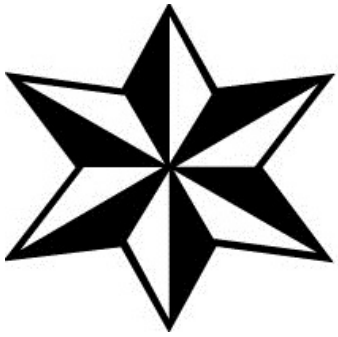

Machine Learning in Drug Discovery and Design

Predicting the Blood Brain Barrier Penetration of Drugs

Ibraheem Ajibola Ganiyu

A dissertation presented for the degree of
BSc. (Hons) Software Engineering.

School of Computing, Engineering and Mathematics
University of Brighton

Contents

List of Figures

List of Tables

Abstract

Drug design and discovery is a very expensive process and lots of new compounds are being developed rapidly. Only roughly about 2% of drugs can pass through the blood brain barrier, this presents a problem in Central Nervous System (CNS) drug development.
This Project aims to develop a solution that can predict with high confidence, the probability of a drug passing through this blood brain barrier in hopes that this can speed up the process of developing a CNS drug.

Chapter 1

Introduction

Chemical data is growing exponentially as there are currently more than 123 million organic and inorganic substances to date [5], which means for drug development purposes, there is an abundance of chemical data to analyse in search for ideal candidates for development.
Machine Learning is a form of artificial intelligence (AI) that enables computer programs to learn concepts from data without being explicitly programmed. This technique of AI can be applied to problems in many domains, which is what this paper aims to do by applying it to chemical data to build computer models, aka Virtual Screening, models which can predict with high accuracy, the probability of a drug to pass through the blood brain barrier (BBB) of living organisms.
In Silico models (computer models) have a profound use in virtual screening as they can enable scientist to scan through a large database of drugs to speed up a time consuming process of analysing drug candidates.
In the context of CNS (Central Nervous System) drug development, when a drug is absorbed into the blood stream, it needs to be able to pass through the Blood Brain Barrier (BBB) to its target which could be the brain or the nervous system, an example would an anti-migraine agent, ergotamine,

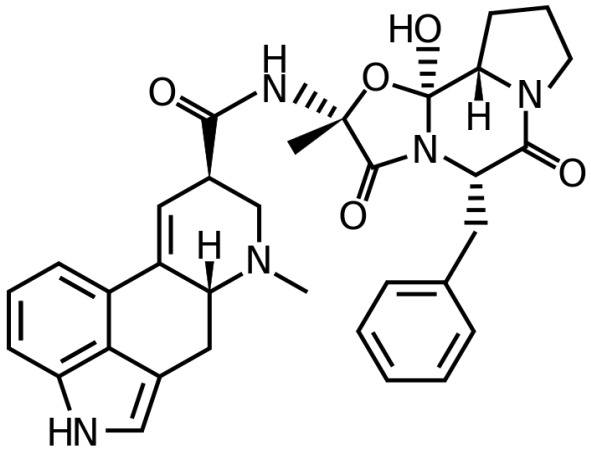

Figure 1.1: Chemical Representation of Ergotamine

which has to pass through the blood brain barrier to the lining of the brain where it constricts the blood vessels there to decrease the pain from migraine headaches [8]. However statistically speaking, only 2% of the currently known molecules can pass through this blood brain barrier, which translates to more time in the drug discovery pipeline spent on analysing drug candidates that can pass through the BBB barrier.

1.1 Proposed Solution

This problem of determining the drug candidates that belong to this 2% is a very challenging one and this paper approaches the problem through the use of computer models built with machine learning techniques, applied to a large database of molecules that pass through the Blood Brain Barrier, with the aim of predicting the probability of an unknown candidate drug passing through the BBB.
Ana et al [9] points out that there is a lack of extensive dataset on BBB prediction as most of them are not comprehensive enough to build complex models out of, as a result, they have compiled a dataset of 2040 molecules for use in BBB prediction. Based on analysis carried out [9] by Ana et al, They show that certain machine learning classifiers such as support vector machines and random forests outperform other classifiers especially for BBB classification tasks. This paper will attempt utilise that analysis as a baseline for developing our computer models whilst also exploring the possibility of Deep neural networks, as Thomas et al [17] show that Deep learning techniques have a significant opportunity in virtual screening.

1.2 Results

Chapter 2

Background

Everything around us is composed of molecules, they are an electrically neutral group of two or more atoms held together by chemical bonds. They are the smallest particle in a compound that exhibit the chemical properties of the compound. Trees can be said to be made up of molecules and historically parts of trees have always been used to cure or alleviate symptoms of illness. Over time, the individual molecules in these herbal medicines were recognized for their effects and they were being produced synthetically, which further gave rise to Modern Drug Design and Discovery.
A potential molecule aka drug candidate is usually screened against a target protein to test its effectiveness and the screening can either be virtual (Virtual Screening) or through a method known as High Throughput Screening (HTS) - a method of experimentation involving the use of robots and control software to conduct millions of scientific tests. These HTS machines can also be credited with the exponential growth in chemical data [7]. Drug discovery is a very expensive and time consuming process; It is usually broken down into numerous stages with the most expensive stage being the clinical trials.
The problem of the Blood Brain Barrier (BBB) prediction evolves around Chemistry and Computer Science being applied to the Drug Design domain with a thorough understanding of the constrainsts imposed by the Central Nervous System. This chapter introduces the necessary concepts needed to understand the solution taken to the BBB prediction problem.

2.1 Cheminformatics

Cheminformatics aka Chemoinformatics and Chemical Informatics can be visualised as a cross domain of Chemistry and Computer Science. As defined by Brown, it is the mixing of numerous information that a scientist needs to trans-

form data into information for the intended purpose of making better decisions in drug lead identification and optimisation [4].
The applications of Cheminformatics that are of particular interest are *Storage and Retrieval of Chemical Data* and *Virtual Libraries and Screening* as they both would enable the manipulation and transformation of molecular information for our machine learning algorithms.

2.1.1 Representing Molecules in Computers

The two most common formats for representing chemical molecules in computers are Simplified Molecular-Input Line Entry System (SMILES) and SMART, with SMART being created by Daylight Chemical Information Systems and both formats being actively supported by them [13].

SMILES are specifications in the typographical notation system that describe the structure of a molecule using short ASCII strings and they are more commonly used. Water can be written in the SMILE format as $[OH2]$ An example would be the SMILE representation of protriptyline

$$CNCCCC1c2ccccc2C = Cc3ccccc13$$

This format can then be utilised by cheminformatics software to extract meaningful chemical information about the molecule. Throughout the project, the Open-Source Cheminformatics Software, RDKit, was used to transform the SMILEs in the dataset and also for the extraction of molecular information, which will be explained below.

2.1.2 Molecular Representation and Similarity

The simplest way to cluster molecules together in a chemical database when performing virtual screening is through the concept of *molecular similarity*. It is the core of Molecular Similarity Analysis (MSA) where the similarity measure, that characterizes the degree of proximity between pairs of molecules are manifested by their "molecular patterns", which are compromised of sets of features (chemical descriptors) [2].
Many different molecular similarity measures exists and some are more suitable to certain virtual screening tasks than another, which is what prompts the discussion of molecular similarity measures in section 2.1.2. The motivation behind the discussion is that it is possible to compare the efficiency of each similarity measure in terms of computer resources (computer memory and time) so as to determine its applicability to the target chemical dataset.

To determine the similarity between molecules, some form of similarity measure has to exist to compare molecules in the same representation. The similarity measures (similarity coefficient) are functions that maps pairs of compatible molecular representation into a real number.

