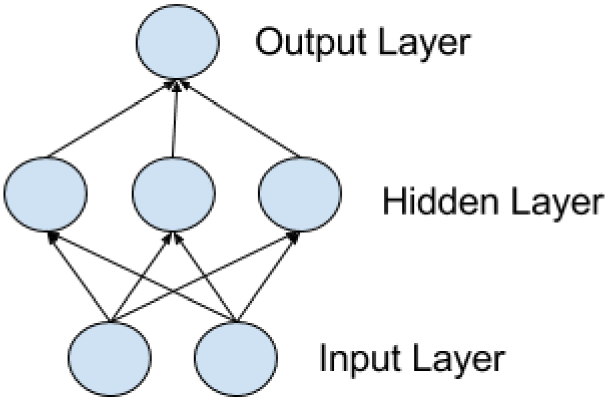
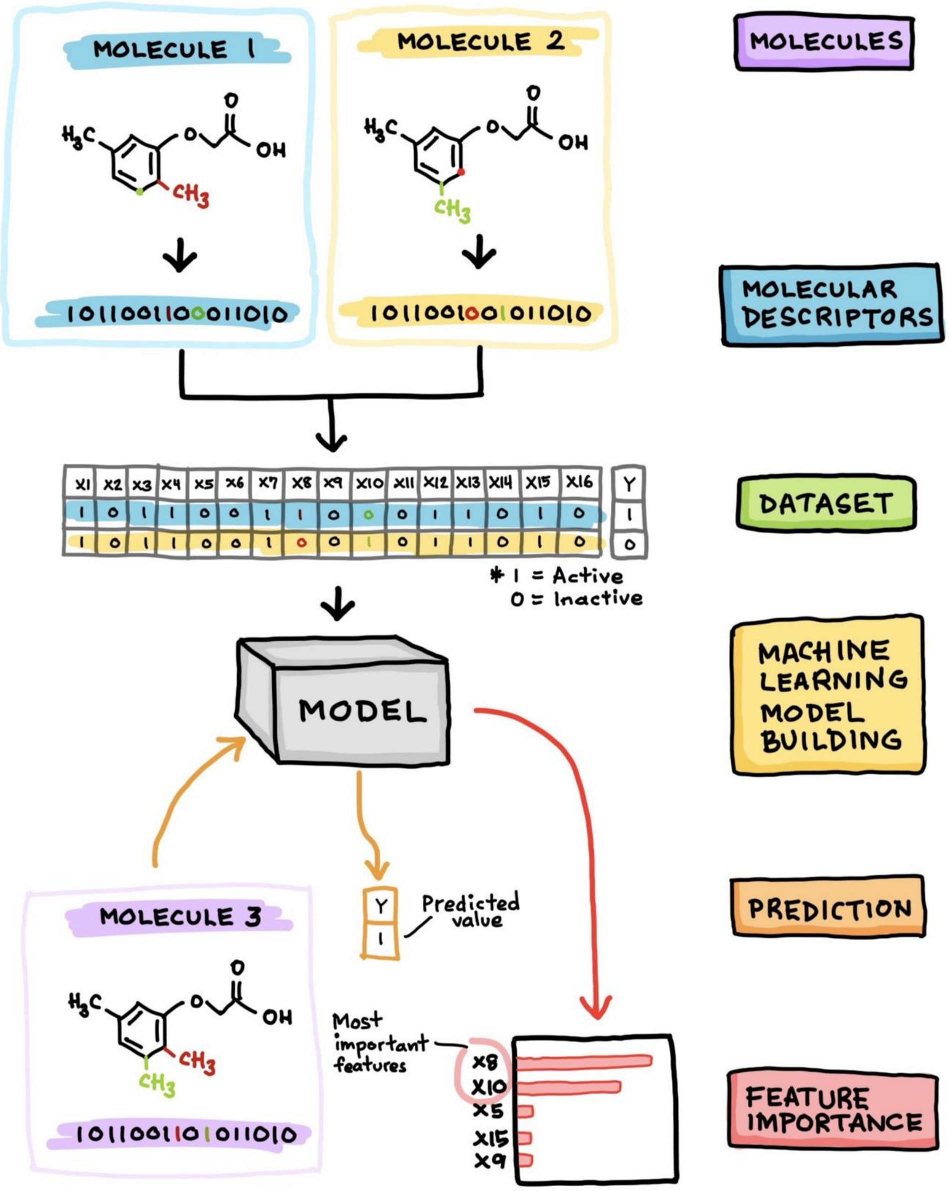


Figure 4.1: Simple Multi-Layer perceptron model [32]

4.5 Computational Drug Discovery Overview

Computational drug discovery has clear advantages over manual drug discovery methods as it accelerates drug discovery and development process and also it is more economical to implement. That is why we chose computational drug discovery using machine learning models for our research purpose.



X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	Y
1	0	1	1	0	0	1	1	0	0	0	1	1	0	1	0	1
1	0	1	1	0	0	1	0	0	1	0	1	1	0	1	0	0

Figure 4.2: Computational drug discovery overview [33]

Computational drug discovery is based on Quantitative structure-activity relationship (QSAR) technique which derives the structural relationship between chemical compounds and their biological activities.

In a drug discovery dataset, multiple molecules with their Simplified Molecular Input Line Entry System (SMILES) notation are present. SMILES notation indicate the simplified structure of chemical compound with the use of line notations.

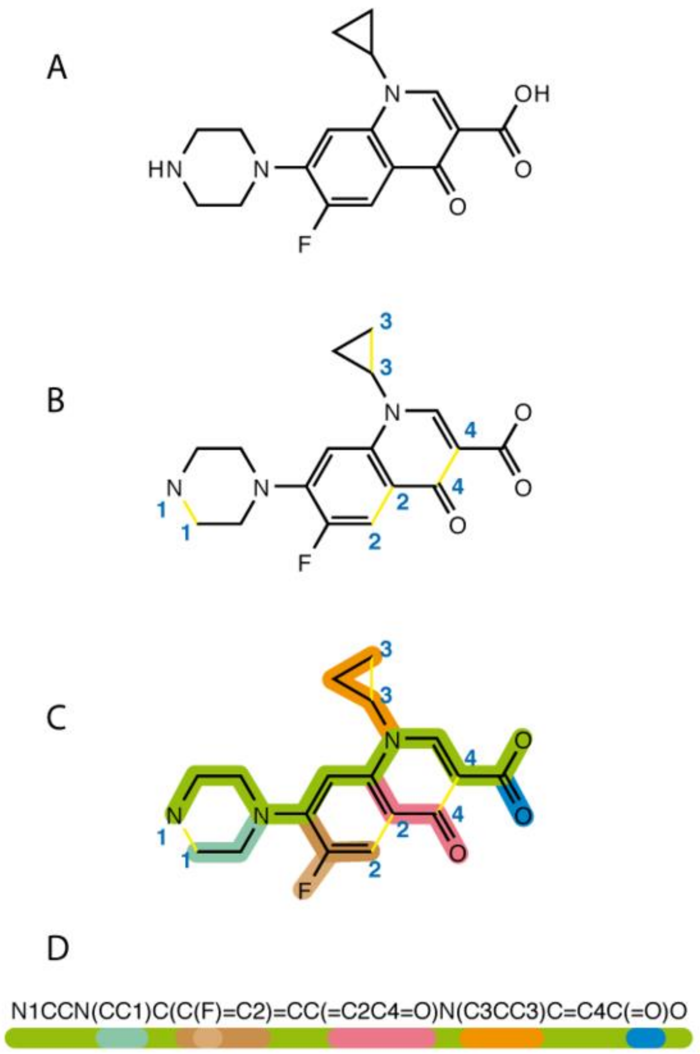


Figure 4.3: SMILES notation of a chemical compound [34]

Molecular descriptors from the molecules can be found out using proper tools. Molecular descriptors are the binary representation of the chemical compounds. ‘1’ and ‘0’ in the molecular descriptor indicates if certain feature is present in the molecule or not.

Bioactivity of a molecule can be found out by analyzing the standard value present in the dataset. Bioactivity is also represented as binary values ’1’ or ‘0’. And along with the molecular descriptors which would be inputs, the bioactivity of the molecules which would be the outputs would be fed into the machine learning and deep learning models for testing.

The trained machine learning and deep learning models can then give prediction on new dataset based on the learning of the previous data. Besides giving prediction, the models can also be used to found out important features which can be used for further research for drug discovery in the future.

4.6 Selected Models

As discussed in the previous chapter, we have selected Random Forest (RF), Support Vector Machine (SVM) and Feed Forward Neural Network (FNN) model for our research. Feed Forward Neural Network (FNN) model have the following characteristics:

- It has 3 hidden layers.
- Hidden layer 1: 116 neurons
- Hidden layer 2: 40 neurons
- Hidden layer 3: 8 neurons
- Output layer: 1 neuron

We have selected 3 hidden layers because for large data with many features more than two hidden layers perform better. There is no optimal way to choose the number of neurons in the hidden layers. Generally the numbers is between the size of input layer and output layer. We have taken this into consideration while selecting the number of neurons for the hidden layers.

4.7 Conclusion

In this chapter, we have discussed about the criteria that we evaluated before model selection for our research. We gave an overview of drug discovery technique. We then discussed about the selected models for our research. Random Forest, Support Vector Machine and feed Forward Neural Network were selected for our research which were discussed in this chapter. In the next chapter, we will discuss about the implementation of the models.

Chapter 5
Implementation

5.1 Introduction

In the previous chapter our proposed models for the research were discussed. The models were Random Forest (RF) classifier, Support Vector Machine (SVM) classifier and Feed Forward Neural Network (FNN) model. In this chapter, we will discuss the implementation of our proposed models.

5.2 Experimental Environment and Tools

We have conducted our experiment in the machine of following configuration:

Table 5.1: System configuration of machine

Processor	Intel Core i7-7500U
CPU	~ 32.5 GHz
RAM	8 GB
System Type	64 bit OS
Hard Disk	2 TB
Operating System	Windows 10, version 21H2

Other important features for our experiment regarding environment and tools:
The Programming language that we used: Python.
Platform: Jupyter Notebook and Google Colab.
We preferred Google Colab because it has some beneficial features which was useful for our experiment. Some of its useful features are:

- Cloud based.
- No extra set up is needed.
- Importing libraries without installing them.

- ➢ Can save the files directly to github.
- ➢ Can save the project as pdf.
- ➢ Provides the use of free GPU.

