A Comparative Study of Machine Learning Algorithms for Leishmanial Activity Prediction based on Molecular Fingerprints.

A Thesis

submitted to

Indian Institute of Science Education and Research Pune in partial fulfilment of the requirements for the BS-MS Dual Degree Programme

by

Pallavi Kiratkar

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

Dec 2023

Supervisor: Dr. Saif Nalband

Certificate

This is to certify Pallavi Nana Kiratkar that this dissertation entitled A Comparative Study of Machine Learning Algorithms for Leishmanial Activity Prediction based on Molecular Fingerprints towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study carried out by Pallavi Kiratkar at Thapar Institute of Engineering and Technology under the supervision of Dr. Saif Nalband, Assistant Professor, Department of Computer Science and Engineering, during the academic year 2018-2023.

Saif Nalband

Dr.Saif Nalband

Committee:

Dr.Saif Nalband

Dr.Krishanpal Karmodiya

This thesis is dedicated to my Parents.

Declaration

I hereby declare that the matter embodied in the report entitled A Comparative Study of Machine Learning Algorithms for Leishmanial Activity Prediction based on Molecular Fingerprints Here are the results of the work carried out by me at the Department of (Computer Science), Thapar Institute of Engineering and Technology, Patiala, under the supervision of Dr.Saif Nalband and the same has not been submitted elsewhere for any other degree.

PKiratkar

Pallavi Kiratkar
Date: 31/10/2023

Contents

List of Tables

List of Figures

Abstract

Leishmania, categorized as a neglected tropical ailment, is instigated by a protozoan belonging to the leishmania genus and is transmitted through sandflies. This disease imposes a significant global health burden, particularly in regions with limited access to healthcare facilities. The parasite undergoes two distinct stages of development: amastigote and promastigote, each playing pivotal roles in the infection process. Various species of sandflies, including Sergentomyia and Phlebotomus, serve as vectors for disease transmission. This study addresses the challenges encountered in drug discovery for leishmaniasis, emphasizing the critical need for effective and safe treatments. Presently available therapeutics exhibit limitations, including adverse side effects and the emergence of drug-resistant strains. Moreover, the pharmaceutical industry's market-driven approach has led to a dearth of innovations for neglected tropical diseases like leishmaniasis.

To confront this issue, we propose an innovative approach combining machine learning with cheminformatics to classify drugs as either active or inactive against leishmania promastigote. The study leverages a dataset comprising 65,057 molecules sourced from the PubChem database, employing the Alamar Blue-based assay to assess their susceptibility to various drugs. Molecular fingerprints, derived from Simplified Molecular Input Line Entry System (SMILES) notations, are employed for data encoding. Three distinct types of fingerprints, namely Avalon Fingerprint, MACCS Key Fingerprint, and Pharmacophore Fingerprint, are utilized to train machine learning models. These models aim to accurately categorize molecules according to their characteristics and chemical structure, potentially revolutionizing the approach to drug discovery for leishmaniasis. The study's significance lies in its potential to expedite the drug discovery process, address the global impact of leishmaniasis, and serve as a model for tackling other neglected tropical diseases.

Acknowledgments

I would like to express my deepest gratitude to those who supported me throughout this thesis. Dr. Saif Nalband, my esteemed supervisor, provided invaluable guidance and unwavering support, shaping the direction of this research. Dr. Krishnapal Karmodiya's profound understanding enriched this thesis significantly. Shwetang and Vaishnavi, your contributions were invaluable, and I'm grateful for your assistance. ChatGPT, the AI platform by OpenAI, played a crucial role in refining my ideas. To my parents and sister, Mayuri, your love, encouragement, and belief in my abilities have been the cornerstone of my academic journey. I owe you a debt of gratitude that words cannot express."

Contributions

Contributor name	**Contributor role**
Pallavi Kiratkar and Dr.Nalband	Conceptualization Ideas
Pallavi Kiratkar and Dr.Nalband	Methodology
Pallavi Kiratkar	Software
Dr.Nalband	Validation
Pallavi Kiratkar and Dr.Nalband	Formal analysis
Pallavi Kiratkar and Dr.Nalband	Investigation
-	Resources
Pallavi Kiratkar	Data Curation
Pallavi Kiratkar	Writing - original draft preparation
Dr.Nalband	Writing - review and editing
Pallavi Kiratkar and Dr.Nalband	Visualization
Dr. Nalband	Supervision
Dr. Nalband and Dr.Krishanpal Karmodiya	Project administration
-	Funding acquisition

This contributor syntax is based on the Journal of Cell Science CRediT Taxonomy[1].

[1] https://journals.biologists.com/jcs/pages/author-contributions

Chapter 1 Introduction

1.1 Leishmaniasis

Leishmaniasis is a neglected tropical disease. The Trypanosomatidae family of protozoa, notably the genus Leishmania, is responsible for this disease. It develops into two stages: promastigote and amastigote. Promastigote are an extracellular form that adheres to the microvilli of insects, whereas amastigotes mostly infect the lysosomal vacuoles of phagocytic cells. Various species of sandflies serve as vectors for disease transmission. Among them, Sergentomyia and Phlebotomus are the most common vectors responsible for transmitting the Old World disease. (Steverding, 2017)

Adult sandflies, much smaller in size than even small mosquitoes, are prone to dehydration, thus they thrive in moist climatic conditions. This explains the distribution of the disease in areas with high humidity. Both male and female sandflies feed on plant juices for carbohydrates, but only female sandfly stand in need of a blood meal. It is this process during which the protozoa are transmitted to the host.

The disease is endemic to regions in Asia, South and Central America, Northern Africa, the Middle East, and the Mediterranean(Alvar *et al.*, 2012) .As only female sandflies require a blood meal, it is during this feeding that the parasites of Leishmania are either passed on to the host or acquired by the fly. The fly uses specialized mouthparts to create a small wound in the skin and then draws blood from injured capillaries. This is how promastigotes enter the foregut of the sandfly and begin replication.(Handman and Bullen, 2002; Rogers *et al.*, 2002; Torres-Guerrero *et al.*, 2017) When the sandfly feeds on another host, which can include canines, humans, marsupials, or rodents, it transmits the disease.(Steverding, 2017) Once the protozoa gain access to the host, they enter the phagolysosomes. Depending on the subtype of cell infected, cutaneous or visceral leishmaniasis can occur.

In cutaneous leishmaniasis, the parasite infects resident macrophages in the skin and begins replicating. Once each compromised cell is filled with amastigotes, it ruptures, releasing them and thus infecting neighbouring macrophages. Visceral Leishmaniasis is caused by amstigotes that transfer hematogenously to the mononuclear cells in the spleen, liver, bone marrow, and intestinal lymph node(Steverding, 2017)

1.2 Motivation

1.2.1The global impact of leishmaniasis:

Leishmaniasis, according to a study, ranks as the ninth most significant global health burden(Alvar *et al.*, 2012). This disease is prevalent in 98 countries and three territories spanning five continents. Government data, "indicates an annual occurrence of over 58,000 cases of visceral leishmaniasis and 220,000 instances of cutaneous leishmaniasis. Estimates propose that there are approximately 0.7 to 1.2 million cases of cutaneous leishmaniasis and 0.2 to 0.4 million cases of visceral leishmaniasis reported each year"(Alvar *et al.*, 2012). A mere six countries contribute to more than 90% of all global instances of visceral leishmaniasis.: Bangladesh, India, Brazil, Sudan, and South Sudan. As per the report, the mortality rate for visceral leishmaniasis in Brazil was 7.2% in 2006, whereas in India it was 1.5% from 2004 to 2008. In Nepal, the rate was 6.2%(2004-2008),and in Bangladesh, it was 2.4%(2004-2008)(Alvar *et al.*, 2012).

1.2.2 The Challenges in Drug Discovery for Leishmaniasis

Unfortunately, leishmaniasis predominantly afflicts impoverished nations. More than 90 percent of the mucocutaneous instances are concentrated in nations including Brazil, Ethiopia,Peru, and Bolivia (Boakye *et al.*, 2005; Maxfield and Crane, 2023). To date, there remains a dearth of both effective and safe treatments for leishmaniasis. Profit-driven entities, such as pharmaceutical companies, not only seek financial gain but also aim to recoup the expenses associated with drug discovery and development. Consequently, these companies have redirected their focus towards innovating drugs for diseases prevalent in high-income regions. This market-oriented approach has led to a grievous imbalance, neglecting diseases of paramount importance to developing countries. From 1975 to 2004, a total of 1556 new molecular entities received approval, with only 21 (a mere 1.3%) being developed for tuberculosis and other neglected tropical diseases(Kameda, 2014; Weng et al., 2018).