Molecular Representation

There also is the notion of a chemical space that maps the chemical features into some form of structure, which are mostly coordinate-based.
Some common representations of molecules include encoding the molecular data into a set, graph, vector or function-based representation that uses distance as a form of molecular similarity [2]. Each representation has its advantages and disadvantages, where graphs have their shortcomings, the molecule could be represented as a feature vector of its chemical properties which can be its molecular fragments, partial atomic charge, molecular weight, logP etc.
For our blood brain barrier prediction task, the molecules in the training set were transformed into their molecular fingerprint representation and simple molecular descriptors feature set. Here the molecule is represented as a point in a 2D or 3D coordinate space based on the derived features by performing Principal Component Analysis (PCA) or Non-linear Mapping (NLM) on the original set of features of the molecule.

Molecular Fingerprints as a Binary-Valued Feature Vector Each fingerprint consists of an n-component bit vector, 2048 bits for a Morgan fingerprint representation but the sizes of the vectors can vary depending on the kernel function used. The vector is given by

$$\vec{V_A} = (v_A(x_1), v_A(x_2), ..., v_A(x_k), ..., v_A(x_n)) \tag{2.1}$$

where x_k indicates the absence or presence of a given feature [2]. i.e

$$v_A(x_k) = \begin{cases} 1 & \text{Feature present} \\ 0 & \text{Feature absent} \end{cases} \tag{2.2}$$

Similarity Measures

For our prediction task, the main use of molecular similarity measures is in similarity searching when performing the nearest neighbour search. Here the molecules are ordered by their chosen chemical descriptors when we apply our similarity measure to calculate some form of structural relatedness between a target molecule and every other molecule in the dataset. The result is then a sorted list of molecules

where the most similar molecules to our target molecule are located at the top of the list and in order of decreasing similarity.
According to Jurgen [2], the most important components of similarity measures are

- The Representation: Which is used to characterize the molecules that are being compared. An example would be molecular fingerprint representation of a molecule.
- The Weighing Scheme: Used to assign differing degrees of importance to the various components of the molecular representation. An extra measure of accuracy in the weighing schemes would be the use of a chemical ontology database (e.g CheBL) when determining the importance of each component [16].
- Similarity Coefficient: A quantitative measure of the degree of structural relatedness between two molecules.

The main application of any similarity measure is to the fact that **structurally similar molecules exhibit similar properties** as stated by Johnson and Maggiora (1990) [10]. However, they also noted that an exception to the rule is that sometimes a small change in the structure of the molecule can result in a radical change in the properties that it exhibits. Which is why Ana Teixeira (2014) [16] notes that the use of a chemical ontology database would be a plausible method to combat this exception. For our prediction task, the exception to this rule is ignored for practical purposes as majority of molecules exhibit similar properties to one another.
In similarity measure calculations, the most common measure for calculating fingerprint similarity is the Tanimoto coefficient, given by

$$Tanimoto(\vec{v_i}, \vec{v_j}) = \frac{\vec{v_i} \bullet \vec{v_j}}{\sum_k v_{ik} + \sum_k v_{jk} - \vec{v_i} \bullet \vec{v_j}} \tag{2.3}$$

where $\vec{v_i}$ and $\vec{v_j}$ are the bit vectors for molecule i and j. There are other similarity coefficients such as the Tversky and Dice, the Dice similarity measure was chose for the k-nearest neighbour calculation of dataset.

2.2 Drug Design and Discovery

According to Dr A.N Boa [3], the different stages of drug discovery are

- Programme Target Selection (Choosing the disease to work on)

- Identification and Validation of the drug target
- Assay Development
- Identification of a Lead Compound
- Lead Optimisation
- Identification of a drug candidate
- Clinical Trials
- Release of the drug
- Follow-up Monitoring

Majority of the targets for the drugs we consume are usually proteins e.g enzymes, receptors and nucleic acids and the structure of the target is confirmed through a virtual screening method known as *molecular docking*; it can be used to predict how the drug will bind to its target protein though various search/optimisation algorithms [2]. Another Virtual Screening technique usually used is *Quantitative Structure-Activity Relationships* (QSAR), here the underlying idea is that molecules with similar structures behave in the same way, as a result, the activity of a protein against a certain group of compounds is recorded and a QSAR model is the constructed from there and used to determine whether a given compound will bind to the target, thus screening the virtual compound library for drugs of interest.

2.3 Central Nervous System Drug Design and the Blood Brain Barrier

CNS Drugs aiming to pass through the Blood Brain Barrier often need to possess certain physical-chemical properties; some of which are Hydrogen bonding, ionization properties, molecular flexibility etc. [14]. The epithelial cells form an interface between the blood and the brain and these are commonly referred to as the Blood Brain Barrier. This interface occurs in other places within the body but what makes the BBB epithelial cells different are the tight junctions they form which makes it harder for drugs to pass through. Hassan and George (2005) further claim in their article that the majority of BBB penetration is through passive diffusion through the cellular membrane.

ADME Properties of CNS Drugs

For a CNS drug to be therapeutically effective, it must be easily disposed asides from having a high degree of potency. The ADME properties of a drug refers to its ability to be easily absorbed, distributed, metabolised and excreted. Some properties of CNS drugs affect their ADME properties, some of which are [14]:

- Solubility: A drug must be very soluble in the blood and still be in high enough concentration at its target, in this case, the Blood Brain Barrier, so that it can easily be absorbed.
- Amount of Protein Binding: Majority of CNS drugs tend to have high binding property towards proteins - this results in the drug being metabolised easily.
- Partition Coefficient (LogP): This is sometimes referred to as the *lipophilicity* of the compound and has served as one of the most important factors in drug design. Higher lipophilicity results in drugs with higher metabolic turnover but lower solubility and absorption [14].

2.4 Machine Learning concepts

2.4.1 Models and Classifiers

2.4.2 Feature Extraction

Chapter 3

Data Representation and Feature Engineering

3.1 Data Preprocessing

3.2 Feature Extraction

Chapter 4

Machine Learning Classifier Training

4.1 Neural Networks

4.2 Ensemble Classifiers

Chapter 5

Model Evaluation and Improvement

5.1 Model Evaluation

5.2 Model Improvement techniques

5.2.1 Grid Search

5.3 Model Persistence

Chapter 6

Conclusion

6.1 Integration into a Web Application

6.2 Deployment

6.3 Areas for improvement

Chapter 7

Notes: Machine Learning in Drug Design (Will be removed later)

With the recent explosion in chemical data available, it is possible to combine them with machine learning techniques to create models of molecules that have a strong binding affinity to their target proteins. According to Antonio (2015) [11], Virtual screening techniques can either be Structure-Based or Ligand-Based. Structural screening used the idea of molecular docking as described earlier whilst Ligand-Based screening uses the idea that similar molecules exhibit similar properties e.g QSAR.
With Ligand-Based Virtual Screening being the core focus of this paper, we can further classify them into either similarity searching or compound classification [11]. Antonio (2015) [11] further highlights that the following are the most popular and successful techniques for Ligand-Based Virtual Screening:

- Support Vector Machines
- Decision Trees
- Random Forests (An Ensemble of Decision Trees)
- Naive Bayesian Classifiers
- k-Nearest neighbours
- Artificial Neural Networks

7.1 Data Preprocessing and Feature Engineering

7.1.1 Data Visualisations

Attempting to explore how the data looks, we start with a scatter plot of the data. Due to the high dimensionality of the data. It is near impossible to plot a scatter plot in 2D, so we apply a Principal Component Analysis (PCA) on the data to extract the two most important component as shown below

Manifold Learning with t-SNE Manifold Learning Algorithms excel at data visualisation tasks, they provide more complex mappings and often better data representations but perform poorly for data classification tasks. They perform poorly because they do not allow transformations of new data once they have been fitted.
The t-SNE algorithm is used because it tries as much as possible to maintain the distance between the data points in the original feature vector space during transformation. It puts more emphasis on points that are close by rather than preserving points that are far apart [1].