Thus, we conducted our experiment in Google Colab.

5.3 Database Description

We gathered our data from ChEMBL database [27]. The ChEMBL database is a database that contains curated bioactivity data of more than 2.3 million compounds. It is compiled from more than 85,000 documents, 1.5 million assays and the data spans 15,000 targets and 2,000 cells and 43,000 indications. [Data as of September 29, 2022; ChEMBL version 31].

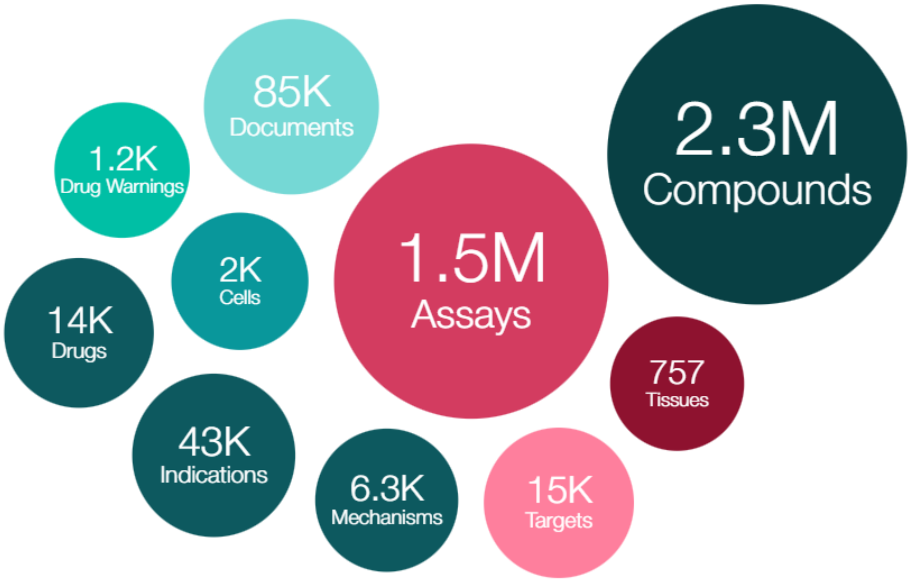


Figure 5.1: Summary of ChEMBL entities and quantities of data for each item in the ChEMBL database [27]

One beneficial feature of ChEMBL database [27] is that ChEMBL web service library is available and it can be imported to the notebook using python.

5.4 Dataset Description

We have used two different datasets for our experiment:

1. Plasmodium Falciparum dataset
2. Leukocyte Elastase dataset.

After installing the ChEMBL web service library in our notebook, we have searched for our datasets in the target section of search option using keywords. ‘Plasmodium Falciparum’ search showed the following results:

Table 5.2: Plasmodium Falciparum targets (Partial table)

Id	Organism	Score	Species group flag	Target CHEMBL id
0	Plasmodium Falciparum	28.0	False	CHEMBL 364
1	Plasmodium Falciparum 3D7	24.0	False	CHEMBL 2366922
2	Plasmodium Falciparum D6	24.0	False	CHEMBL 2367107
3	Plasmodium Falciparum NF54	24.0	False	CHEMBL 2367131
4	Plasmodium Falciparum FcB1/Columbia	19.0	False	CHEMBL 612608
…	…	…	…	…
75	Plasmodium Falciparum	8.0	False	CHEMBL 2169724

As our research is based on bioactivity prediction from target proteins we only searched in the target section. And to maintain the uniformity data we only chose data which were categorized based on IC50 unit. IC50 unit shows the concentration of a drug to inhibit a particular biological process by 50%.

After the searches we selected the target ChEMBL id of ‘364’ for Plasmodium Falciparum dataset and ChEMBL id of ‘248’ for Leukocyte Elastase dataset.

Plasmodium Falciparum dataset:

Plasmodium Falciparum dataset originally contained 43324 targets.

A visual representation of few of the rows and columns of Plasmodium Falciparum dataset is shown in the following table:

Table 5.3: Plasmodium Falciparum (ChEMBL id: 364) dataset

activity_id	assay_chembl_id	assay_type
32325	CHEMBL764090	F
32480	CHEMBL760652	F
33480	CHEMBL764090	F
34739	CHEMBL764090	F
34877	CHEMBL760652	F
34878	CHEMBL760652	F
35971	CHEMBL764090	F
35972	CHEMBL764090	F
37170	CHEMBL764090	F
37171	CHEMBL764090	F
37172	CHEMBL764090	F
37328	CHEMBL760652	F
37329	CHEMBL760653	F
38263	CHEMBL762998	F
38264	CHEMBL762999	F
38744	CHEMBL763697	F

38745	CHEMBL763696	F
38747	CHEMBL763697	F
38748	CHEMBL763696	F

Leukocyte Elastase dataset:

Leukocyte Elastase dataset originally contained 3147 targets.

A visual representation of the first few rows and columns of Leukocyte Elastase dataset is shown in the following table:

Table 5.4: Leukocyte Elastase (ChEMBL id: 248) dataset

activity_id	assay_chembl_id	bao_label
32937	CHEMBL675350	single protein format
35188	CHEMBL676202	single protein format
35192	CHEMBL676203	tissue-based format
35193	CHEMBL752072	tissue-based format
35431	CHEMBL675350	single protein format
46161	CHEMBL675350	single protein format
46174	CHEMBL675350	single protein format
50141	CHEMBL675350	single protein format
51352	CHEMBL675350	single protein format
54831	CHEMBL676202	single protein format
54835	CHEMBL878396	tissue-based format
54836	CHEMBL752072	tissue-based format
54864	CHEMBL676202	single protein format
54868	CHEMBL676203	tissue-based format
54869	CHEMBL752072	tissue-based format

59807	CHEMBL676202	single protein format
59811	CHEMBL878396	tissue-based format
59812	CHEMBL752072	tissue-based format
60174	CHEMBL675350	single protein format
66748	CHEMBL675350	single protein format
72782	CHEMBL676197	single protein format

5.5 Data Preprocessing

The raw data collected from a source is often not in the state such that machine learning models or deep learning models can be implemented on them. So some sort of data processing is needed on the raw data. Data preprocessing refers to some sort of operation perform on the raw data to produce another form of data before the actual data preprocessing takes place. Data preprocessing coverts the raw data into a clean state. The transformed data after the data preprocessing is more useful and efficient to use.

The raw data for both Plasmodium Falciparum dataset and Leukocyte Elastase dataset needed data preprocessing before the actual implementation of the models for our research. Firstly, we handled the missing data from the dataset. If any compounds had missing value for the standard value and canonical smiles column then the rows were dropped. We did it because the standard value and canonical smiles column give important information for the dataset which we needed for further operation in future. Then, duplicate canonical smiles data were dropped to create a more balanced dataset.

After handling the missing data and removing the duplicated canonical smiles, size of the both of datasets were significantly reduced. The Plasmodium Falciparum dataset was reduced from 43324 entries to 20664 entries and Leukocyte Elastase dataset was reduced from 3147 entries to 2030 entries.

Then, we combined the values of the three following columns: molecule_chembl_id, canonical_ smiles and standard value into a single data frame for both of the datasets. By this way, unnecessary columns were removed and dataset became much cleaner and easier to understand.

Next, we did further operation to find the active and inactive class of the datasets. The molecules needed to be classified into two classes ‘active’ and ‘inactive’ for our research. The value we chose to divide the classes is the standard value column of our dataset. The bioactivity data is in the IC50 unit. Compounds having values of less than or equal 1000 nM were considered to be active while those greater than 1000 nM were considered to be inactive. It is a standard way of classifying active and inactive compounds as, the lower the IC50 value, the more potent the drug is. Then the Standard IC50 value was changed into smaller pIC50 unit for the ease of operation. The pIC50 value is the negative log of the IC50 value when converted to molar.

After all these operations the datasets looked like the following:

Table 5.5: First few entries of the preprocessed dataset of Plasmodium Falciparum

Id	molecule_chembl_id	canonical_smiles	class	pIC50
0	CHEMBL77052	C[C@@H]1CC[C@H]2[C@@H](C)[C…	active	8.37161
1	CHEMBL307145	Oc1cccc(O)c1O	inactive	5.24718
2	CHEMBL16300	O=C(NO)c1ccccc1	inactive	4.75448
3	CHEMBL307153	C[C@@H]1CC[C@H]2[C@@H]©….	active	8.01999
4	CHEMBL339049	CC(C)(C)NCc1cc(Nc2ccnc3cc(Cl)….	active	7.74472
5	CHEMBL316098	CC(C)(C)NCc1cc(Nc2ccnc3cc(Cl)….	active	8.83268
6	CHEMBL93286	CC(C)(C)NCc1cc(Nc2ccnc3cc(….	active	8.16749
7	CHEMBL305899	C[C@@H]1CC[C@H]2[C@@H]©…	active	8.11520
8	CHEMBL337981	CC(C)(C)NCc1cc(Nc2ccnc3cc(Cl)..	active	7.23657
9	CHEMBL252518	C[C@@H]1CC[C@H]2[C@@H]…	active	8.95860
10	CHEMBL306763	C[C@@H]1CC[C@H]2[C@@H..	active	8.43415
11	CHEMBL327539	CCc1cc(Nc2ccnc3cc(Cl)ccc23)..	active	8.52724
12	CHEMBL287556	O=c1cc(CO)occ1O	inactive	2.80134

Table 5.6: First few entries of the preprocessed dataset of Leukocyte Elastase

Id	molecule_chembl_id	canonical_smiles	class	pIC50
0	CHEMBL97927	O=C1C[C@H](Cc2ccccc2)…	inactive	4.124939
1	CHEMBL543416	CC(C)[C@@H]1C(=O)N(S..	active	7.958607
2	CHEMBL95260	CS[C@@H]1C(=O)N(C(=O)…	active	6.69897
3	CHEMBL95813	C=CC[C@@H]1C(=O)N(C….	active	6.522879
4	CHEMBL330647	C[C@@H]1C(=O)N(C….	inactive	4.60206
5	CHEMBL318052	C[C@H]1C(=O)N(C….	inactive	4.443697
7	CHEMBL2368616	CC(C)[C@@H]1C(=O)….	active	7.657577
8	CHEMBL554013	CC(C)[C@@H]1C(=O….	active	7.431798
9	CHEMBL542949	CC(C)[C@@H]1C(=O)….	active	8
11	CHEMBL318482	CO[C@@H]1C(=O)N(….	active	6.045757
12	CHEMBL311340	CCCO[C@H]1C(=O)N2C..	inactive	4.39794

For both table 5.3 and table 5.4, full canonical_smiles could not be shown because of space shortage as the notations are huge.