Medications such as antimonial compounds, meglumine antimonate that have been used for more than 70 years, exhibit serious adverse side effects. Additionally, their treatment exposure is extensive, leading to the rapid development of strains that are antimonial-resistant. However, alternative medications such as amphotericin B (as deoxycholate), pentamidine, and liposome based formulations are recommended. Yet, they come with severe toxicity and are expensive. The sole orally active treatment, which received approval in 2014 and was initially designed for cancer treatment, demonstrates effectiveness against infections caused by L. panamensis, L. braziliensis, and L. guyanensis (Dorlo *et al.*, 2012; Monge-Maillo and López-Vélez, 2015). However, there is limited data regarding the efficacy of miltefosine against Old World leishmaniasis ((Haldar *et al.*, 2011; Kevric *et al.*, 2015; Monge-Maillo and López-Vélez, 2015; van Griensven *et al.*, 2016))Due to the restricted availability of effective therapeutics against leishmaniasis, there is a substantial disease burden with an escalating development of resistance. Strategies aimed at addressing this issue are faltering in their efforts to introduce new drugs, leading to a widening gap in available therapeutics.

1.3 Significance of the Study.

Combining machine learning with cheminformatics for classifying drugs as active or inactive against leishmaniasis holds significant importance for several reasons:

A) This study addresses leishmaniasis, a neglected tropical disease that poses a global burden, particularly affecting impoverished and vulnerable populations.(Alvar *et al.*, 2012) It is prevalent in regions with limited access to healthcare facilities. The application of machine learning aims to streamline the research and development process of therapeutics in a economical way, revolutionizing the approach to drug discovery for leishmaniasis.

B) This approach significantly reduces the time required for the drug discovery and development process given the leishmania parasite's complex life cycle and the diverse range of disease-causing species, traditional methods of drug discovery are time-consuming and resource-intensive.(Pushpakom *et al.*, 2019) By utilizing machine learning with large datasets to analyze patterns and relationships in molecules, the process becomes faster. This accelerated pace aids in controlling the further spread of the disease and alleviating human suffering.

C) Insights gleaned from these studies serve as a testament to the potential of combining machine learning with chemoinformatic for drug discovery. (Pushpakom *et al.*, 2019)Consequently, these methods and approaches can be extrapolated to address other global health challenges posed by various diseases.

1.4 Objective:

This study aims to utilize three distinct types of molecular fingerprints to train a range of machine learning models. The objective is to differentiate molecules as either active or inactive against leishmania promastigote. The ultimate aim is to develop a reliable classifier proficient in precise categorization of molecules, relying on their chemical composition and characteristics.

- Train a machine learning model using 65,057 molecules.

- Utilize different types of fingerprints for each molecule.

- Assess the effectiveness of different machine learning models separately for each fingerprint type.

- check the performance of the best performing model on the unseen data(FDA approved drugs).

Chapter 2 Materials and Methods

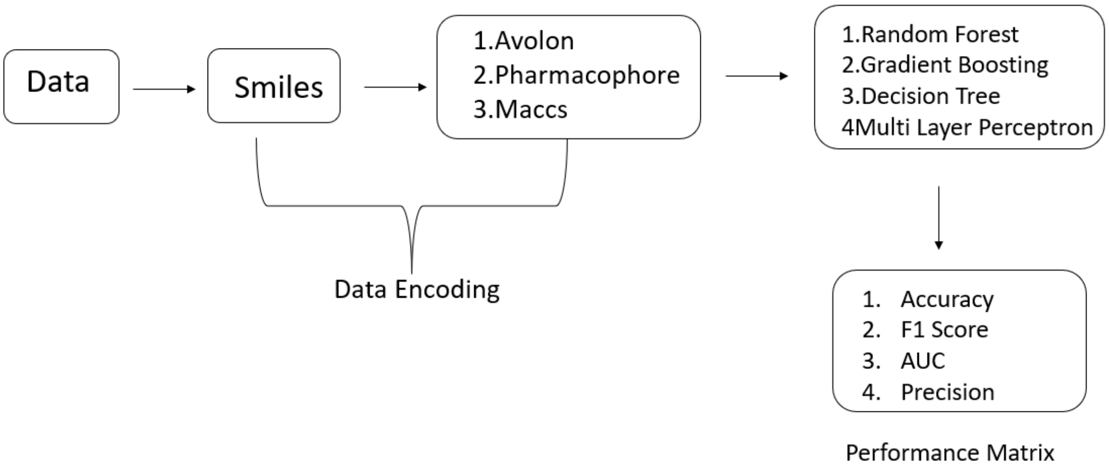


Figure 1 Methodology

2.1 Resources and tools:

The entire study was conducted within a Kaggle notebook, utilizing Python version 3.10. The RDKit library, version 2023.3.3, was employed for gathering molecular fingerprints. Additionally, data visualization was facilitated by the use of Matplotlib (version 3.7.2) and Seaborn (version 0.12.2). Fundamental libraries such as NumPy and Pandas were extensively utilized throughout the project.

2.2 Dataset:

The dataset comprises a list of 65,057 molecules sourced from the PubChem database, specifically referenced as AID 1063. This data originates from an experiment where an Alamar Blue-based assay was employed to assess the susceptibility of Leishmania parasites to various drugs. The assay yielded a binary outcome, distinguishing between "antileishmanial" (active) and "leishmanial" (inactive) states, representing the growth and viability of the Leishmania parasite. 1 denotes active compound and 0 denotes inactive compound.47427 compounds are inactive and 17630 compounds are active. Figure(1) is the graphical representation of data distribution. List of Food and Drug Administration (FDA) approved was taken from github.

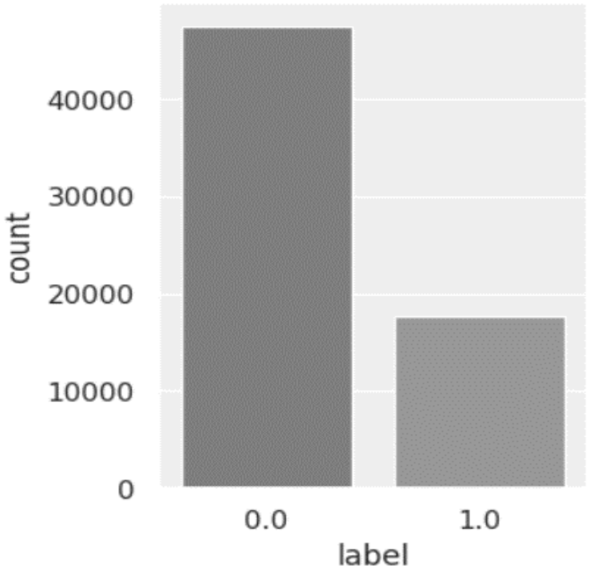


Figure 2 Graphical representation of data

2.2.1 Alamar blue-based assay:

"The Alamar Blue assay is mainly employed for studying the in vitro cytotoxicity of compounds(Fields and Lancaster, 1993; Ahmed *et al.*, 1994). This method relies on metabolism. as indicated by its name, It is based on a blue-coloured compound that is non-fluorescent known as resazurin, which functions as a fluorometric redox indicator. Once taken up by the cell, resazurin undergoes reduction in the cell's reducing environment, transforming into the fluorescent compound resorufin. This conversion process is facilitated by diaphorases present inside the cell, in conjunction with NADH or NADPH as the reductant(O'Brien *et al.*, 2000). Resorufin emits a bright red coloured fluorescence with an range of emission between 580-610nm and an excitation of range 530-570nm. The intensity of this fluorescence is measured to determine cell viability. Additionally, absorbance at 570nm, using 600nm as a reference, can also be employed for reading the test.

2.3 Data encoding

The chemical structures of the molecules were encoded using the Simplified Molecular Entry Line System (SMILES). Subsequently, these SMILES notations were transformed into three distinct types of fingerprints: Avalon Fingerprint, MACCS Key Fingerprint, and Pharmacophore Fingerprint. The dataset was initially partitioned into a training set of 70% and a temporary set of 30%. The test and validation sets were made by further splitting of a temporary set, each receiving 50% of the data. This approach ensures a comprehensive evaluation of the model's performance.