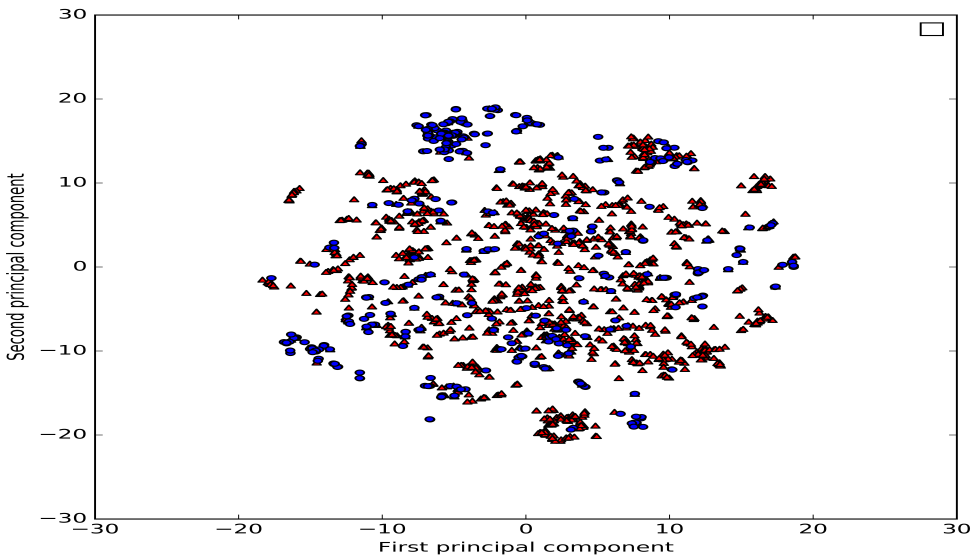

Figure 7.1: 2D Scatter Plot after applying the t-SNE Algorithm

Looking at the t-SNE scatter plot, there is a small cluster of the molecules that do not pass through (B) at the top and some randomly distributed throughout the plot, whilst the ones that pass through the BBB barrier (A) maintain an

even distribution around the feature space, it is worth mentioning that performing passing the simple molecular data to the algorithm, it was preprocessed with a MaxAbsoluteScaler that scales each feature by its maximum value. This doesn't mean that it is impossible to separate the data, as an accuracy of $\sim 89\%$ will be achieved in later sections. This result is to be expected as the data is biased towards molecules that cross the barrier and it also hints at the fact that the molecules that do not cross the barrier might share a lot in common with molecules that cross the barrier.

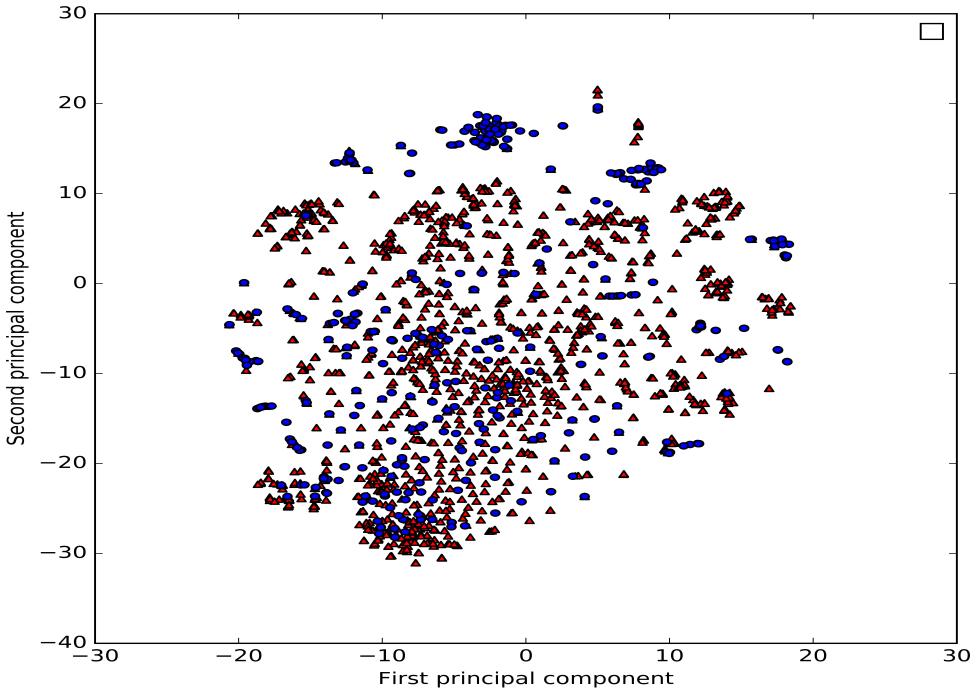

Figure 7.2: 2D Scatter Plot after applying the t-SNE Algorithm (Morgan Fingerprint dataset)

The scatter plot has no significant difference with the Morgan Fingerprint dataset, it highlights the clusters of the B molecules more visibly. Further analysis would need to be carried to create a more linear separable representation.

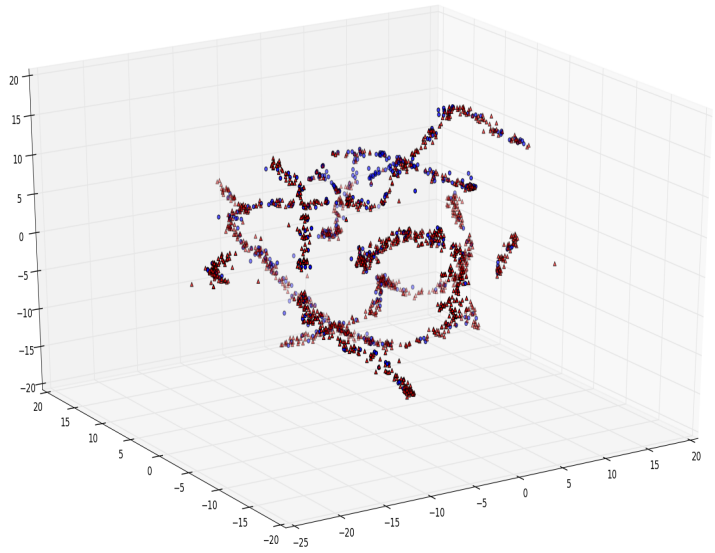

Figure 7.3: 3D Scatter Plot after applying the t-SNE Algorithm

The 3D scatter plot of the simple molecular descriptor data looks slightly more separable than its 2D counterpart.

Principal Component Analysis (PCA) In the 2D representation shown below, the data points are tightly clustered in the middle, and looks like it is linearly inseparable.