5.6 Feature Extraction

Feature extraction is an important part of a bioactivity prediction result. Features are directly fed into machine learning or deep learning models as inputs. Different methods are available for feature extraction from molecules [35]. For molecules, molecular features are known as molecular descriptors.

We chose the PaDEL descriptor for feature extraction from the molecules of our dataset. The

PaDEL descriptor is a software for calculating molecular descriptors and fingerprints. The software currently calculates 797 descriptors (663 1D, 2D descriptors, and 134 3D descriptors) and 10 types of fingerprints. These descriptors and fingerprints are calculated mainly using The Chemistry Development Kit [36].

Advantages of the PaDEL descriptor were taken into consideration for selecting it in our research. Some of the advantages of the PaDEL descriptor is:

- It is open source, therefore, free and easily accessible.
- Provides GUI.
- Provides command line interfaces.
- Supports more than 90 different molecular file formats.
- Can work on any platforms.
- Very fast calculation speed.
- Reliable.

From the PaDEL descriptor we have used the PubChem fingerprint as the molecular descriptor for our research. The PubChem fingerprint is a type of database fingerprint (DFB) which encodes molecular fragments information with 881 binary digits [37]. The 881 binary digits indicates absence or presence of certain features in the molecules. The features are useful for drug discovery. For our research, we did not use the low variance features. The number of features were reduced to 174 by setting the variance threshold value to 0.16. So the threshold values equal to or lower than 0.16 were removed. Threshold value of 0.16 was selected as this value gave the best outcome after few experiments with different values.

Some important features of PubChem fingerprints were considered to choose it as molecular descriptor. Some of the features are:

- Useful for similarity neighboring.
- Useful for similarity searching.
- Binary representation is simple.
- Able to handle different elements.
- Gives very good results.

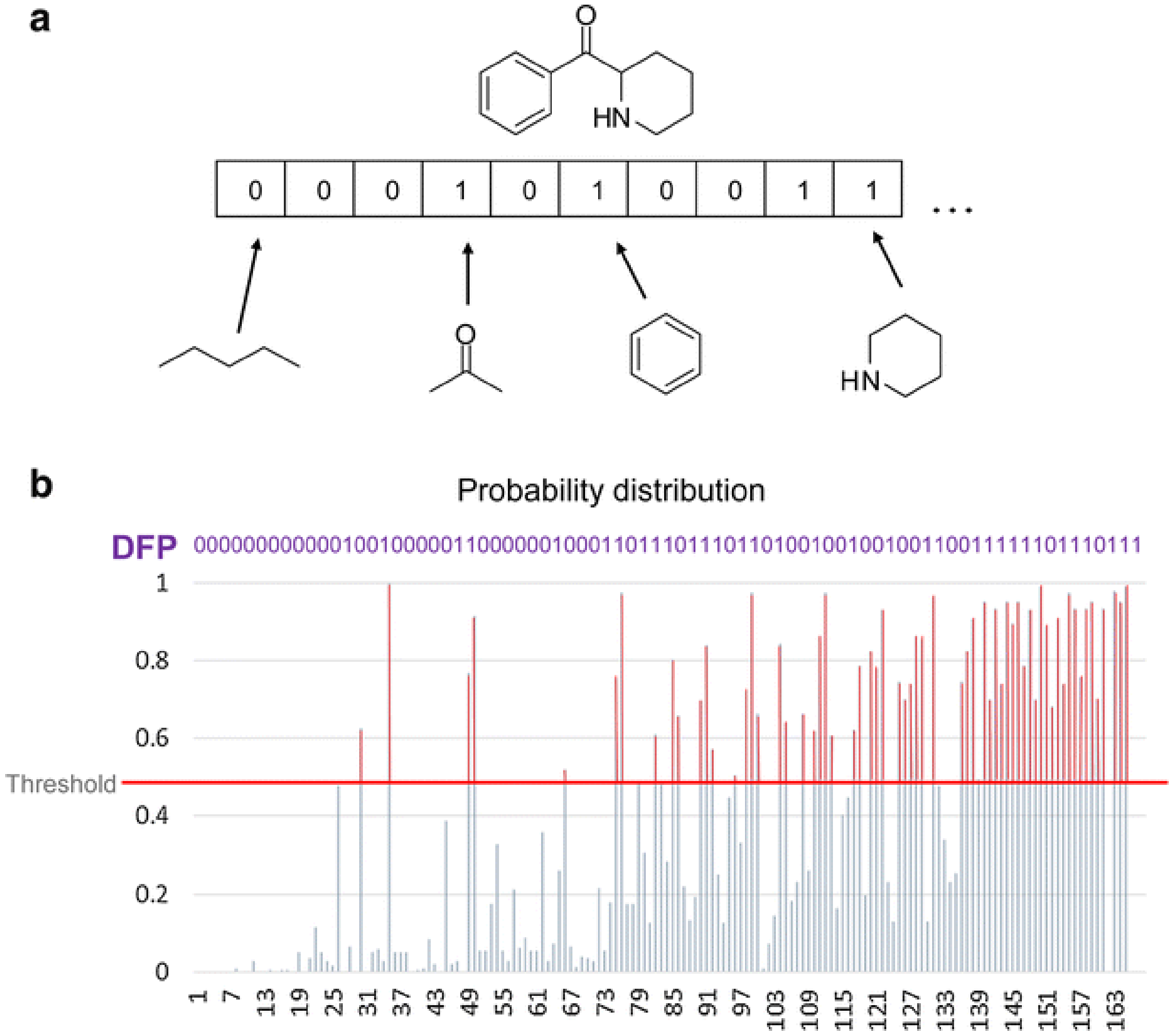


Figure 5.2: (a) Schematic representation of a binary and dictionary-based molecular fingerprint. (b) Schematic representation of a database fingerprint (DFP) [37].

We generated molecular descriptor data for both Plasmodium Falciparum dataset and Leukocyte Elastase datasets. An example of the generated dataset is given in the form of a table below:

Table 5.7: Sample of Molecular descriptor data for Plasmodium Falciparum

Chembl_id	Feature_1	Feature_2	Feature_3	Feature_4	Feature_5
CHEMBL4634678	1	1	1	0	0
CHEMBL4649215	1	1	1	0	0
CHEMBL4639076	1	1	1	0	0
CHEMBL4642967	1	1	1	0	0
CHEMBL4646244	1	1	1	0	0
CHEMBL4649077	1	1	1	0	0
CHEMBL4649433	1	1	1	0	0
CHEMBL4647889	1	1	1	0	0
CHEMBL4644427	1	1	1	0	0
CHEMBL4639851	1	1	1	0	0
CHEMBL4648455	1	1	1	1	0
CHEMBL4641754	1	1	1	1	0
CHEMBL4641338	1	1	1	0	0
CHEMBL4638969	1	1	1	0	0
CHEMBL4640761	1	1	1	0	0
CHEMBL4634976	1	1	1	0	0
CHEMBL4648980	1	1	1	0	0
CHEMBL4634698	1	1	1	0	0
CHEMBL4647784	1	1	1	0	0
CHEMBL4648000	1	1	1	0	0
CHEMBL4642389	1	0	0	0	0
CHEMBL4639642	1	0	0	0	0

5.7 Random Forest (RF) Classifier Implementation

We used built in python libraries for implementing Random Forest classifier (RF) model. Implementation procedure is as follows:

- Split the data into 80/20 ratio for training and testing purpose.
- Set the n estimator to 100.
- Fed the data into the classification model.
- Found out the output result in the form of accuracy.
- Generate the confusion matrix.
- Printed the classification report for RF model.

5.8 Support Vector Machine (SVM) Classifier Implementation

Similarly, we used built in python libraries for implementing Support Vector Machine (SVM) classifier model. Implementation procedure is as follows:

- Split the data into 80/20 ratio for training and testing purpose.
- Set Kernel to linear.
- Set the value of C to 0.1.
- Set gamma to 1.
- Fed the data into the classification model.
- Found out the output result in the form of accuracy.
- Found out confusion matrix.
- Printed the classification report for SVM model.

5.9 More machine learning models implementation

Using Lazy Predict classifier [38] from python library we also implemented and compared different machine learning models with each other.

5.10 Feed Forward Neural Network (FNN) model Implementation

Implementation procedure of Feed Forward Neural Network (FNN) model is as follows:

- Split the data into 80/20 ratio for training and testing.

- Added 3 hidden layers.
- Set Hidden Layer 1: 116 neural nodes, Activation: ReLU.
- Set Hidden Layer 2: 40 neural nodes, Activation: ReLU.
- Set Hidden Layer 1: 8 neural nodes, Activation: ReLU.
- Set Output Layer: 2 neural nodes. Activation: Sigmoid
- Set Loss: binary cross entropy
- Set Optimizer: Adam
- Fed the inputs to the model.
- Set Batch size: 32
- Set Epoch: 200
- Found out the accuracy.
- Plotted epoch versus accuracy graph for training and testing data.

5.11 Conclusion

In this chapter, we have discussed about the environmental tools, the database, the dataset, data preprocessing, molecular descriptor, and implementation of the machine learning models. We discussed in details about ChEMBL database and the two datasets Leukocyte Elastase and Plasmodium Falciparum which were collected from the database. We also discussed about feature extraction method using PaDEL descriptor in the form of PubChem fingerprints. In the next chapter, experimental results and performance analysis will be discussed.

CHAPTER 6

Experimental Results and Performance Analysis

6.1 Introduction

In this section, we discuss the experimental results and performance of the models after implementing them. Different evaluation metrics like accuracy, precision, recall etc. have been used for performance evaluation of the models. Graphical representation of the ROC curve, histogram analysis of the model results have also been used for performance analysis and comparison of the models.