2.3.1 SMILES:

The Simplified Molecular Input Line Entry System (SMILES) is a type of a notation that enables users to represent the structure of a molecule. this format of the molecule can be interpreted by computers.(Alvar *et al.*, 2012) It adheres to five fundamental syntax rules for representing a molecule. These rules are as follows:

Simple Chains:
The simple chain structure is represented by combining bond symbols and atomic symbols. The software comprehends the potential number of connections each atom can make, considering all the elements permitted in SMILES notation. In this method, molecules are represented without explicitly including hydrogen atoms. If the bonds represented by SMILES notation do not appear to be sufficient, it is assumed that these connections are satisfied by hydrogen atoms.

For example:

CC	CH_3CH_3	Ethane
CBr	CH_3Br	Bromomethane

Atoms and Bonds:
SMILES represents atoms and bonds using their respective atomic symbols. Aromatic atoms are denoted by uppercase letters, while non-aromatic atoms are represented by lowercase letters. If the symbol of an atom contains more than one alphabetic then, the 2nd alphabetic must be in lowercase. For instance, in SMILES notation, "CC" implies that 2 non-aromatic atoms are attached by a single bond. By default, single bonds are assumed. The same convention applies for other types of bonds as well.

Following are the bond representation:

Single bond	(-)
double bond	(=)
Aromatic bond	(*)
Triple bond	(#)
Disconnected structures	(.)

Rings:

In SMILES notation, the opening and closing of a ring are indicated by numbers. If there are multiple rings, they are distinguished by representing each ring with different numbers. When a ring closure is represented by double bond,(-) the symbol of bond is placed prior to the number denoting ring closure. For instance, in the SMILES string "C1CCCC1", the first carbon is assigned the number one and is connected to the last carbon by a single bond. The last carbon is also numbered as one.

Some more examples:

C1OC1CC	Ethyloxirane
c1cc2ccccc2cc1	Napthalene.

Branches:

In a molecular chain,a branch is denoted by enclosing the SMILES symbol(s) for the branch within parentheses. within the parentheses,string is positioned immediately after the atomic to which it is attached. symbol directly follows the left parenthesis, If the branch is connected by a double or triple bond,

Example:

CC(CC)C	2 Methylbutane
c1c(N(=O)=O)cccc1	Nitrobenzene

Charged Atoms:

For charged atoms the brackets are placed after the atom .its atomic charge is shown inside the closed bracket.

For example:

CCC(=O)O{-}	propanoic acid in the ionized form.
c1ccccn{+1}1CC(=O)O	1-Carboxylmethyl pyridinium.

3.3.2 MOLECULAR FINGERPRINTS:

A molecular fingerprint is a condensed, mathematical representation of a molecule's structure. It encodes crucial structural characteristics, patterns, and properties in a format conducive to computational analysis. These fingerprints play a vital role in cheminformatics and computational chemistry, supporting tasks like similarity analysis, quantitative structure-activity relationship (QSAR) and virtual screening, modeling. They encapsulate information which tells if the specific substructure is present or absent, as well as various physicochemical properties of the molecule. In summary, a molecular fingerprint serves as a distinctive numerical pattern or bitstring, acting as a unique identifier for a molecule. This enables efficient comparison and analysis across extensive datasets of chemical compounds. Depending on the specific application, different types of fingerprints may emphasize varying aspects of a molecule's structure or properties.

Avalon Fingerprint:

Information about a molecule's characteristics and the existence or lack of particular chemical substructures is encoded in the Avalon fingerprint.. These substructures encompass both aromatic and non-aromatic rings, functional groups, and other molecular motifs. Represented as a fixed-length bit vector consisting of 0s and 1s, where each bit denotes the existence or lack of a particular substructure. Specifically, a bit is set to 1 if the corresponding substructure is present, otherwise, it is set to 0. Moreover, the Avalon fingerprint provides insights into the way atoms are arranged in a space within a molecule, including considerations of chirality. It also encapsulates details about atom connectivity, bond arrangements around atoms, bond types, and hydrogen bonding sites. This information is pivotal in comprehending the molecule's biological activity.

3D Pharmacophore Fingerprint:

The 3D pharmacophore fingerprint encodes comprehensive information regarding the nature and 3D arrangement of chemical functionalities within a ligand. This encompasses characteristics such as aromatic rings, hydrophobic areas, hydrogen bond donors, hydrogen bond acceptors, as well as positively and negatively charged ionizable groups. Specifically, when it comes to hydrogen bonds and aromatic interactions, it provides details about the directional preferences of these features. To delve into the details, let's consider 2 features of pharmacophore fingerprint denoted as A and B. By takin into account every conceivable pairing of two or three features, along with two distance bins, we can discern several scenarios.

Three pairs (BB, AA, and AB) and four triplets (BBB, AAA, AAB, and ABB) may be created by combining Features A and B. The first bin (0), which indicates two or less bonds, and the second bin (1), which indicates more than two bonds, may be used to classify the distance between these feature pairs. This suggests that two bits describe a single two-point pharmacophore. Due to the presence of three inner distances, these pharmacophores are represented by eight bits. These can fall within either the first or the second distance bin. Altogether, 38 bits are employed to represent the complete signature.

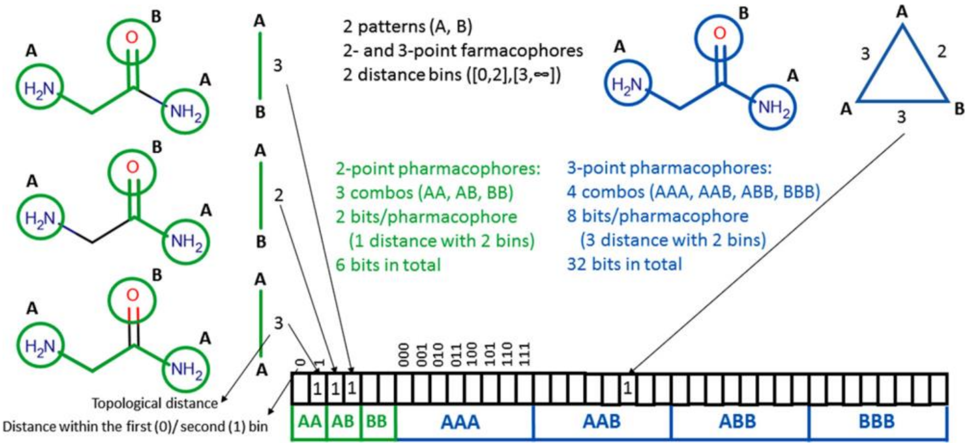
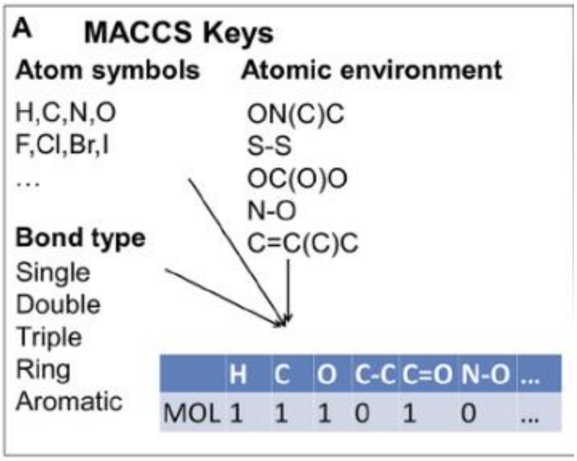


Figure 3 Pharmacophore print generation (Warszycki et al., 2021)

MACCS Fingerprint:

The Molecular ACCess System (MACCS) key fingerprints are utilized to measure molecular similarity and operate within a 2D structural framework. Each feature in the fingerprint is represented by either '1' or '0', where ‘1’ denotes the presence and ‘0’ denotes the absence of specific substructures. These fingerprints represent whether specific pre-defined structural patterns or substructures are present or not. There exist two sets of MACCS keys: one with 960 keys and the other with 166 keys. These keys account for the count of substructures, encompassing a variety of non-hydrogen atoms (Maggiora *et al.*, 2014).To calculate how similar two molecules are to one another, the Tanimoto coefficient is used. This coefficient is defined as the ratio of the number of shared "1" bits to the total number of "1" bits present in at least one of the fingerprints.



	H	C	O	C-C	C=O	N-O	...
MOL	1	1	1	0	1	0	...