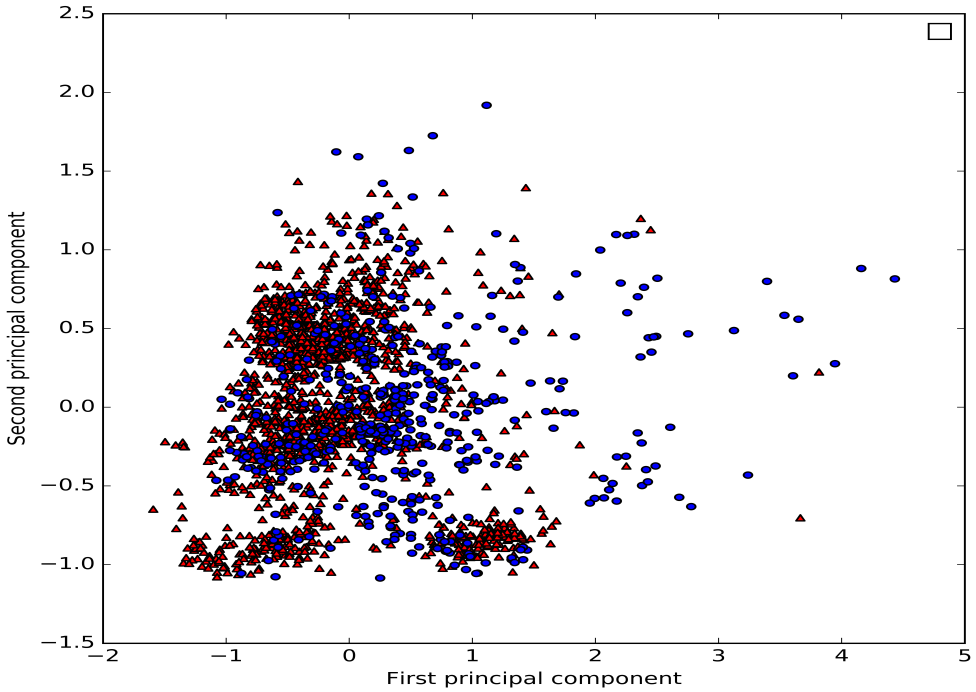

Figure 7.4: 2D Scatter Plot after applying the PCA Algorithm

Remembering that the dataset is biased towards the molecules that pass through the barrier, as they represent around 70% of the data. A significant proportion of the molecules that do not pass through the barrier(B) are clustered in the same space as the ones that pass through the barrier (A). Although majority of A stays clustered in the same space with the exception of a few outliers.
To further examine the data, we look at creating a 3D representation, the next 3 principal components are selected to create data for the 3D plot. The 3D plot has shown below also has the same clustering of a significant proportion of the B molecules in the same space as the A molecules but looks more linearly separable than the 2D plot

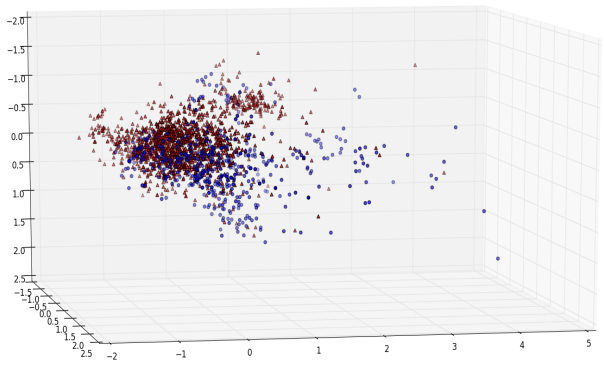

Figure 7.5: 3D Scatter Plot after applying the PCA Algorithm (I)

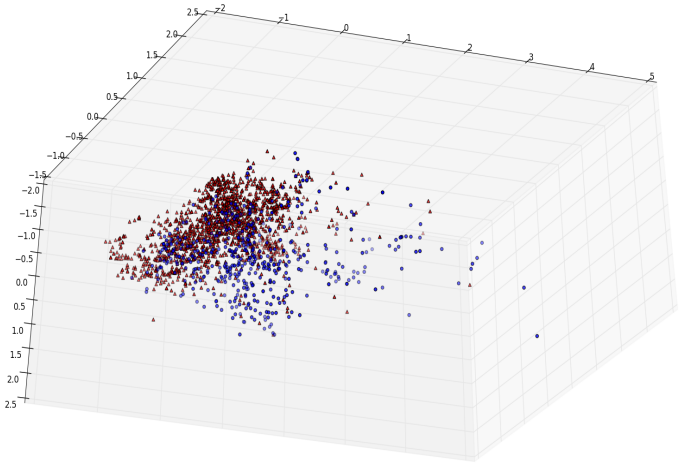

Figure 7.6: 3D Scatter Plot after applying the PCA Algorithm (II)

Clustering

k-Means Clustering In our further attempts to have some form of linearly separable data, we select the 2 principal components of our datasets and apply the k-means clustering algorithm on them.

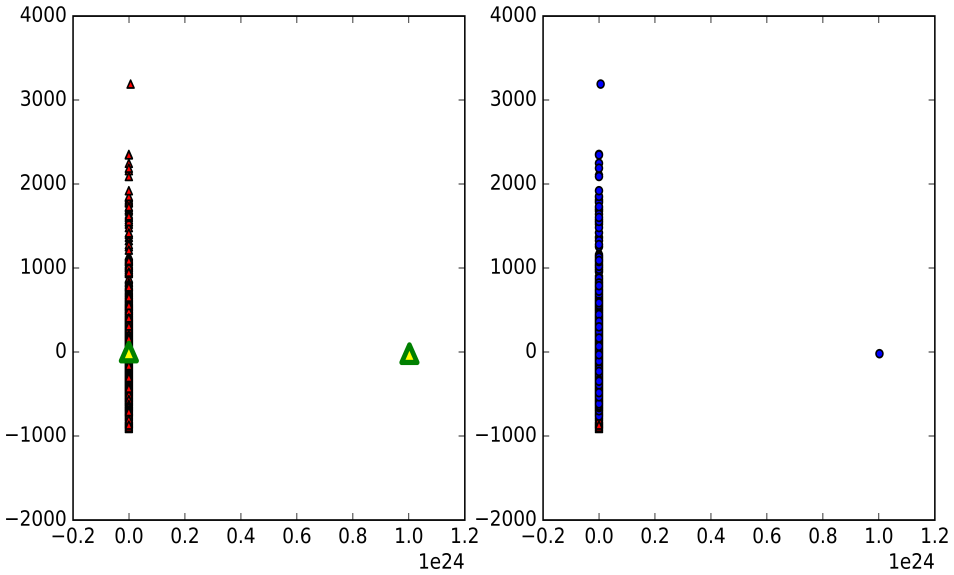

Figure 7.7: 2D Scatter Plot found by k-means on Simple Molecular Descriptors. *Original data points on the right and clusters found on the left*

We can see that the clustering algorithm performs poorly on the simple molecular descriptor dataset. Examining the data, it clustered 2039 molecules to class 0 and clustered just 1 molecule to class 1.
However, promising results were achieved using the Morgan fingerprint dataset.

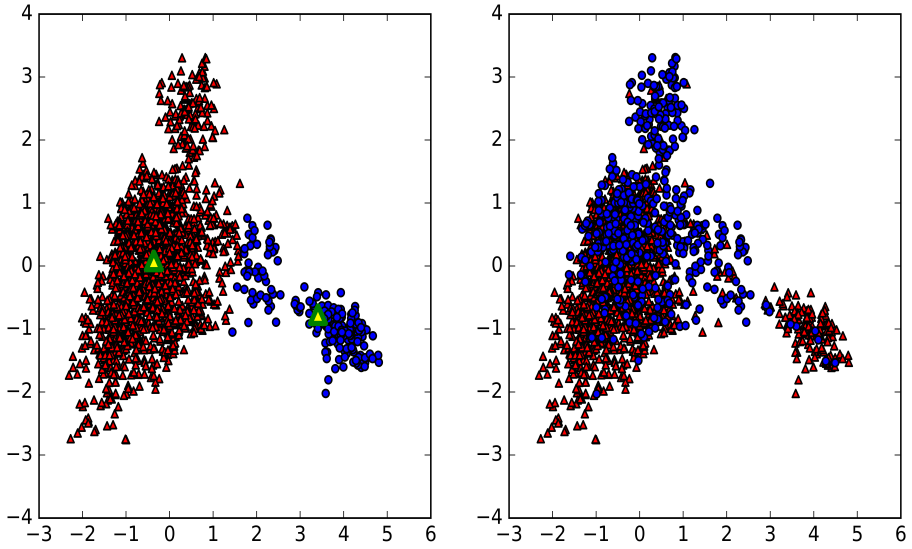

Figure 7.8: 2D Scatter Plot found by k-means on fingerprint dataset. *Original data points on the right and clusters found on the left*

The plot looks linearly separable now, an accuracy of 84% was achieved classify the red points (molecules that pass through the barrier (A)) and a 59% accuracy of classifying the blue points.The yellow triangles in the diagram represent the centers of each clusters.