6.2 Evaluation Metrics

Comparing the actual value and the predicted value, the following four results can appear:

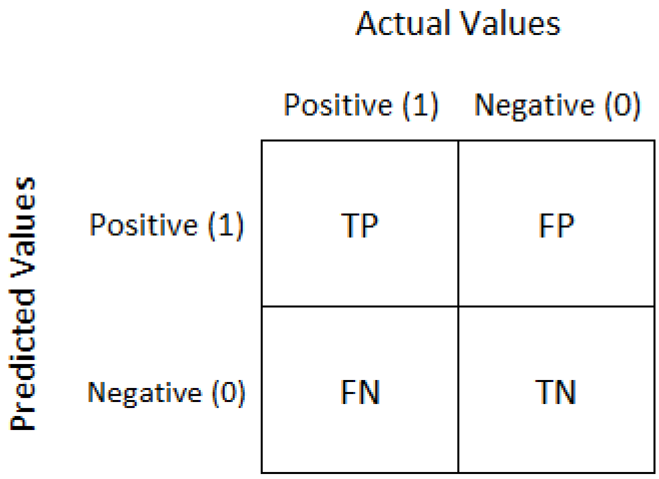
(1) True Positive (TP)

(2) True Negative (TN)

(3) False Positive (FP)

(4) False Negative (FN)

Considering these values a confusion metrics of the result can be formed.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 6.1: Confusion Matrix

Confusion matrix a graphical representation of the four values TP, TN, FP and FN. From these values different evaluation metrics like accuracy, precision, recall, specificity, F-score can be generated. These are briefly discussed below:

Accuracy:

The accuracy is calculated by dividing the correctly predicted items by total predicted items, both accurate and inaccurate. The formula for calculating accuracy is:

$$\textit{\textbf{Accuracy}} = (TP+TN) / (TP+FP+FN+TN) \quad (6.1)$$

Precision:

To calculate the value of precision, we divide the whole number of correctly predicted positive examples by the whole number of predicted positive examples. Higher precision intimates an instance specified as positive is certainly positive.

$$\textit{\textbf{Precision}} = TP / (TP+FP) \quad (6.2)$$

Recall:

Recall means out of the total positive classes, how much the model predicted as positive. The recall is interpreted as the rate of the whole number of accurately predicted positive samples divided by a total number of positive samples. High Recall value intimates the class is precisely identified. It should be as high as possible.

$$\textit{\textbf{Recall}} = TP / (TP+FN) \quad (6.3)$$

Specificity:

Specificity (SP) shows the ratio of accurately predicted negative items upon all negative data. The formula for calculating the specificity value is:

$$\textit{\textbf{Specificity}} = TN / (TN+FP) \quad (6.4)$$

F1 Score:

It is tough to differentiate two models with high precision and low recall or vice versa. So, in order to make them relative, we implement the F1-Score. F1-score aids to calculate Precision and Recall at the same time. It utilizes the Harmonic Mean in the position of Arithmetic Mean by correcting the absolute values.

$$\textit{\textbf{F1 Score}} = 2*(\textit{Recall} * \textit{Precision}) / (\textit{Recall} + \textit{Precision}) \qquad (6.5)$$

6.3 Random Forest (RF) Model Performance

As we have implemented our model on two different datasets, Plasmodium Falciparum and Leukocyte Elastase, model evaluation is presented accordingly.

Performance on Plasmodium Falciparum dataset:

Our Random Forest model gave 81% accuracy when implemented on the Plasmodium Falciparum dataset. The confusion matrix that was generated from the experiment is given below:

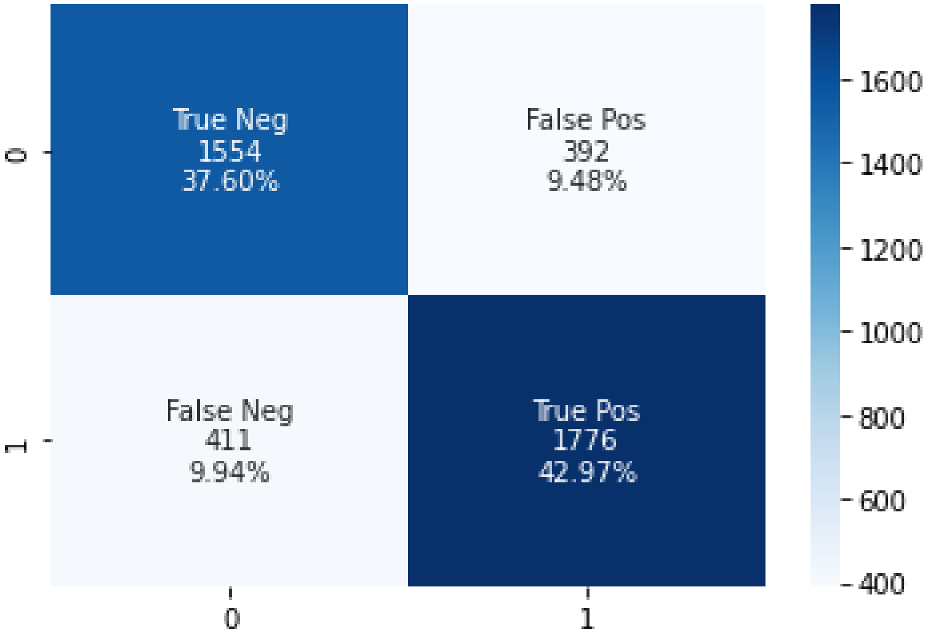


Figure 6.2: Confusion Matrix of RF on Plasmodium Falciparum dataset

The result of other evaluation metrics is as follows:

Table 6.1: Evaluation metrics of RF model on Plasmodium Falciparum dataset

	Precision	Recall	F1-score	Support
0	0.79	0.80	0.79	1946
1	0.82	0.81	0.82	2187
Accuracy	-	-	0.81	4133
Macro avg.	0.81	0.81	0.81	4133
Weighted avg.	0.81	0.81	0.81	4133

Performance on Leukocyte Elastase dataset:

Our Random Forest model gave 90% accuracy when implemented on the Leukocyte Elastase dataset. The confusion matrix that was generated from the experiment is given below:

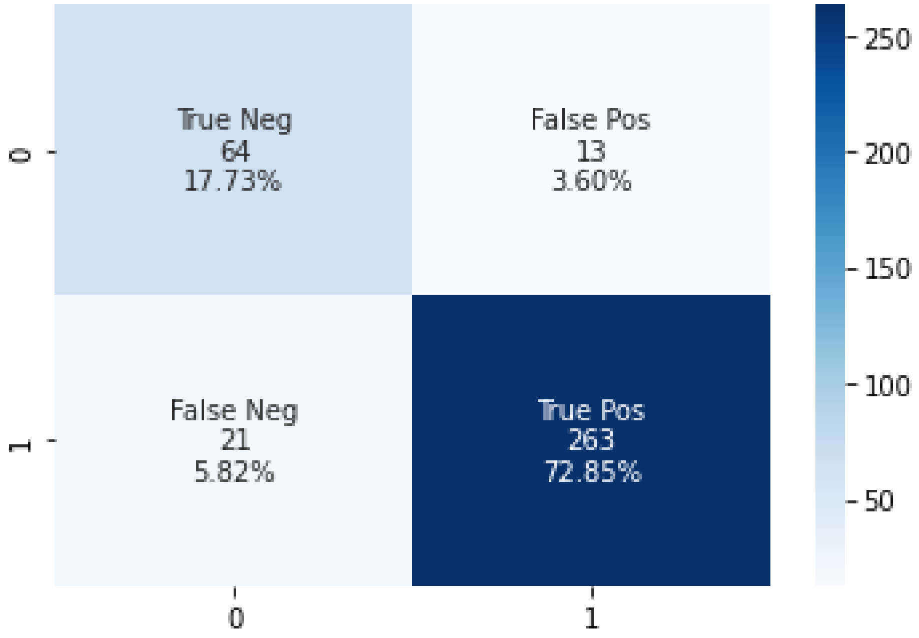


Figure 6.3: Confusion Matrix of RF on Leukocyte Elastase dataset

The result of other evaluation metrics is as follows:

Table 6.2: Evaluation metrics of RF model on Leukocyte Elastase dataset

	Precision	Recall	F1-score	Support
0	0.80	0.78	0.79	87
1	0.93	0.94	0.93	274
Accuracy	-	-	0.91	361
Macro avg.	0.87	0.86	0.86	361
Weighted avg.	0.90	0.90	0.90	361

6.4 Support Vector Machine (SVM) Model Performance

As we have implemented our model on two different datasets, Plasmodium Falciparum and Leukocyte Elastase, model evaluation is presented accordingly.

Performance on Plasmodium Falciparum dataset:

Our Support Vector Machine model gave 80% accuracy when implemented on the Plasmodium Falciparum dataset. The confusion matrix that was generated from the experiment is given below:

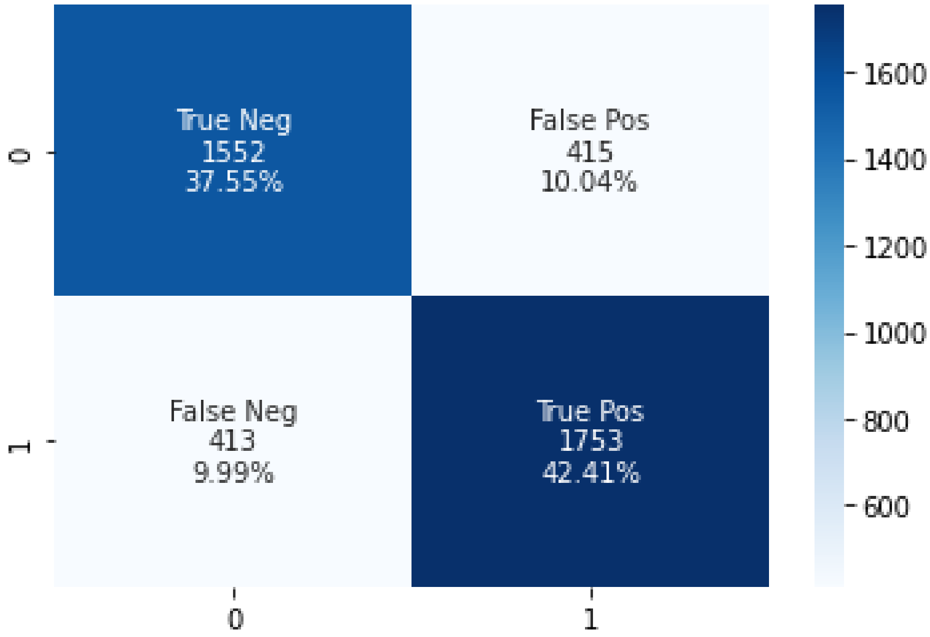


Figure 6.4: Confusion Matrix of SVM on Plasmodium Falciparum dataset

The result of other evaluation metrics is as follows:

Table 6.3: Evaluation metrics of SVM model on Plasmodium Falciparum dataset

	Precision	Recall	F1-score	Support
0	0.79	0.79	0.79	1967
1	0.81	0.81	0.81	2166
Accuracy	-	-	0.80	4133
Macro avg.	0.80	0.80	0.80	4133
Weighted avg.	0.80	0.80	0.80	4133

Performance on Leukocyte Elastase dataset:

Our Support Vector Machine model gave 91% accuracy when implemented on the Leukocyte Elastase dataset. The confusion matrix that was generated from the experiment is given below:

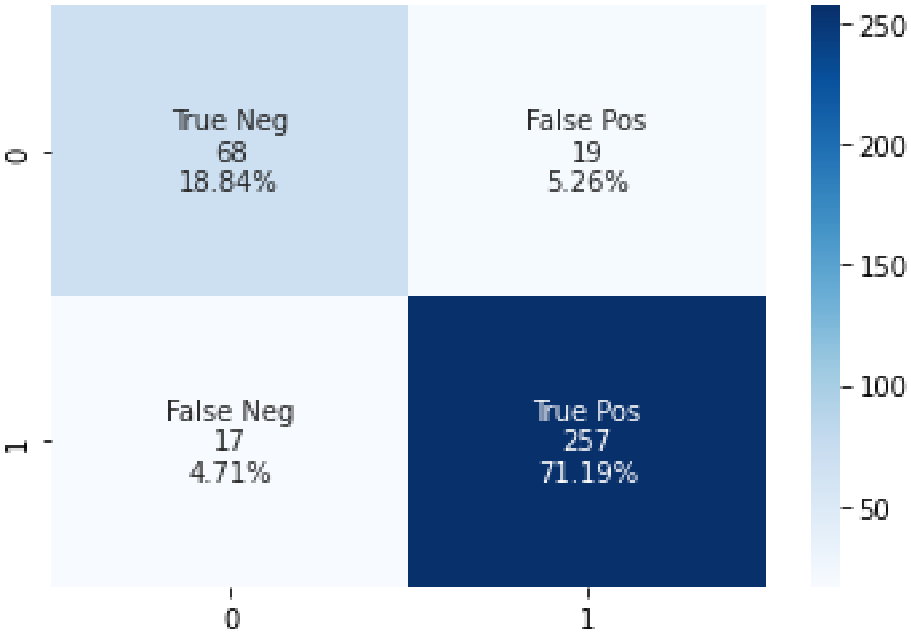


Figure 6.5: Confusion Matrix of SVM on Leukocyte Elastase dataset

The result of other evaluation metrics is as follows:

Table 6.4: Evaluation metrics of SVM model on Leukocyte Elastase dataset

	Precision	Recall	F1-score	Support
0	0.82	0.80	0.81	88
1	0.93	0.95	0.94	273
Accuracy	-	-	0.91	361
Macro avg.	0.88	0.87	0.88	361
Weighted avg.	0.91	0.91	0.91	361

6.5 ROC Curve of RF and SVM Models

ROC curve for Random Forest (RF) and Support Vector Machine (SVM) were generated for the two datasets, Plasmodium Falciparum and Leukocyte Elastase.

ROC curve for Plasmodium Falciparum dataset:

Random (chance) Prediction: AUROC = 0.500

Random Forest: AUROC = 0.871

Support Vector Machine: AUROC = 0.745

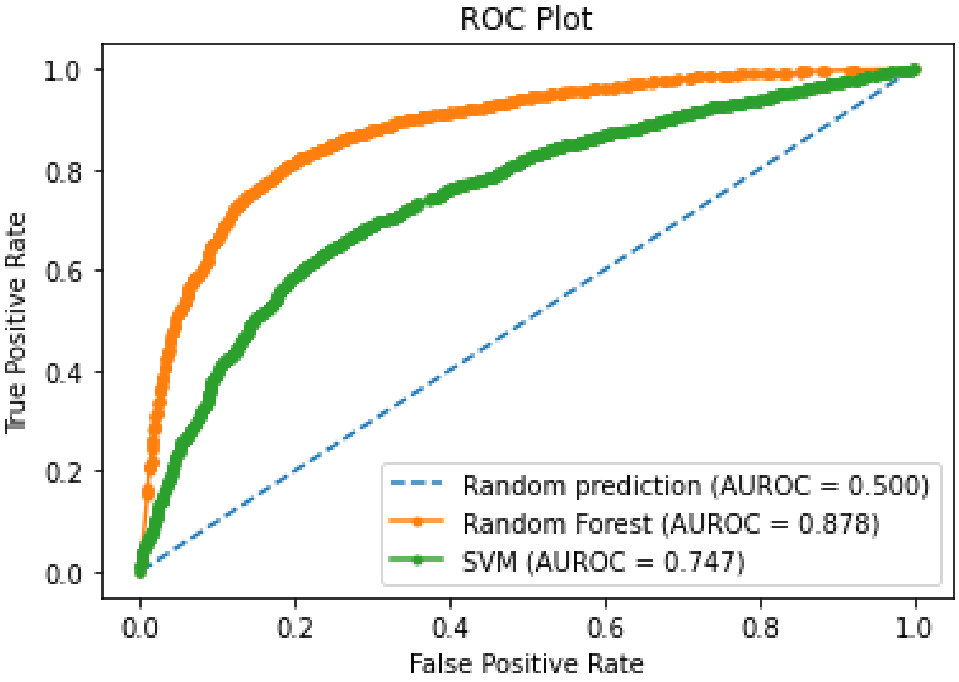


Figure 6.6: RF vs SVM ROC for Plasmodium Falciparum dataset

Here, Random Forest clearly outperforms Support vector machine on AUROC scores.

ROC curve for Leukocyte Elastase dataset:

Random (chance) Prediction: AUROC = 0.500

Random Forest: AUROC = 0.959

Support Vector Machine: AUROC = 0.935

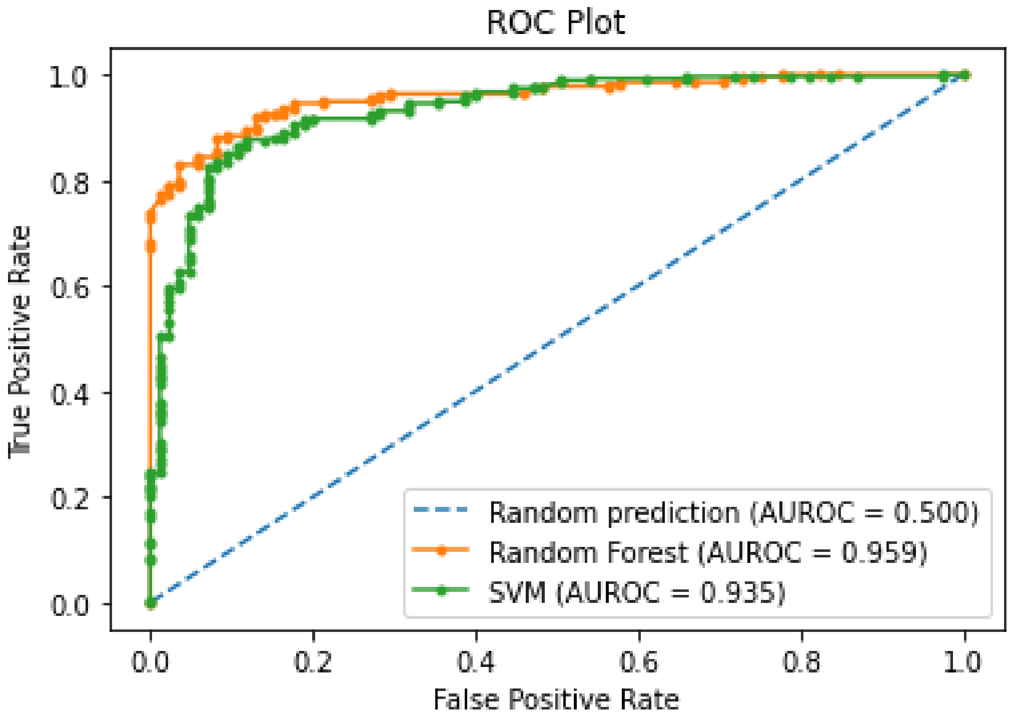


Figure 6.7: RF vs SVM ROC for Leukocyte Elastase dataset

Here, Random Forest clearly and Support vector machine shows similar AUROC scores where Random Forest slightly outperforms Support vector machine.

6.6 Feed Forward Neural Network (FNN) model performance

As we have implemented our model on two different datasets, Plasmodium Falciparum and Leukocyte Elastase, model evaluation is presented accordingly.

Performance on Plasmodium Falciparum dataset:

After running the FNN model on the dataset we got accuracy of 94% on training data and accuracy of 97% on the testing data.

The resultant curve is shown below:

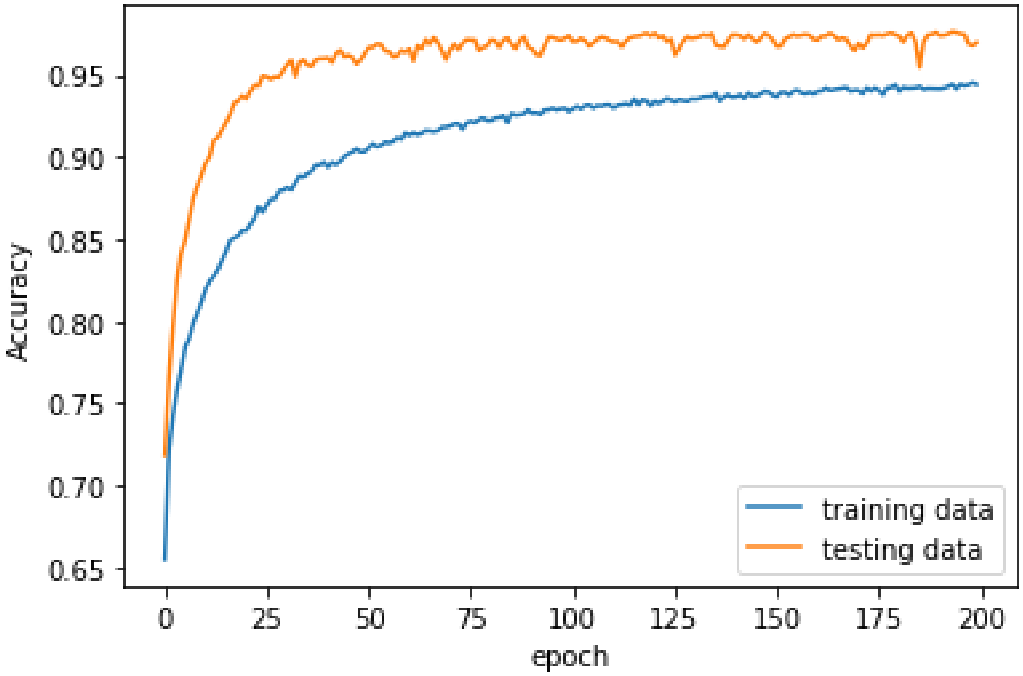


Figure 6.8: Epoch vs Accuracy curve for Plasmodium Falciparum dataset

From the graph, we can see that at first the spike is higher for both training and testing data. Then after few epochs the curves become steady as there is not much change in the accuracy.