Figure 4 MACCS Key representation

Tanimoto (A, B) = (A ∩ B) / (A ∪ B)

A: set of “1 ” bits in the fingerprint of a molecule A.

B: set of “1 ” bits in the fingerprint of molecule B.

A ∩ B: number of shared “1” bits.

A ∪ B: total number of “1” bits in A or B.

2.4 MACHINE LEARNING:

Our study encompassed the implementation and evaluation of four distinct machine learning algorithms. These models were meticulously trained utilizing molecular fingerprints as inputs, with the aim of effecting assigning molecules to either the active or inactive category through a binary classification process. To diversify our approach, we employed three distinct types of fingerprints: MACCS key fingerprint, Avalon fingerprint, and pharmacophore fingerprint. The algorithms selected for this study comprised Random Forest (RF), Gradient Boosting (GB), and Decision Tree (DT), each chosen for their unique characteristics and proven efficacy in similar classification tasks. Data was split into train (70%) test(15%) and validation(15%) ,followed by balancing that was done using SMOTE(Synthetic minority oversampling technique).

A comprehensive training regimen was undertaken, involving the application of each of the aforementioned algorithms to all three types of fingerprints. This rigorous approach was undertaken to ensure a robust evaluation of their performance across diverse input data. To test the model, unseen data was given to the trained Random forest model. The RF model, previously trained on a separate dataset, was applied to the FDA-approved drugs dataset. This allowed us to assess the model's predictive capability on previously unseen compounds.To quantitatively assess the models' performance, we calculated key metrics including accuracy, precision, F1 score, and Area Under the Curve (AUC) for all combinations of algorithms and fingerprints. We employed box plots, heatmaps, and bar graphs to provide clear and intuitive visualizations of the performance metrics. These visual aids not only facilitate a rapid grasp of the relative performance of different models, but also serve as valuable tools for conveying our results effectively to a wider audience.

2.4.1 Machine learning algorithms:

Random Forest:

One of the method of supervised learning is Random Forest. It is based on the idea of ensemble learning, which combines several classifiers to solve a challenging issue and enhance the model's functionality. To improve the dataset's predicting accuracy, Random Forest utilizes multiple decision trees on different parts of the input data, subsequently combining their outcomes. Based on the majority votes of these predictions, the Random Forest algorithm uses the predictions from these several decision trees to decide the final result. With the accuracy increases with the number of trees in the forest.Moreover, it effectively prevents the model from overfitting.In addition to its high accuracy and overfitting prevention capabilities, Random Forest excels in handling large datasets with high dimensions.

Working of Random Forest Algorithm:

Bagging, exemplified by techniques like Random Forest, begins with the creation of different training subsets from the sample dataset. These subsets are formed by row sampling. This selection results in what is termed Bootstrap samples, drawn from the original data. Following this, models are created independently for each bootstrap sample through raining. This step involves training each model on its respective bootstrap sample. Consequently, results are generated for each model.To determine the final output, the results from all the models are combined through a process called majority voting. This entails selecting the outcome that is most commonly predicted. This process of merging outcomes determined by a majority vote to generate the final output is known as aggregation.

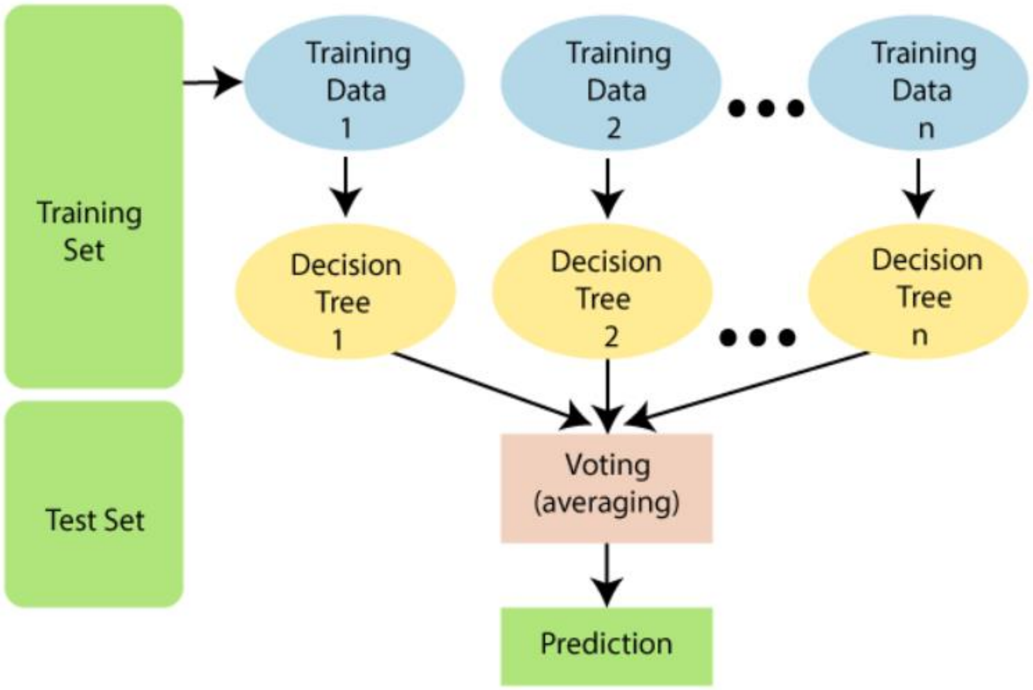


Figure 5 workflow of Random Forest classifier

DECISION TREE

One of type of supervised learning technique is the decision tree. It builds a structure like a tree, beginning with the root node and continuing into branches; this is how it gets its name. It can be applied to regression issues as well, albeit its main usage is in classification problems. Essentially, it is a tree-structured classifier. The dataset's characteristics are represented by the internal nodes of this tree, decision rules by the branches, and outcomes by each leaf node. A decision tree has two different kinds of nodes: decision nodes and leaf nodes. Decision nodes possess multiple branches and are instrumental in making decisions. Conversely, leaf nodes do not have any further branches and they hold the output of the decision. The dataset contains features that serve as the basis for the decisions or tests performed. The fundamental principle it operates on involves posing questions with binary answers (yes or no). Depending on the response, it splits the tree into subtrees.The algorithm employed for building a tree is known as the Classification and Regression Tree algorithm (CART). The root node of the Decision Tree is where the tree starts for a classification . It compares the values of the root attribute with the real dataset's equivalent attribute. It follows the branch that goes to the next node based on this comparison. This procedure keeps on until it reaches a leaf node.

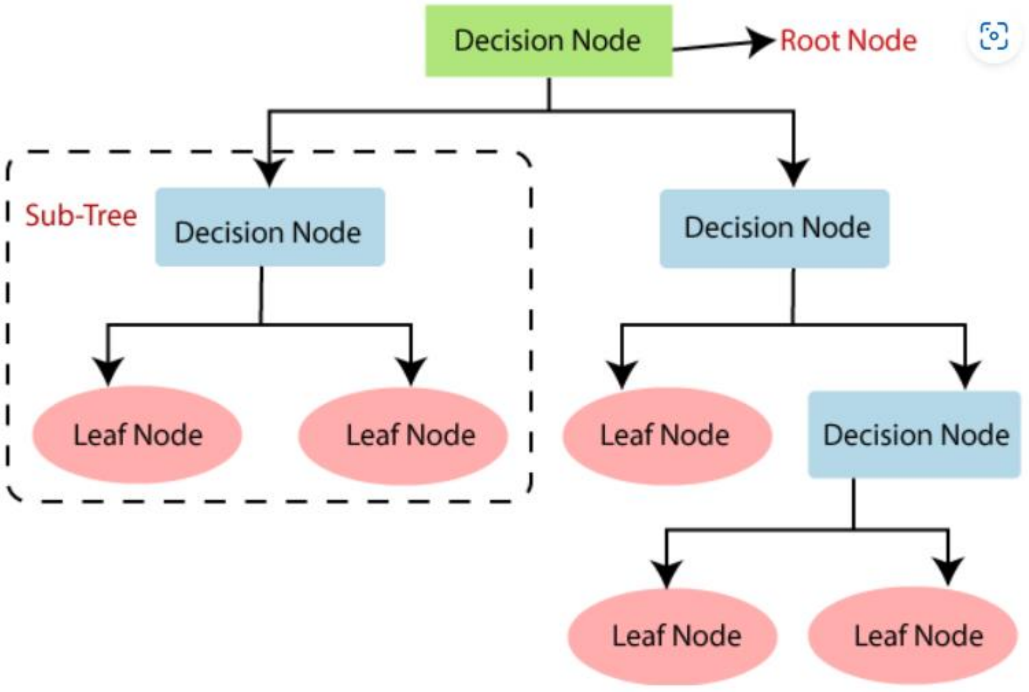


Figure 6 workflow of Decision tree classifier

To outline the steps:

- The tree starts with the root node, denoted as S, which encompasses the entire dataset.
- By employing an attribute selection measure, the best attribute in the dataset is identified.
- To obtain the potential values for this best attribute, S is divided into subsets.
- A decision tree node is then generated, containing the chosen best attribute.
- Utilizing the subset of the dataset created in step 3, a new decision tree is formed. This process iterates until a stage is reached where further classification is not possible, culminating in the final node referred to as the leaf node.