Agglomerative Clustering For the Simple Molecular dataset, the entire dataset was scaled to have a minimum of -1 and a maximum of 1. Then the 2 principal components were selected, a ward agglomerative cluster algorithm was applied and it created 2 clusters with cluster 1 having 1442 samples and cluster 2 having 598 samples. The original dataset had 1563 molecules passing through the barrier and 477 molecules not passing through, most likely cluster 1 represents the molecules that pass through

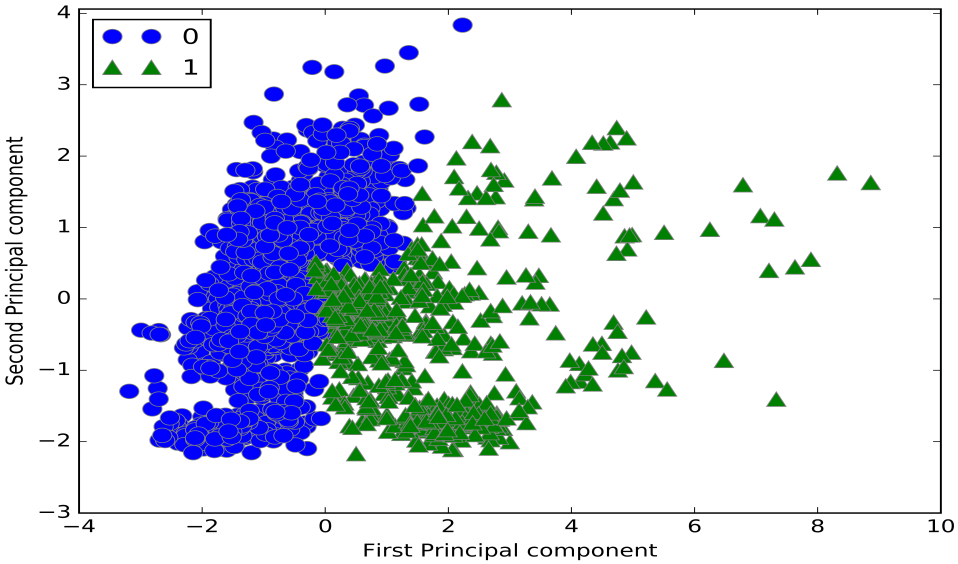

Figure 7.9: 2D Scatter Plot of clusters found by the agglomerative ward algorithm on the simple molecular dataset

Applying the same cluster to the Morgan fingerprint dataset, without preprocessing the data, it created 2 clusters; cluster 1 with 1894 samples and cluster 2 with 146 samples. The simple molecular descriptor dataset proves to be a better dataset for this clustering algorithm as it achieves a lower error rate

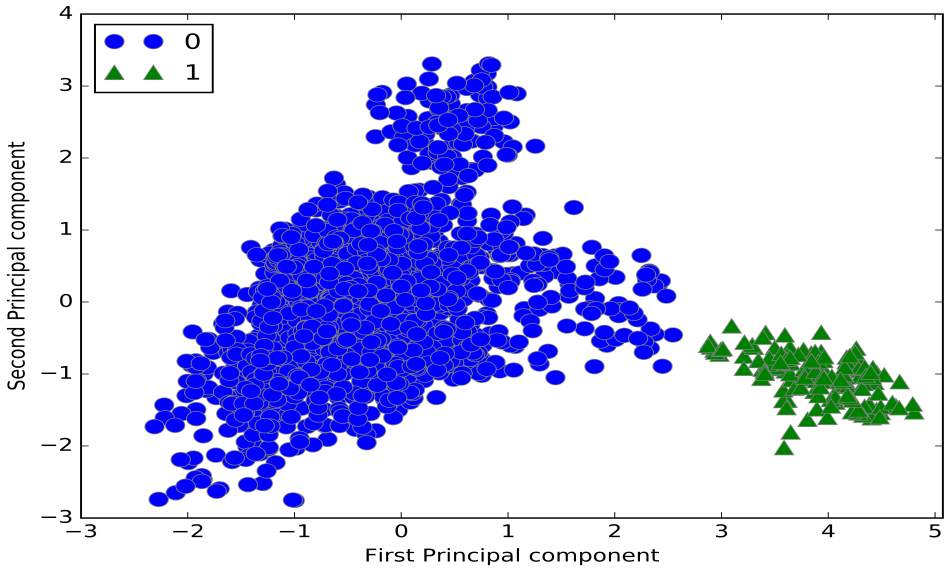

Figure 7.10: 2D Scatter Plot of clusters found by the agglomerative ward algorithm on the Morgan Fingerprint dataset

7.2 Machine Learning Classifiers/Models

A Detailed explanation of the selected classifiers based on literature review

7.2.1 Decision Tree

Decision trees are an example of non-parametric supervised learning algorithms. After applying the decision tree classifier, an accuracy of $\sim$ 85% was achieved. Examining the internal workings of the Tree, based on the simple molecular descriptor dataset, we can see that the total polar surface area (TPSA) is the defining factor in determining whether a molecule crosses the blood-brain barrier.

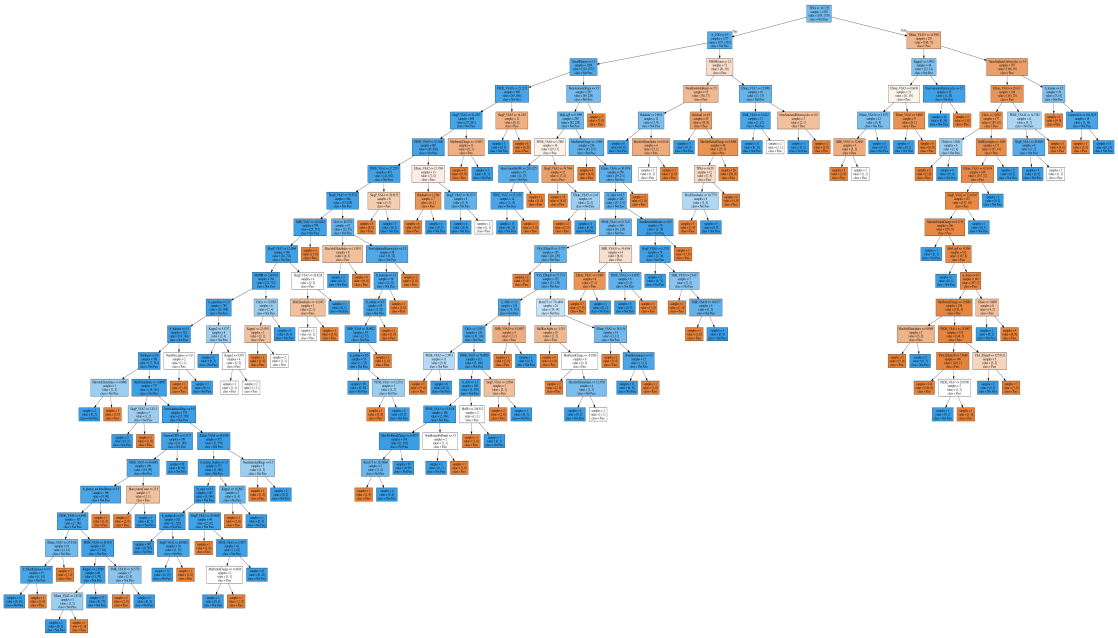

Figure 7.11: Decision Tree: Simple Molecular Descriptors

Random Forests

An accuracy of $\sim$ 87% was achieved using the random forest classifier on the simple molecular descriptor dataset whilst an accuracy of $\sim$ 89% was achieved on the Morgan fingerprint dataset. The number of trees in the random forest was set to 20.