Performance on Leukocyte Elastase dataset:

After running the FNN model on the dataset we got accuracy of 97% on training data and accuracy of 99% on the testing data.

The resultant curve is shown below:

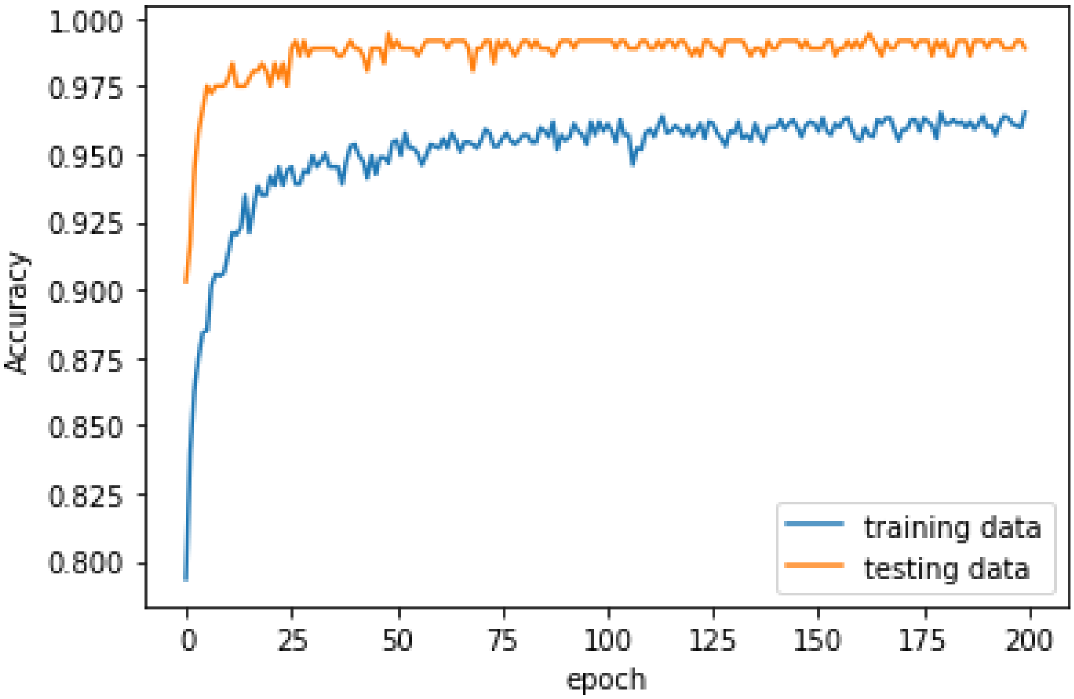


Figure 6.9: Epoch vs Accuracy curve for Leukocyte Elastase dataset

From the graph, we can see that at first the spike is higher for both training and testing data. Then after few epochs the curves become steady as there is not much change in the accuracy.

6.7 Model Comparison for Plasmodium Falciparum Dataset

Here is a histogram showing the model comparison:

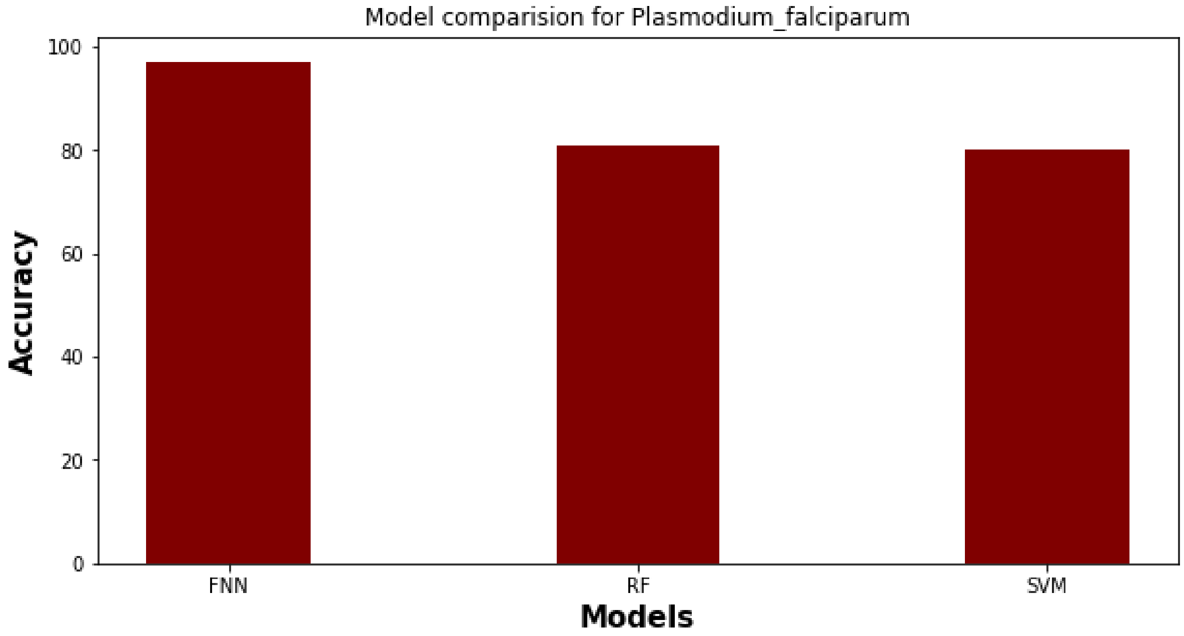


Figure 6.10: Model Comparison for Plasmodium Falciparum Dataset

From the histogram, we can see that Random Forest (RF) and Support Vector Machine (SVM) performs similarly where Feed Forward Neural Network (FNN) model significantly outperforms the other two models for the Plasmodium falciparum dataset. This shows deep learning models have a significantly higher advantage over machine learning models.

6.8 Model Comparison for Plasmodium and Leukocyte dataset

Here is a histogram showing the model comparison:

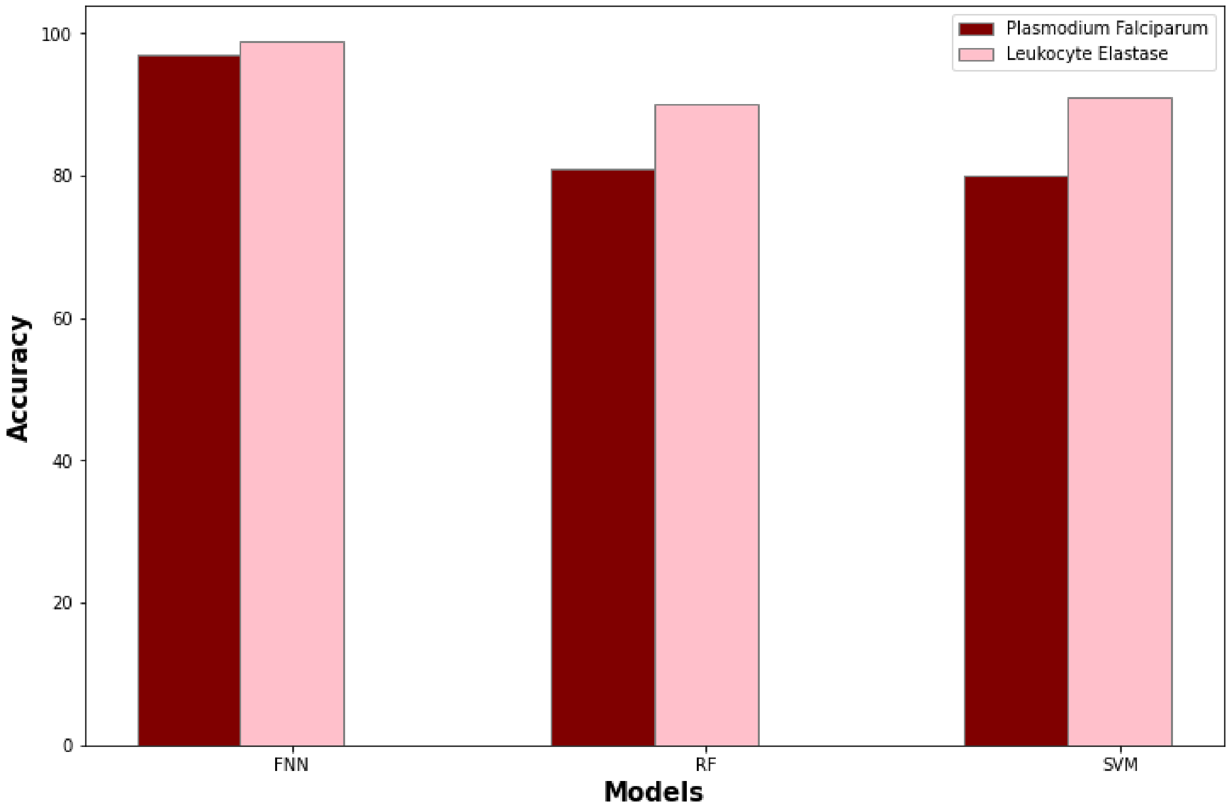


Figure 6.11: Model Comparison for both Dataset

From the histogram, we can see that Random Forest (RF) and Support Vector Machine (SVM) performs similarly where Feed Forward Neural Network (FNN) model significantly outperforms the other two models for the Plasmodium falciparum dataset and the Leukocyte Elastase dataset. This shows deep learning models have a significantly higher advantage over machine learning models.