Gradient Boosting:

Gradient boosting combines several weak learners to create a strong learner, making it one of the most powerful tools in machine learning algorithms.Using gradient descent, each new model in this method is trained to minimize the loss function—such as cross entropy—of the preceding model. The approach calculates the gradient of the loss function in relation to the current ensemble predictions for every iteration.. It then proceeds to train the new model. The predictions from this new model are added to the ensembled predictions of the previous models, and the process iterated to meet the specified criteria.(Li)

Algorithm:

To understand it stepwise:

We begin with calculation of the error also called the residual(r_i) .

$r_i = y_i - F(x_i)$

y_i is the value for the data point i.and $F(x_i)$ is the prediction made by the current model for the data point i.

to fix the mistakes made by the previous model we create a new model .

$h_m (x_i)=r_i$ where $h_m (x_i)$ is the prediction of the new model for datapoint i in round m.

r_i is the residual calculated for the i data point.

We combine model by adding the predictions from the corrective models to the predictions of the previous model. The overall improved predictions is represented by

$F_{m+1(}x_i) = F_m(x_i) + h_m(x_i)$

Where, $F_{m+1(}x_i)$ is combined predictions $h_m(x_i)$,for the i datapoint in m+1 round.

$F_m(x_i)$ is the prediction of current ensemble of models for i datapoint in m round.

$h_m(x_i)$, is the prediction of the new model in m round for the datapoint i.

Multilayer Perceptron:

The input layer, output layer, and hidden layer are the three layers that make up the Multilayer Perceptron.It serves as an extension of the feedforward neural network. The input layer receives the signals, while the outer layer handles tasks such as classification and prediction.The Multilayer Perceptron's real computational engine is the arbitrary number of hidden layers positioned between the input and outer layers. Like a feedforward neural network, data moves from the input layer to the output layer. The backpropagation technique is used in the Multilayer Perceptron to train its neurons.. It is designed in a way that enables it to solve any continuous function. This architecture finds significant application in classification, prediction, recognition, and approximation (Abirami and Chitra, 2020).

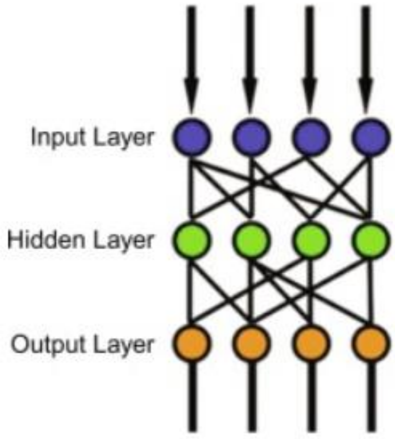


Figure 7 Working Pathway Multilayer Perceptron

The computation that takes place in the hidden layer and at every neuron at the output is as follows:

O(x) =G(b(2))+ W (2)h(x))

h(x)= Φ(x)=s(b(1) +w(1))

here b1 and b2 are are bias vectors,

W(1) and W(2) are weight matrices,

G and s are activation functions.

The set of parameters to learn is set theta which includes {W(1),W(2),b(1),b(2)}.

Choice for s includes tanh function. That is tanh(a)=(e^{a}-e^{-a}) /(e^{a}+e^{-a})

Or the sigmoid function with sigmoid(a)=1/(1+e^{-a}).

Chapter 3 Results

3.1 AVALON FINGERPRINT:

Table:1 Performance Metrics of Machine Learning Models Using Avalon fingerprint

Models Avalon	Accuracy	Precision	AUC	F1-Score
Random Forest	0.83	0.90	0.83	0.81
Gradient Boosting	0.82	0.88	0.82	0.80
Decision Tree	0.76	0.76	0.75	0.75
Multilayer Perceptron	0.69	0.76	0.71	0.69

The performance of 4 different types of machine learning models using Avalon fingerprints for predicting molecular activity is summarized in Table 1. Among the models, Random Forest exhibited the highest accuracy at 0.83, along with notable precision of 0.90, indicating a low rate of false positives. This model also demonstrated an AUC of 0.83, signifying its ability to effectively distinguish between active and inactive molecules. The F1-Score, which balances precision and recall, was measured at 0.81 for Random Forest. Gradient Boosting showed similar performance, with an accuracy of 0.82, precision of 0.88, AUC of 0.82, and F1-Score of 0.80. Decision Tree exhibited an accuracy of 0.76, along with precision and AUC values of 0.76 and 0.75, respectively. The F1-Score for Decision Tree was recorded at 0.75. Multilayer Perceptron, while demonstrating a lower accuracy of 0.69, showed a competitive precision of 0.76 and an AUC of 0.71. The F1-Score for Multilayer Perceptron was 0.69, indicating a balanced performance in terms of precision and recall.

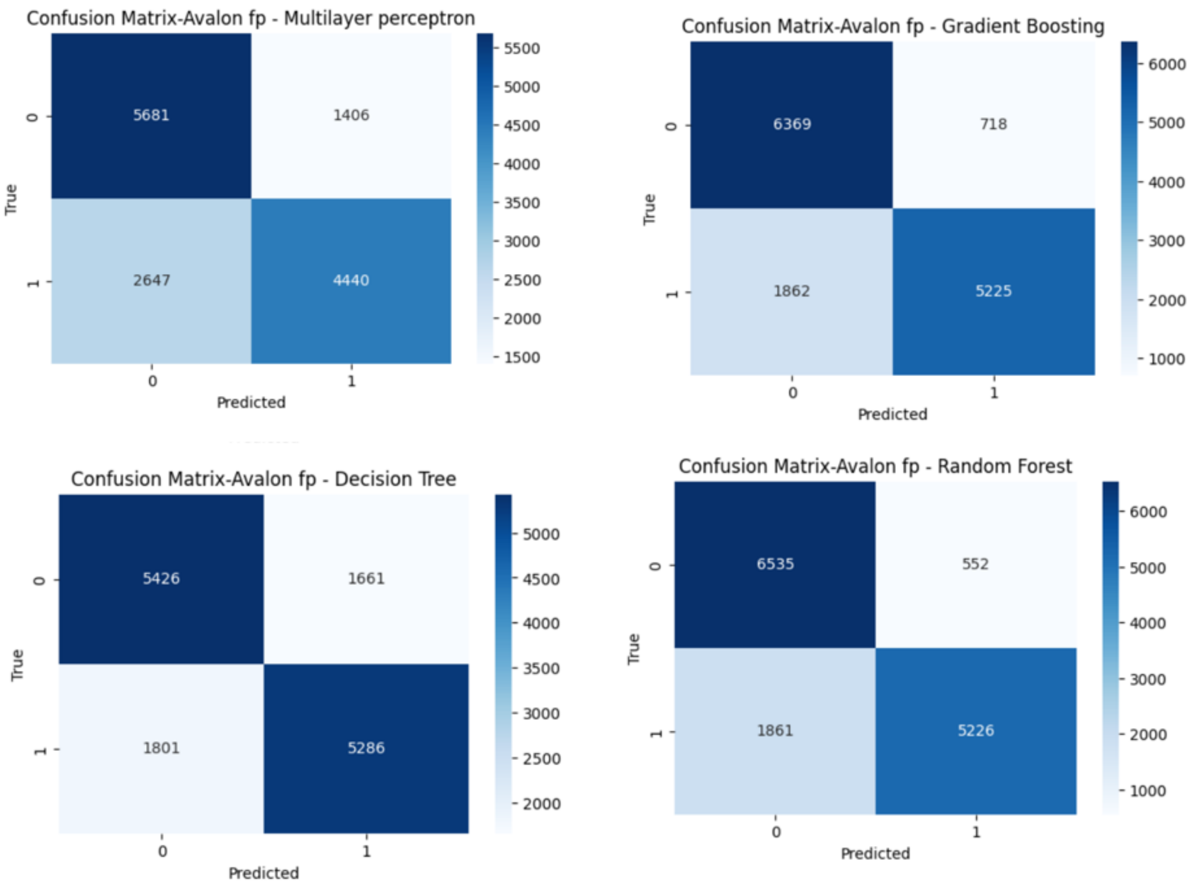


Figure 8 Confusion Matrix for machine learning models using Avalon Fingerprint

The confusion matrix in Figure 8 illustrates the performance of four machine learning models (Random Forest, Gradient Boosting, Decision Tree, and Multilayer Perceptron) applied to the Avalon fingerprint for predicting molecular activity. The matrix provides a detailed breakdown of the model's predictions, showing the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) instances. This allows us to assess the model's performance in classifying molecules as active or inactive.