Extra Trees

The extra trees tend to perform slightly better than the random forests. On the simple molecular dataset, the extra trees score an accuracy of $\sim$ 90% and on the Morgan fingerprint dataset, the extra trees score an accuracy of $\sim$ 89%, performing just slightly less than the random forest on the same dataset. The number of extra trees was set to 20

AdaBoost Decision Trees

Applying the AdaBoost algorithm to the decision trees, an accuracy of $\sim$ 83% was achieved on the simple molecular descriptor dataset, and an accuracy of $\sim$ 81% on the Morgan fingerprint dataset. Both rounds of classification with number of trees = 200. Strangely, the Adaboosted decision trees ought to provide better results than the random forest or extra trees classifiers.

Gradient Tree Boosting

With a little bit of tuning the parameters of the gradient boosted classifier, by increasing the number of estimators to around 150 and reducing the learning rate to 0.45. An accuracy score of $\sim 86\%$ was achieved on the simple molecular descriptors and $\sim 82\%$ on the Morgan fingerprint dataset.

7.2.2 K-Nearest Neighbours (kNN)

The kNN classifier, at $k = 10$ has an accuracy of roughly around $\sim 85\%$ using only the simple molecular descriptors. By engineering the feature vector to have unit variance and a Gaussian mean of 0, we then have an accuracy of around $\sim 89\%$ at 10 neighbours.

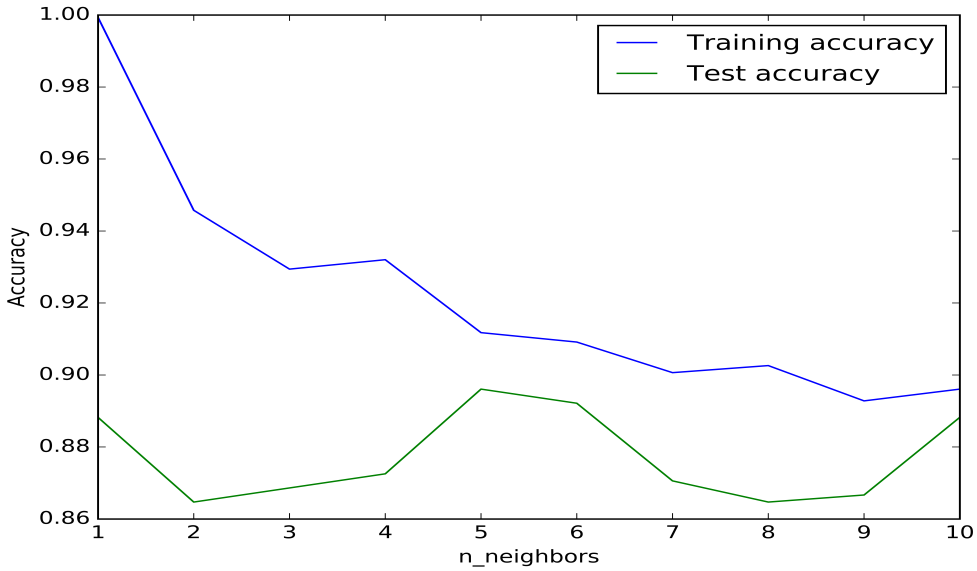

Figure 7.12: kNN classifier for 10 neighbours using the Simple Molecular Descriptors

Also using the Morgan fingerprint has the data for the kNN classifier, we have an accuracy of around $\sim 89\%$. Without scaling the simple molecular descriptors, the Morgan fingerprint dataset provides better results.

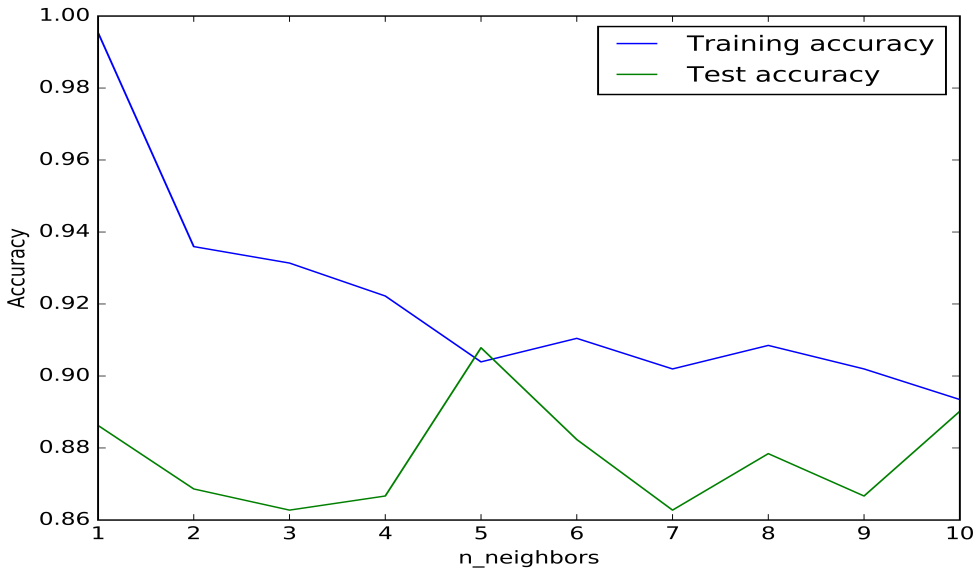

Figure 7.13: kNN classifier for 10 neighbours using the Morgan Fingerprint dataset

7.2.3 Support Vector Machines

Scaling the data using a Standard Scaler and selecting the top 100 principal component, a mean accuracy of 84.9% was achieved using both datasets

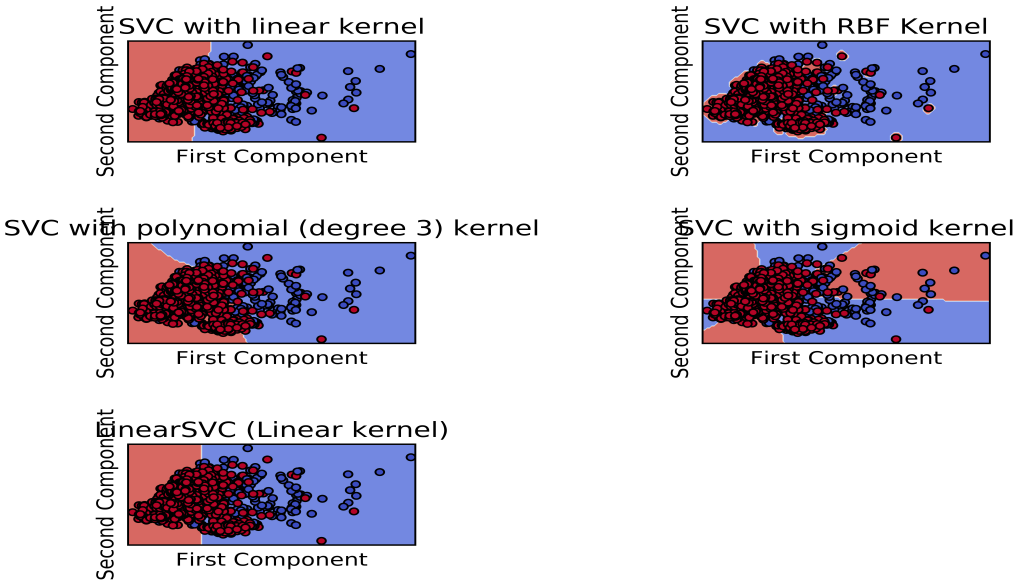

Figure 7.14: Support Vector Machines with different kernel functions on Simple Molecular Dataset