Also the gap of accuracy FNN and the machine learning models for Plasmodium falciparum dataset is higher than that of Leukocyte Elastase dataset. It is because the number of samples are higher for Plasmodium falciparum, almost ten times higher than Leukocyte Elastase dataset. This shows with larger samples deep learning models tend to perform at higher level whereas the performance of the machine learning models significantly degrades.

6.9 Comparison across Studies

To get a better understanding of the performance of our work, the comparison of the results between our work with past works is shown below:

Table 6.5: Comparison of the results across studies

Research	Model	Target id	Cross Validation	Accuracy (%)
Present	FNN	CHEMBL364	-	97
	SVM			80
Robinson et al. [18]	FNN	CHEMBL1794580	3 Fold	90
	SVM			92
Mayr et al. [19]	Deep learning model	CHEMBL1794352	2 Fold	87

From the table, we can see that Feed Forward Neural Network (FNN) model significantly outperforms Support Vector Machine (SVM) for our research whereas, the models perform similarly for Robinson's [18] research. The deep learning model of Mayr's [19] research also lags behind in the performance in comparison with our research.

This is due to the fact that our dataset is more uniform than theirs and our model is also better optimized.

6.10 Comparison of Different Machine Learning Models for Plasmodium Falciparum

Using the Lazy Predict [38] feature of python we evaluated the results between different machine learning models:

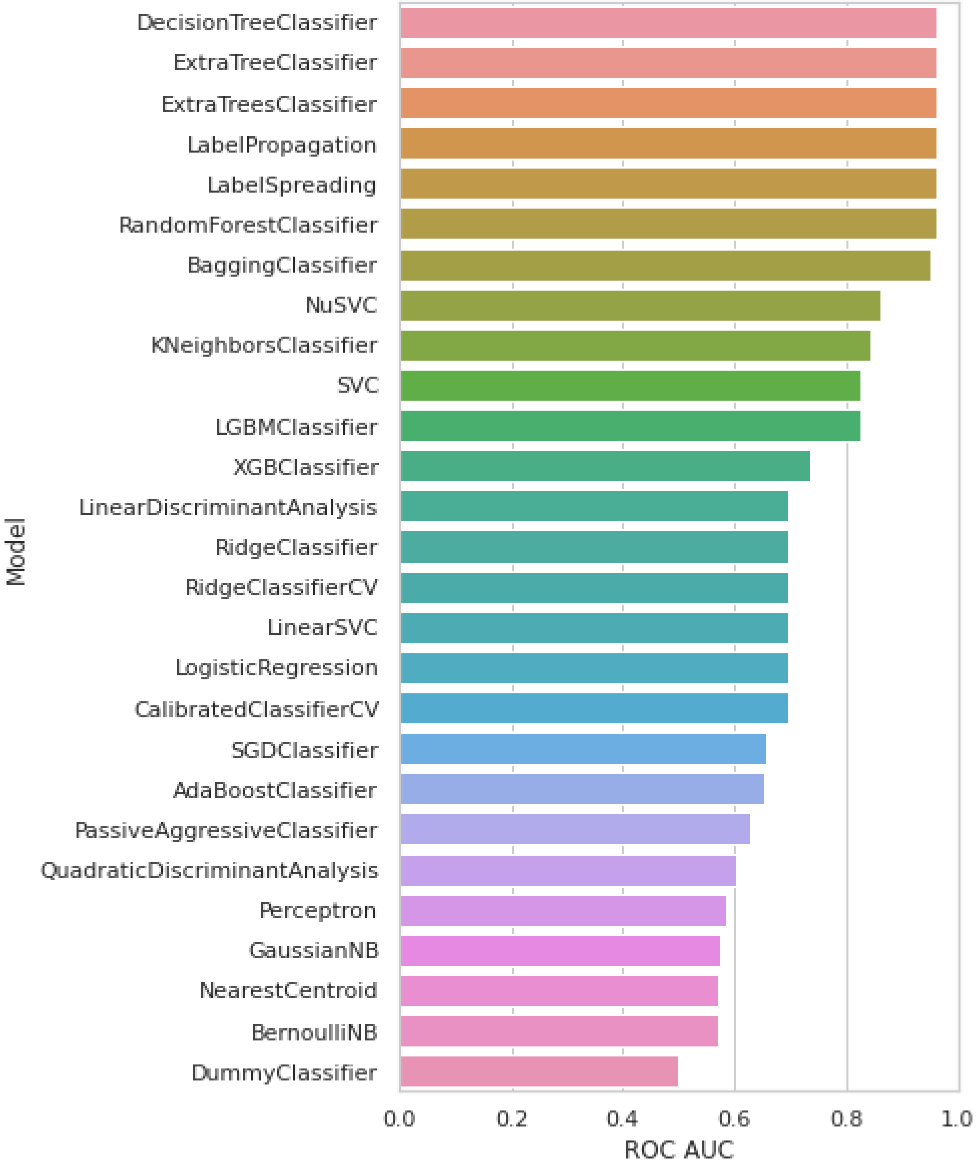


Figure 6.12: Comparison of ML models for Plasmodium Falciparum dataset

Decision Tree classifier models performed best. Random Forest classifier is not far behind.

6.11 Comparison of Different Machine Learning Models for Leukocyte Elastase

Using the Lazy Predict [38] feature of python we evaluated the results between different machine learning models:

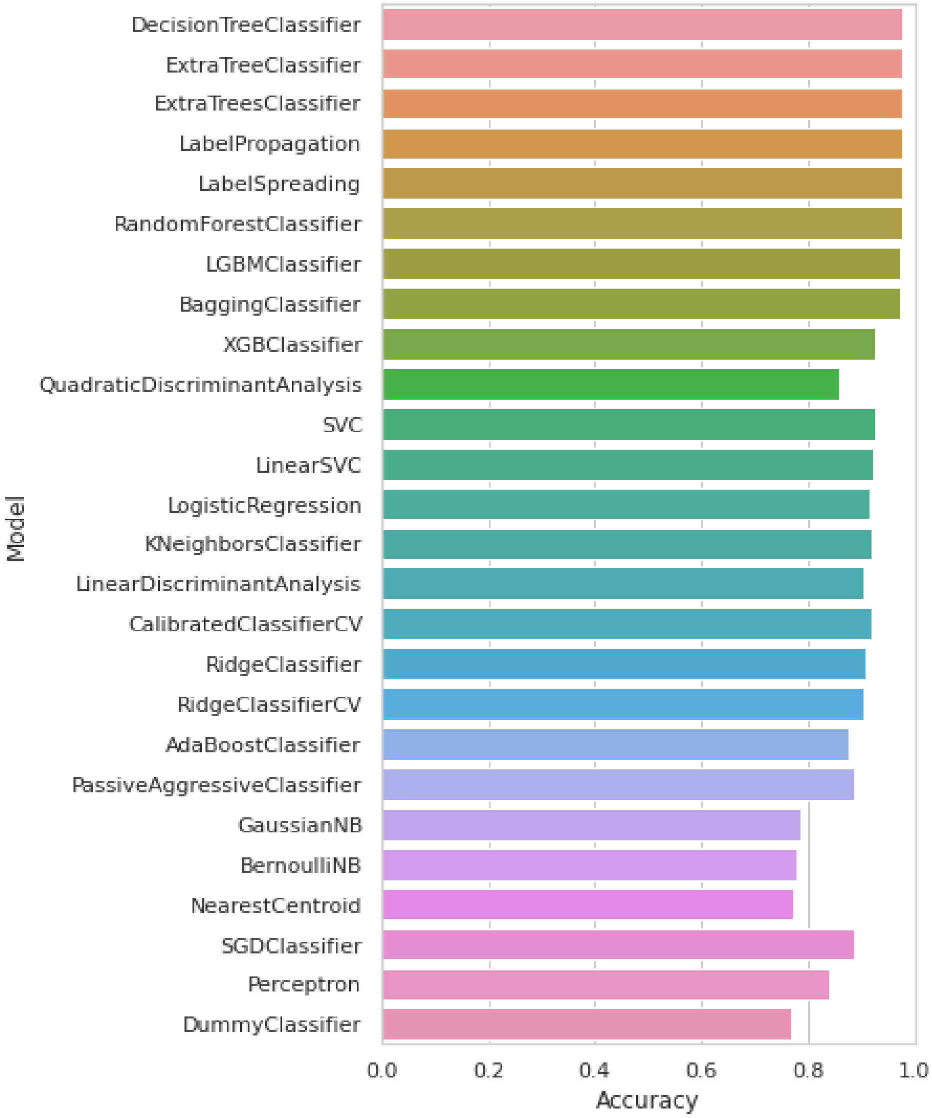


Figure 6.13: Comparison of ML models for Leukocyte Elastase dataset

Here also, Decision Tree classifier models performed best. Random Forest classifier is not far behind.

6.12 Conclusion

In this chapter, we have shown the experimental results and given an analysis. We found out that Feed Forward Neural Network (FNN) outperformed the other machine learning models for both of the datasets. The difference in performance is even higher when the number of samples rises in a dataset. Our results of our research also outperforms results from past researches. This is because our dataset is more uniform and more balanced compared to the datasets of other studies. The next chapter will present a conclusion of our work and will also discuss the future scopes of our work.

CHAPTER 7

Conclusion and Future Works

7.1 Introduction

In the previous chapter, we have shown the experimental results and discussed the performance analysis of our experiments. In this chapter, we have first given a summary of our whole thesis research. Then we have briefly discussed about our contribution in the research. Our research contribution is followed by the limitations of our work. In the next section, directions towards possible future works are provided. Finally, the chapter ends with a final conclusion.

7.2 Thesis Summary

Our thesis can be summarized in the following points:

- Our thesis was on bioactivity prediction using computational approach.
- We collected our data from ChEMBL database.
- We selected Plasmodium Falciparum and Leukocyte Elastase datasets.
- We selected Random Forest and Support Vector Machine classifiers and Feed Forward Neural Network as our machine learning models.
- We preprocessed the data.
- We used PaDEL descriptor to extract features or molecular descriptors from the molecules of the datasets.
- We extracted molecular descriptors or features from the data in the form of pubchem fingerprints.
- We implemented our models on the datasets and found out FNN outperforms the other two machine learning models.
- For the Leukocyte Elastase dataset the difference in performance between Feed Forward Neural Network model and the other two models were not that much.
- For Plasmodium Falciparum dataset the difference in performance between Feed Forward Neural Network model and the other two models were a lot higher.
- It showed when the data sample is larger the difference in the performance of FNN and machine learning models is more significant.