PHARMACOPHORE FINGERPRINT:

Table 2: Performance Metrics of Machine Learning Models Using Pharmacophore fingerprint

Models	Accuracy	Precision	AUC	F1-Score
Random Forest	0.82	0.87	0.82	0.81
Gradient Boosting	0.80	0.82	0.80	0.79
Decision Tree	0.71	0.73	0.72	0.70
Multilayer Perceptron	0.82	0.86	0.82	0.81

The table presents the performance metrics of various machine learning models trained on the pharmacophore fingerprint for predicting molecular activity. Random Forest achieved an accuracy of 0.82, with a precision of 0.87, an AUC of 0.82, and an F1-Score of 0.81. Gradient Boosting demonstrated similar performance with an accuracy of 0.80, a precision of 0.82, an AUC of 0.80, and an F1-Score of 0.79. Decision Tree yielded an accuracy of 0.71, a precision of 0.73, an AUC of 0.72, and an F1-Score of 0.70. Multilayer Perceptron showcased competitive performance with an accuracy of 0.82, a precision of 0.86, an AUC of 0.82, and an F1-Score of 0.81. These metrics collectively provide a comprehensive assessment of the models' performance in classifying molecules based on the pharmacophore fingerprint. The high values for accuracy, precision, and AUC indicate robust predictive capabilities, demonstrating the efficacy of the pharmacophore fingerprint in this predictive modeling task.

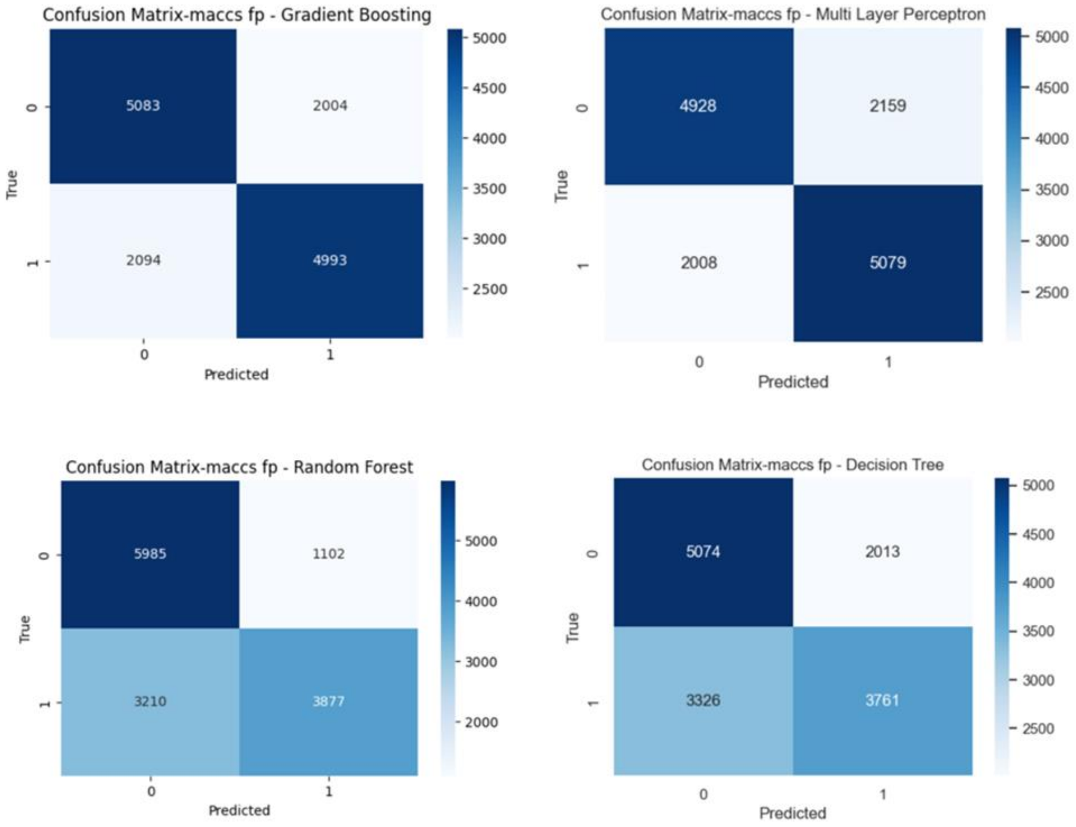


Figure 9 Confusion Matrix for machine learning models using pharmacophore fingerprint

MACCS FINGERPRINT:

Table.3 Performance Metrics of Machine Learning Models Using MACCS fingerprint

Models	Accuracy	Precision	AUC	F1-Score
Random Forest	0.70	0.78	0.70	0.64
Gradient Boosting	0.71	0.71	0.71	0.71
Decision Tree	0.62	0.65	0.62	0.58
Multilayer Perceptron	0.71	0.70	0.71	0.71

The models' performance was evaluated using a range of metrics including accuracy, precision, area under the curve (AUC), and F1-Score. Random Forest exhibited an accuracy of 0.70, a precision of 0.78, an AUC of 0.70, and an F1-Score of 0.64. Gradient Boosting demonstrated consistent performance across metrics, achieving an accuracy of 0.71, a precision of 0.71, an AUC of 0.71, and an F1-Score of 0.71. The Decision Tree model showed slightly lower performance with an accuracy of 0.62, a precision of 0.65, an AUC of 0.62, and an F1-Score of 0.58. Similarly, the Multilayer Perceptron model achieved an accuracy of 0.71, a precision of 0.70, an AUC of 0.71, and an F1-Score of 0.71. These results provide a comprehensive assessment of the models' predictive capabilities, indicating that Gradient Boosting and Multilayer Perceptron models exhibited particularly robust performance across the evaluated metrics.