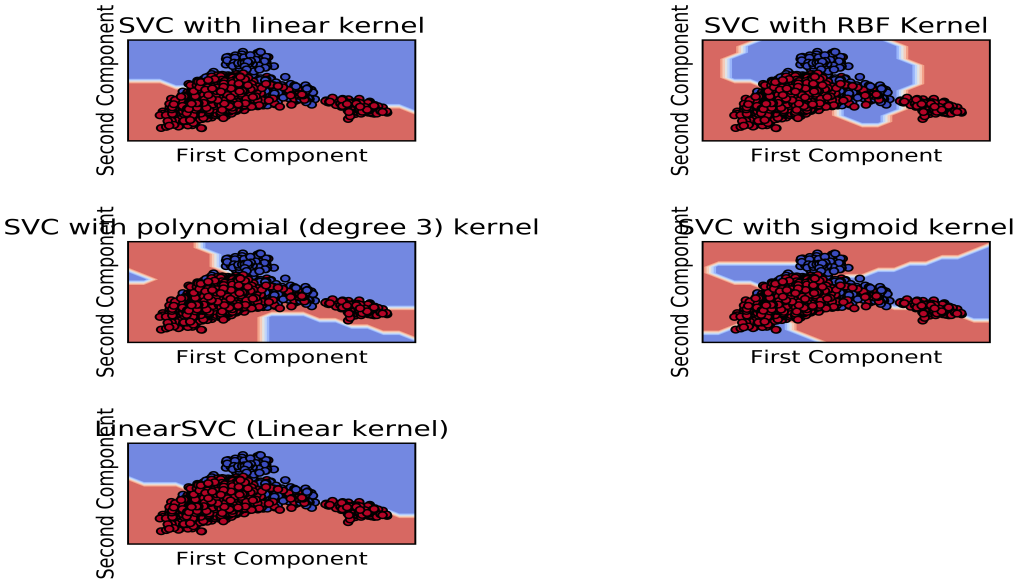

Figure 7.15: Support Vector Machines with different kernel functions on Morgan Fingerprint Dataset

7.2.4 Neural Networks

The Human brain is a very fascinating and robust system, the basic processing unit of the brain is the nerve cell or neuron and there are around 100 billion neurons in the brain all connected together by synapses. These neurons are electrically excitable, they pass signals from one neuron to the next by electrical and chemical signals when the membrane potential reaches some threshold, which can be inhibited if the threshold is not exceeded. A Neural Network is a computational model inspired by the human brain, it consists of a network of neurons connected together by axons. The inspiration behind the model is that if learning occurs in the brain and if we can represent this in a computational model then we can achieve some form of intelligence in our systems.

Hebbian Learning This is a learning concept of the brain based on the mechanism neural plasticity proposed by Donald Hebb. The rule formulated states that changes in the strength of synaptic connections are proportional to the correlation in the firing of the two connecting neurons, the strength of the connection is proportional to the frequency of excitement between them. However, the connection between two neurons that never fire will eventually be broken.
The idea of repetition enhances retention can also be applied here, where actions that are repeated continuously will create retention in the brain [6]. An example is in the experiment carried out by Ivan Pavlov, where his dogs were conditioned to associate the ringing of a bell to the presence of food. As a result, the neurons responsible for salivating over food and for hearing the bell would have a very strong connection to the extent that merely hearing the bell was sufficient to cause the salivating neurons to fire.

Multi Layer Perceptron

The Perceptron can also be referred to a collection of neurons along with a vector of inputs and a vector of weights to fasten the inputs to the neuron. This neuron in this network consists of its inputs, weights, threshold and activation function. Each neuron,receives an input along with the strength of the input i.e weight, which it multiplies together and if the value is greater than the threshold, the neuron fires.
For our dataset the input $X = [x_1, x_2, x_3, ..., x_n]$ to a Perceptron could be either a simple molecular feature vector value e.g the number of hydrogen atoms in the molecule or vectors of 1s and 0s if we're using the Morgan fingerprint dataset to train the network. Each neuron is also given a vector of small random positive

and negative values as its weights. The neuron calculates its activation value as

$$h = \sum_{i=1}^{m} w_i x_i \tag{7.1}$$

This activation value is then passed on to the activation function $f(h)$ which fires if the activation value is greater than the threshold. A threshold value of 0 was used irrespective of the training dataset. The output value of the function is given by

$$output = f(h) = \begin{cases} 1 & \text{if h}>\theta \\ 0 & \text{if h} \leq \theta \end{cases} \tag{7.2}$$

Multi Layer Perceptron The Input layer to the neuron is the same as the length of the feature vector plus one (the fixed bias input). For the simple molecular feature dataset, the number of inputs n to the network is 196, as that is the number of simple chemical descriptors that were calculated. For the Morgan Fingerprint dataset, $n = 2048$ where n represents the length of the bit vector.
The bias input is used as a control value to adjust the value that the neuron fires at. An example is when the feature vector to a neuron is all zeros and the neuron should fire. By our activation function, the neuron wouldn't fire because its activation value does not exceed the threshold. By adding a fixed value (bias) of ± 1 to the feature vector, we can then adjust the weight to make the neuron fire or not.

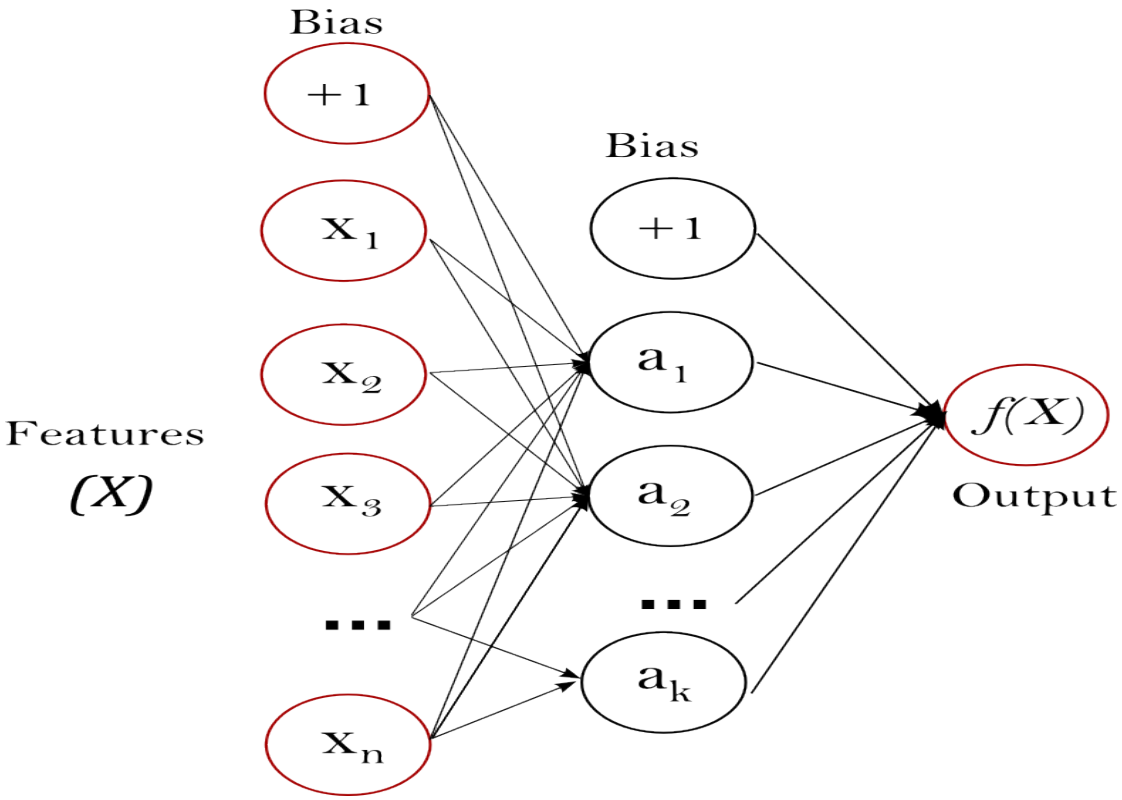

Figure 7.16: Multi Layer Perceptron [15]