- We concluded that, FNN is the best model for bioactivity prediction from target proteins.

7.3 Research Contribution

We set some objectives before the start of our research work. We believe have been successful in completing those objectives by the end of it. Our main goal was to perform bioactivity prediction from target proteins using different machine learning and deep learning models. We achieved that by the end of our research. Some of the research contribution of our thesis are:

- We performed bioactivity prediction for two different datasets: Plasmodium Falciparum and Leukocyte Elastase.
- We performed our research in a customized dataset only considering IC50 values of the targets which made our datasets uniform.
- We implemented three different models Random Forest, Support Vector Machine and Feed Forward Neural Network.
- Our Feed Forward Neural Network Model gave 97% accuracy for Plasmodium Falciparum dataset and 99% accuracy for Leukocyte Elastase dataset.
- Our Feed Forward Neural Network model outperformed the neural network model from Robinson's [18] research.
- We did a comparison of the results of the three models that we used.
- We found out that Feed Forward Neural Network Model outperformed the other two machine learning models.
- We highlighted some of the limitations of other researches in the field of bioactivity prediction.
- Robinson's research [18] concluded that Support Vector Machine performed on par with deep learning models in bioactivity prediction. We challenged their conclusion and proved their conclusion wrong by showing that Feed Forward Neural Network outperforms Support Vector machine when the dataset is more balanced and uniform
- We showed that the performance of the machine learning models, Random Forest and Support Vector Machine, decreases with the increase of data size of the dataset.
- We showed that Feed Forward Neural Network model performs similarly even with larger datasets.

7.3 Research Limitations

Our research has some limitations. Some of the limitations are:

- Could not work with larger dataset.
- Larger dataset were available but data was not uniform on those datasets.
- Could only work with IC50 standard value.
- Did not consider other factors besides IC50 like the potency, EC50 values etc.
- Only considered active and inactive classes, did not consider the inactive class for drug discovery.
- Could not implement different deep learning models and compare them for the same datasets for complexity.
- GPU had limitations.

7.4 Future Works

Bioactivity prediction is an emerging field in the research area. So there are much scope for future researches in this field which can bring more advancement in drug discovery.

Datasets of targets are increasing rapidly. And much of that data is publicly available in ChEMBL database and on other sources. Further research can be done in the field of drug discovery with the use of new datasets.

Variety of data is also increasing. Scope for implementing different machine learning and deep learning models in diverse will be possible in future.

In future, we wish to apply our model on newly available datasets. We larger datasets. We also wish to try out different deep learning models for same dataset and to compare their results.

As we only considered IC50 values in this research, we wish to work with different standard values like potency, inhibition, EC50 if larges sample sizes with these values are available in the future.

7.5 Conclusion

Drug discovery is a noble cause for mankind. Using computational approaches the drug discovery process have been significantly accelerated. The drug discovery process can further be accelerated and improved using powerful machine learning and deep learning models. Moreover, newer datasets requires newer researches. In this paper, we implemented Random Forest, Support Vector Machine and Feed Forward Neural Network model to do drug discovery in two different target protein datasets. We concluded, Feed Forward Neural Network as the best performing model for drug discovery among the other tested methods, giving accuracy of 97% for large dataset of Plasmodium Falciparum.

REFERENCES

[1] Ho, B., Baryshnikova, A. and Brown, G., 2018. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Systems*, 6(2), pp.192-205.e3.

[2] Staszak, M., Staszak, K. and Wieszczycka, K., 2021. Machine learning in drug design: Use of artificial intelligence to explore the chemical structure–biological activity relationship. *WIREs Computational Molecular Science*, 12(2).

[3] Vos, Theo, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, 386(9995), pp.743-800.

[4] "Who coronavirus (COVID-19) dashboard," *World Health Organization*. [Online]. Available: https://covid19.who.int/. [Accessed: 8-Oct-2022].

[5] Baskin, I., 2020. The power of deep learning to ligand-based novel drug discovery. *Expert Opinion on Drug Discovery*, 15(7), pp.755-764.

[6] Zeng, X., Zhu, S., Lu, W., Liu, Z., Huang, J. and Zhou, Y., 2020. Target identification among known drugs by deep learning from heterogeneous networks. *Chemical Science*, 11(7), pp.1775-1797.

[7] Husic, B., Charron, N., Lemm, D., Wang, J. and Pérez, A., 2020. Coarse graining molecular dynamics with graph neural networks. *The Journal of Chemical Physics*, 153(19), p.194101.

[8] Stokes, J., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N., MacNair, C., French, S., Carfrae, L., Bloom-Ackermann, Z. and Tran, V., 2020. A Deep Learning Approach to Antibiotic Discovery. *Cell*, 180(4), pp.688-702.e13.

[9] Wang, Y., You, Z., Yang, S. and Yi, H., 2020. A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. *BMC Medical Informatics and Decision Making*, 20(S2).

[10] Chen, C., Shi, H., Jiang, Z., Salhi, A. and Chen, R., 2021. DNN-DTIs: Improved drug-target interactions prediction using XGBoost feature selection and deep neural network. *Computers in Biology and Medicine*, 136, p.104676.

[11] Lee, I., Keum, J. and Nam, H., 2019. DeepConv-DTI: Prediction of drug-target interactions

via deep learning with convolution on protein sequences. *PLOS Computational Biology*, 15(6), p.e1007129.

[12] Zhang, J., Jiang, Z., Hu, X. and Song, B., 2020. A novel graph attention adversarial network for predicting disease-related associations. *Methods*, 179, pp.81-88.

[13] Zhang, C., Lu, Y. and Zang, T., 2022. CNN-DDI: a learning-based method for predicting drug–drug interactions using convolution neural networks. *BMC Bioinformatics*, 23(S1).

[14] Tolles, J. and Meurer, W., 2016. Logistic Regression. *JAMA*, 316(5), p.533.

[15] Ghahramani, Z., 2015. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553), pp.452-459.

[16] Cano, G., Garcia-Rodriguez, J. and Garcia-Garcia, A., 2017. Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert Systems with Applications*, 72, pp.151-159.

[17] Zane, P., Gieschen, H., Kersten, E. and Langguth, P., 2019. In vivo models and decision trees for formulation development in early drug development: A review of current practices and recommendations for biopharmaceutical development. *European Journal of Pharmaceutics and Biopharmaceutics*, 142, pp.222-231.

[18] Robinson, M., Glen, R. and Lee, A., 2020. Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. *Journal of Computer-Aided Molecular Design*, 34(7), pp.717-730.

[19] Mayr, A., Klambauer, G., Unterthiner, T. and Steijaert, M., 2018. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science*, 9(24), pp.5441-5451.

[20] Richardson, A., Pollak, E., Williams, D. and Smith, M., 2010. Intrauterine Infection. *Comprehensive Toxicology*, pp.239-258.

[21] “Plasmodium falciparum,” *Wikipedia*, 30-Aug-2022. [Online]. Available: https://en.wikipedia.org/wiki/Plasmodium_falciparum. [Accessed: 11-Oct-2022].

[22] LEE, W. and DOWNEY, G., 2001. Leukocyte Elastase. *American Journal of Respiratory and Critical Care Medicine*, 164(5), pp.896-904.

[23] “Machine learning random forest algorithm - javatpoint,” *www.javatpoint.com*. [Online]. Available: https://www.javatpoint.com/machine-learning-random-forest-algorithm. [Accessed: 10-Oct-2022].

[24] García-Gonzalo, E., Fernández-Muñiz, Z., García Nieto, P., Bernardo Sánchez, A. and Menéndez Fernández, M., 2016. Hard-Rock Stability Analysis for Span Design in Entry-Type Excavations with Learning Classifiers. *Materials*, 9(7), p.531.

[25] "Multilayer feed-forward neural network in Data Mining," *GeeksforGeeks*, 08-Sep-2022. [Online]. Available: https://www.geeksforgeeks.org/multilayer-feed-forward-neural-network-in-data-mining/. [Accessed: 11-Oct-2022].

[26] Reker, D., 2019. Practical considerations for active machine learning in drug discovery. *Drug Discovery Today: Technologies*, 32-33, pp.73-79.

[27] "Chembl Database," *EBI*. [Online]. Available: https://www.ebi.ac.uk/chembl/. [Accessed: 10-Oct-2022].

[28] Sarker, I., 2021. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, 2(6).

[29] Sarker, I., 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3).

[30] Carracedo-Reboredo, P., Liñares-Blanco, J., Rodríguez-Fernández, N., Cedrón, F. and Novoa, F., 2021. A review on machine learning approaches and trends in drug discovery. *Computational and Structural Biotechnology Journal*, 19, pp.4538-4558.

[31] "Difference between ANN, CNN and RNN," *GeeksforGeeks*, 24-Aug-2022. [Online]. Available: https://www.geeksforgeeks.org/difference-between-ann-cnn-and-rnn/. [Accessed: 10-Oct-2022].

[32] J. Brownlee, "When to use MLP, CNN, and RNN Neural Networks," *Machine Learning Mastery*, 15-Aug-2022. [Online]. Available: https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/. [Accessed: 10-Oct-2022].

[33] C. Nantasenamat, "How to build your first machine learning model using no code," *Medium*, 25-Jun-2021. [Online]. Available: https://towardsdatascience.com/how-to-build-your-first-machine-learning-model-using-no-code-a7cf8db37dd1. [Accessed: 10-Oct-2022].

[34] "Simplified molecular-input line-entry system," *Wikipedia*, 21-Jul-2022. [Online]. Available: https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system. [Accessed: 10-Oct-2022].

[35] Alsenan, S., Al-Turaiki, I. and Hafez, A., 2020. Feature Extraction Methods in Quantitative Structure–Activity Relationship Modeling: A Comparative Study. *IEEE Access*, 8, pp.78737-78752.

[36] Yap, C., 2010. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7), pp.1466-1474.

[37] Fernández-de Gortari, E., García-Jacas, C., Martinez-Mayorga, K. and Medina-Franco, J., 2017. Database fingerprint (DFP): an approach to represent molecular databases. *Journal of Cheminformatics*, 9(1).

[38] “Welcome to lazy predict's documentation!” *Welcome to Lazy Predict's documentation! – Lazy Predict 0.2.12 documentation*. [Online]. Available: https://lazypredict.readthedocs.io/en/latest/. [Accessed: 10-Oct-2022].