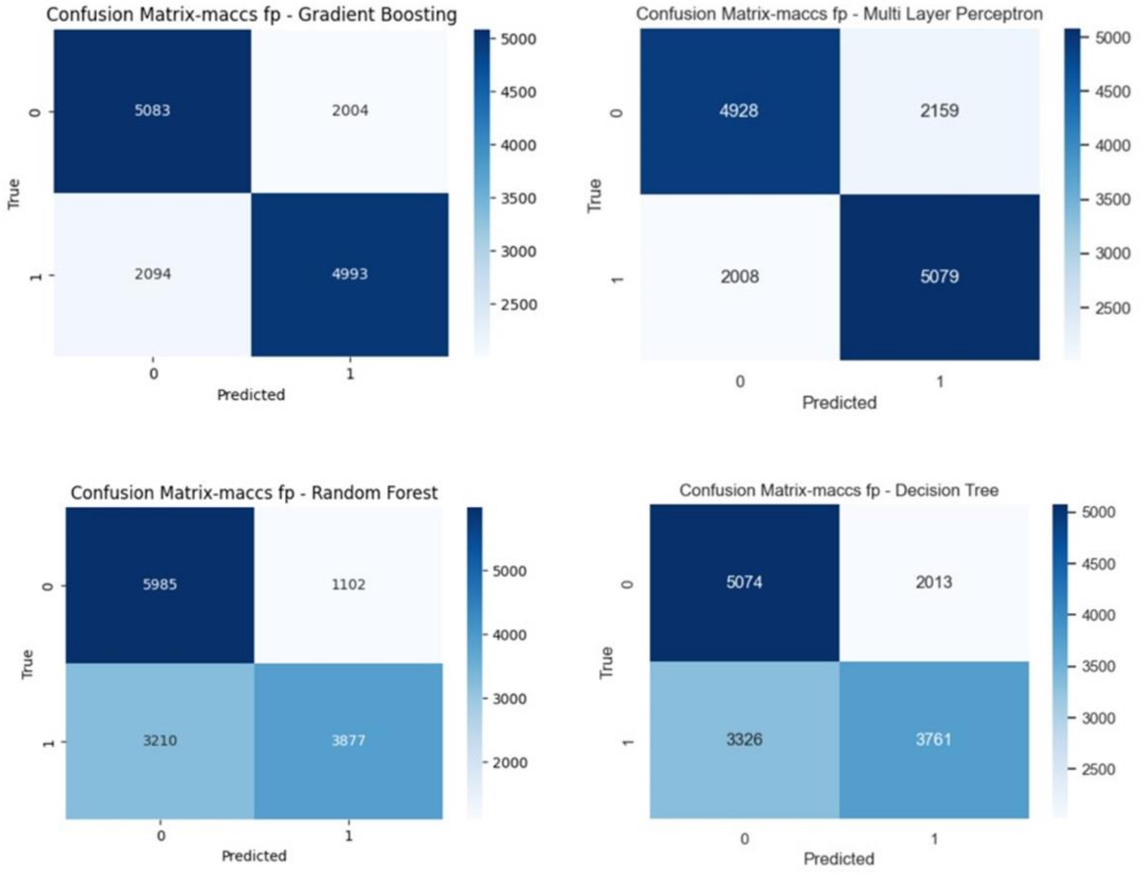


Figure 10 Confusion Matrix for machine learning models using MACCS fingerprint

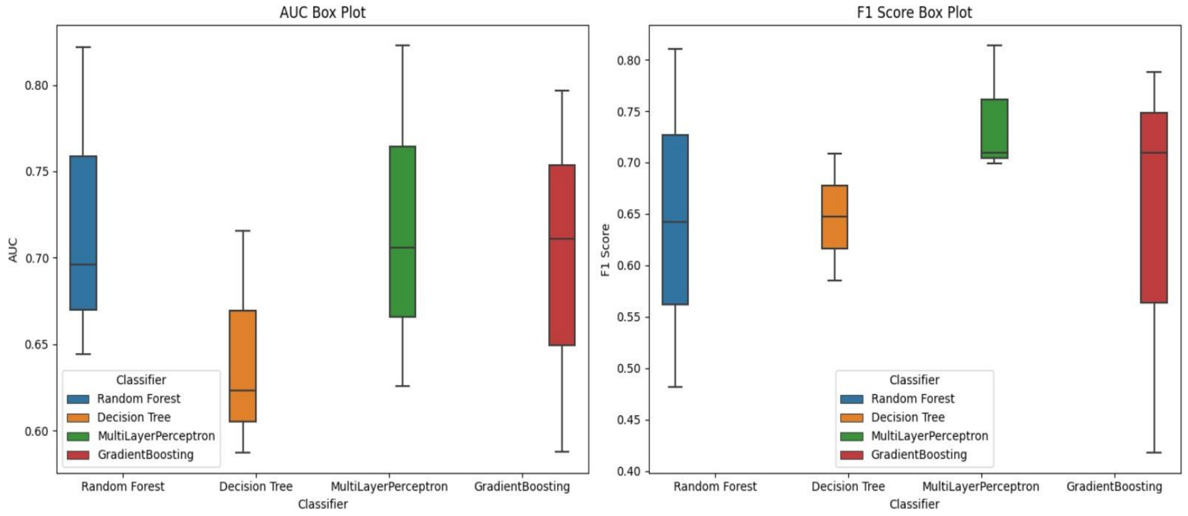


Figure 11 AUC and F1 score comparison Across Different Fingerprint Types

The box plots presented in Figure 11 illustrate the distribution of performance metrics, specifically the Area Under the Curve (AUC) and F1 Score, across different classifiers using the three types of fingerprints: MACCS, Avalon, and Pharmacophore. In the AUC box plot (Figure 11, left panel), it is observed that Random Forest consistently exhibits higher median AUC values compared to the other classifiers across all fingerprint types. Gradient Boosting also demonstrates competitive performance, particularly with Avalon and Pharmacophore fingerprints. Decision Tree shows slightly lower median AUC scores, while Multilayer Perceptron exhibits a wider spread of AUC values.

Turning to the F1 Score box plot (Figure 11, right panel), similar trends are observed. Random Forest consistently achieves higher median F1 Scores,demonstrating its efficiency in striking a balance between recall and accuracy. Gradient Boosting performs well, especially with Avalon and Pharmacophore fingerprints. Decision Tree and Multilayer Perceptron show comparable performance, with slightly lower median F1 Scores.Overall, these box plots provide valuable insights into the comparative performance of classifiers across different fingerprint types. Random Forest emerges as a robust choice, particularly when combined with the Avalon fingerprint, showcasing its potential for accurate classification of active and inactive molecules against Leishmania."

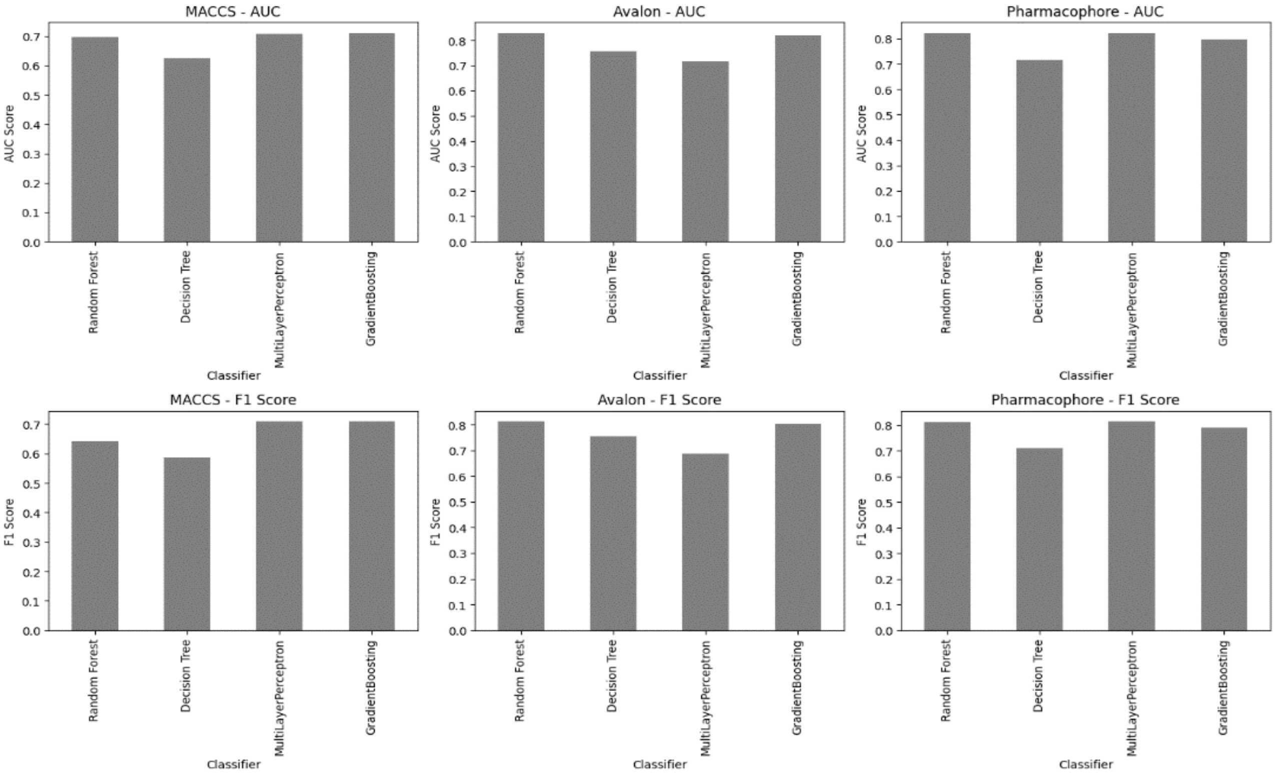


Figure 12 Classifier Performance Comparison Across Different Fingerprint Types

The subplots in Figure 12 provide a comprehensive comparison of classifier performance across three different fingerprint types

Names	Prob
Tolbutamide	0.97
Difluprednate	0.96
Halobetasol Propionate	0.96
Pyrazinamide	0.95
Lidocaine	0.93
Pentoxifylline	0.93
Podofilox	0.92
Caffeine	0.92
Clobetasol	0.92
Fluorouracil	0.92
Lisdexamfetamine	0.92
Mefenamic acid	0.92

Table 4. Predicted Probabilities of FDA Approved Drugs Against Leishmania

The table presents the predicted probabilities of FDA approved drugs potentially exhibiting activity against leishmania. The listed drugs, including Tolbutamide, Difluprednate, Halobetasol Propionate, Pyrazinamide, Lidocaine, Pentoxifylline, Podofilox, Caffeine, Clobetasol, Fluorouracil, Lisdexamfetamine, and Mefenamic acid, are ranked based on their respective probabilities, suggesting their potential suitability for repurposing efforts against leishmania.