Application to the Blood Brain Barrier prediction

We are undertaking a form of supervised learning in the BBB prediction. The dataset we have compiled is labelled with the correct class and this data is fed to the neural network in order for the it to generalise what it has learned about our data. The network will then find patterns in the molecules that cross the blood-brain barrier and predict if an unknown molecule will cross it or not.
Let $X = [x_1, x_2, x_3, ..., x_n]$ be the input vector to our network for n features and $W = [w_1, w_2, w_3, ..., w_n]$, the weights. Each neuron calculates whether it should fire or not based on its activation vector. We then examine the wrong neurons i.e the neurons that fired when they should have not. For each wrong neuron i, we calculate the difference in weights for it $\Delta w_i = -(y_i - t_i) * x_i$ where y_i is the output and t_i is the target output, the weight for neuron i is then updated as $w_i = w_i + \Delta w_i \eta$, where η is the learning rate (usually around $0.1 < \eta < 0.4$) which determines how much to change the weights by and as a result, determines how fast the network learns. The training is run again till the algorithm converges and the network gets all of its input right.

The Learning Algorithm [12]

- **Initialisation**
  n small (positive and negative) random numbers are assigned as weights w_{ij} where i is the number of weights and j represents the neuron being examined.

- **Training** for T iterations or until all the outputs are correct

  - For each input vector $X = [x_1, x_2, x_3, ..., x_n]$
    * Compute the activation value (h) of each neuron j:

$$h = \sum_{i=0}^{n} w_{ij} x_i \tag{7.3}$$

    * Calculate the activation (y) of neuron j using the activation function g:

$$y_j = g(h) = \begin{cases} 1 & \text{if h}>\theta \\ 0 & \text{if h} \leq \theta \end{cases} \tag{7.4}$$

    * Update the weights of neuron j using the formula:

$$\Delta w_{ij} = -\eta(y_j - t_j) * x_i \tag{7.5}$$
$$w_{ij} = \Delta w_i j + w_i j \tag{7.6}$$

- **Recall or Prediction**
  To predict the value of new molecule with feature vector X using:

$$y_j = g(\sum_{i=0}^{m} w_{ij} x_i) = \begin{cases} 1 & \text{if } w_{ij} x_i > 0 \\ 0 & \text{if } w_{ij} x_i \leq 0 \end{cases} \tag{7.7}$$

  where 1 represents a molecule that passes through the barrier and 0 a molecule that doesn't.

Model Training For both datasets (simple molecular descriptors and fingerprint), they were rescaled to have a minimum of -1 and a Maximum of 1. The transformed data was then trained on a neural network and an accuracy of $\sim 86\%$ was achieved.

Model Selection on the datasets for training The following model selection techniques were applied to both datasets

- Univariate Model Selection: Here for each feature, we compute whether there is a statistical relationship between the feature and the target. The features which are calculated to be highly unrelated to the target are then dropped from the dataset. Applying this model selection to the molecular descriptor dataset and using a percentile of 50 gives us the features in figure 7.17. The

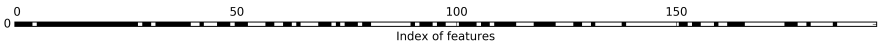

Figure 7.17: Matrix diagram of selected features

neural network was then trained with the new feature vector and a percentage increase in accuracy was achieved - with a new accuracy of $\sim 87\%$ on the molecular descriptor dataset and an accuracy of $\sim 88\%$ on the fingerprint dataset.

7.2.5 Classifier Performance Comparison

Table 7.1: Classifier average performance on different datasets

Classifier	Simple Molecular descriptors(%)	Morgan Fingerprint(%)
Decision Tree	80	79
Extra Trees	86	84
Random Forests	85	84
AdaBoost Decision Trees	84	82
Gradient Boosted Trees	86	82
k-Nearest Neighbours	87	89
Support Vector Machines	84.9	84.9
Neural Networks	87	88

7.3 Model Evaluation and Improvement

Bibliography

[1] Sarah Guido Andreas C. Müller. *Introduction to Machine Learning with Python: A Guide for Data scientists*. O'Reilly - O'Reilly Media, 1st edition, 2016.

[2] Jrgen Bajorath. *Chemoinformatics: Concepts, Methods and Tools for Drug Discovery*. Humana Press, Totowa, New Jersey, 2004.

[3] Dr A.N Boa. Introduction to drug discovery. http://www.hull.ac.uk/php/chsanb/DrugDisc/Introduction%20to%20Drug%20Discovery%202012.pdf. Accessed: 2016-11-28.

[4] F.K. Brown. Chemoinformatics: What is it and how does it impact drug discovery. *Annual Reports in Medicinal Chemistry*, 33:375–384, 1998.

[5] Chemical abstracts service home page CAS. Cas registry - the gold standard for chemical substance information, 2016, accessed November 7, 2016. http://www.cas.org/content/chemical-substances.

[6] Williams DM Cunningham TF, Healy AF. Effects of repetition on short-term retention of order information. *Journal of Experimental Psychology: Learning, Memory and Cognition.*, 10.

[7] Petra Fey Takashi Gojobori Linda Hannick Winston Hide David P. Hill Renate Kania Mary Schaeffer Susan St Pierre Simon Twigger Owen White Doug Howe, Maria Costanzo and Seung Yon Rhee1. Big data: The future of biocuration. *PubMed Central (PMC)*, 455.

[8] Drugs.com. Caffeine/ergotamine: Indications, side effects, warnings - drugs.com. https://www.drugs.com/cdi/caffeine-ergotamine.html. Accessed: 2016-12-11.

[9] Luis Pinheiro Ines Filipa Martins, Ana L. Teixeira and Andre O. Falcao. A bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of Chemical Information and Modelling*, 52:16861697, 2012.

[10] M. A. Johnson and G. M. Maggiora. Concepts and applications of molecular similarity. 1990.

[11] Antonio Lavecchia. Machine learning approaches in drug discovery: methods and applications. *Drug Discovery Today*, 20:318–331, 2015.

[12] Stephen Marshall. *Machine Learning: An Algorithmic Perspective*. Chapman and Hall/CRC, 2nd edition, 2014.

[13] OpenBabel, 2017, accessed March 8, 2017. https://openbabel.org/wiki/SMARTS.

[14] Hassan Pajouhesh and George R. Lenz. Medicinal chemical properties of successful central nervous system drugs. *The American Society for Experimental NeuroTherapeutics*, 2:541–553, 2005.

[15] Scikit-Learn, 2016, accessed November 7, 2016. http://scikit-learn.org/.

[16] Ana L. Teixeira. *Machine learning methods for quantitative structure-property relationship modelling*. PhD thesis, Universidade De Lisboa, Faculdade De Cincias, 2014.

[17] Gnter Klambauer Marvin Steijaert Jrg Wegner Hugo Ceulemans Sepp Hochreiter Thomas Unterthiner, Andreas Mayr. Deep learning as an opportunity in virtual screening. *Conference, Neural Information Processing Systems Foundation*, 2014.