Chapter 4 Discussion

In this study, we conducted a comparative analysis of various machine learning algorithms for predicting leishmanial activity based on molecular fingerprints. It was observed that Random Forest exhibited high accuracy across all three fingerprints (Avalon, Pharmacophore, and MACCS). The Multilayer Perceptron demonstrated accuracy close to that of Random Forest, but it did not surpass it. Both Random Forest and Multilayer Perceptron appeared to outperform Gradient Boosting and Decision Tree in terms of accuracy, precision, AUC, and F1-Score. This could be attributed to the fact that Random Forest is an ensemble method, utilizing multiple decision trees for prediction. This allows it to capture intricate relationships within the data, which is crucial for classifying molecules based on their structure. Additionally, Random Forest is less susceptible to overfitting, providing an advantage. The size and quality of the data also play vital roles in the model's performance. Random Forest tends to be more resilient to noisy and incomplete data compared to other machine learning models, which could contribute to its higher accuracy.The performance of the machine learning models is influenced by the choice of the molecular fingerprint.

The accuracy on Avalon fingerprint is the highest among all the fingerprints, with the accuracy of the pharmacophore being slightly closer to that of Avalon. Notably, the precision is highest for the Avalon fingerprint. This can be attributed to the fact that, compared to the other two fingerprints, Avalon captures a wide array of structural features such as bond types, atom environments, and information about substructures. We can infer that the structural characteristics encoded by the Avalon fingerprint are relevant for predicting activity against Leishmania. This suggests that the presence of certain motifs or the arrangement of certain motifs in a molecule is linked to its biological activity against the Leishmania promastigote. Additionally, Avalon fingerprint allows for detailed analysis of the substructures present in a molecule, encoding information about specific molecular fragments. Certain ring structures or spatial arrangements may be particularly effective in interacting with key targets in the Leishmania parasite. It may be possible to gain information on the potential mechanism of compounds that are active and inactive against Leishmania by analysing the specific structural features highlighted by Avalon fingerprints. If a certain set of substructures are consistently associated with high accuracy, it cannot be ruled out that they are involved in interactions with biological targets in the promastigote. This understanding of the association of specific structural elements with the highest activity against Leishmania allows for targeted lead optimization efforts. Further modifying or improving these key features may enhance the potency and efficiency of the compounds. This can be instrumental in developing new therapeutics. The higher precision shown by the Avalon fingerprints is a positive sign of the model's robustness, indicating a lower chance of falsely identifying inactive molecules as active. In the process of drug discovery, where false positives can lead to harmful and costly consequences, the precision metric demonstrates its importance. Given the high accuracy achieved by our trained Random Forest (RF) model, we sought to validate its performance on previously unseen data. To do so, we utilized a list of FDA-approved drugs obtained from a reputable source on GitHub.Upon subjecting this dataset to our RF model, we observed that the model accurately identified three drugs with a predicted probability of activity exceeding 0.92. This notable result underscores the robustness and reliability of our model in predicting the activity of pharmaceutical compounds.

All the findings align and support previous studies highlighting the importance of selecting appropriate molecular fingerprints for predictive modeling. This study conducted addresses the question of how different molecular fingerprints, when combined with machine learning algorithms, impact predictive modeling.The results highlight that with further research into feature engineering and in-depth analysis of the chemical features captured by each fingerprint type, we could uncover specific structural motifs that will influence the predictive task of the machine learning model. Using the knowledge of important structural motifs identified through Avalon fingerprints, further research can delve into scaffold hopping — identifying chemically distinct compounds that have similar biological activity against Leishmania, thereby increasing the scope of potential lead compounds.

As for the result in Table (4), higher probabilities indicate a stronger confidence in the prediction, they do not inherently guarantee efficacy against leishmania. Consulting domain experts in parasitology or pharmacology is of great significance to validate the model's predictions. Their specialized knowledge can provide invaluable context and assist in deciphering the practical implications of the results.

Chapter 5 References:

1. 1. Ahmed, SA, Gogal, RM, and Walsh, JE (1994). A new rapid and simple non-radioactive assay to monitor and determine the proliferation of lymphocytes: an alternative to [3H]thymidine incorporation assay. J Immunol Methods 170, 211–224.

2. Alvar, J, Vélez, ID, Bern, C, Herrero, M, Desjeux, P, Cano, J, Jannin, J, Boer, M den, and Team, the WLC (2012). Leishmaniasis Worldwide and Global Estimates of Its Incidence. PLOS ONE 7, e35671.

3. Boakye, D, Wilson, M, and Kweku, M (2005). A Review of Leishmaniasis in West Africa. Ghana Med J 39, 94–97.

4. Dorlo, TPC, Balasegaram, M, Beijnen, JH, and de Vries, PJ (2012). Miltefosine: a review of its pharmacology and therapeutic efficacy in the treatment of leishmaniasis. Journal of Antimicrobial Chemotherapy 67, 2576–2597.

5. Fields, RD, and Lancaster, MV (1993). Dual-attribute continuous monitoring of cell proliferation/cytotoxicity. Am Biotechnol Lab 11, 48–50.

6. van Griensven, J, Gadisa, E, Aseffa, A, Hailu, A, Beshah, AM, and Diro, E (2016). Treatment of Cutaneous Leishmaniasis Caused by Leishmania aethiopica: A Systematic Review. PLoS Negl Trop Dis 10, e0004495.

7. Haldar, AK, Sen, P, and Roy, S (2011). Use of antimony in the treatment of leishmaniasis: current status and future directions. Mol Biol Int 2011, 571242.

8. Handman, E, and Bullen, DVR (2002). Interaction of Leishmania with the host macrophage. Trends in Parasitology 18, 332–334.

9. Kevric, I, Cappel, MA, and Keeling, JH (2015). New World and Old World Leishmania Infections: A Practical Review. Dermatol Clin 33, 579–593.

10. Li, C A Gentle Introduction to Gradient Boosting.

11. Maggiora, G, Vogt, M, Stumpfe, D, and Bajorath, J (2014). Molecular Similarity in Medicinal Chemistry. J Med Chem 57, 3186–3204.

12. Maxfield, L, and Crane, JS (2023). Leishmaniasis. In: StatPearls, Treasure Island (FL): StatPearls Publishing.

13. Monge-Maillo, B, and López-Vélez, R (2015). Miltefosine for visceral and cutaneous leishmaniasis: drug characteristics and evidence-based treatment recommendations. Clin Infect Dis 60, 1398–1404.

14. O’Brien, J, Wilson, I, Orton, T, and Pognan, F (2000). Investigation of the Alamar Blue (resazurin) fluorescent dye for the assessment of mammalian cell cytotoxicity. Eur J Biochem 267, 5421–5426.

15. Pushpakom, S et al. (2019). Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov 18, 41–58.

16. Rogers, ME, Chance, ML, and Bates, PA (2002). The role of promastigote secretory gel in the origin and transmission of the infective stage of Leishmania mexicana by the sandfly Lutzomyia longipalpis. Parasitology 124, 495–507.

17. Steverding, D (2017). The history of leishmaniasis. Parasit Vectors 10, 82.

18. Torres-Guerrero, E, Quintanilla-Cedillo, MR, Ruiz-Esmenjaud, J, and Arenas, R (2017). Leishmaniasis: a review.

19. Warszycki, D, Struski, Ł, Śmieja, M, Kafel, R, and Kurczab, R (2021). Pharmacoprint: A Combination of a Pharmacophore Fingerprint and Artificial Intelligence as a Tool for Computer-Aided Drug Design. J Chem Inf Model 61, 5054–5065.