Summer Training Report

On

Development of a Machine Learning Model for Accelerating Drug Discovery in the Treatment of High-Priority Diseases

Submitted in partial fulfillment of the requirements for the completion of one month's summer internship/training [ART 355]

Name: Parth

Enrollment Number: 05519051922

Under the supervision of

Dr. Sushobhan Chowdhury

UNIVERSITY SCHOOL OF AUTOMATION AND ROBOTICS GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY EAST DELHI CAMPUS, SURAJMAL VIHAR, DELHI-110032

Certificate

This is to certify, that the project report submitted by Part is an outcome of his work that was done during the summer internship period.The work was original and done by the candidate. He has duly acknowledged all the sources from which the ideas and extracts have been taken.

Name. : Parth

Designation : Student(AI-DS, 2022-26)

Institution : University School of Automation and Robotics,
Guru Gobind Singh Indraprastha University, East Delhi Campus,
Vishwas Nagar Extension, Shahdara, Delhi - 110032

Title of the Report: Development of a Machine Learning Model for Accelerating Drug Discovery in the Treatment of High-Priority Diseases

Email : parthsharma23212@gmail.com

Contact No. : 8076759218

Declaration

I hereby declare that the Summer Training Report entitled (``Development of a Machine Learning Model for Accelerating Drug Discovery in the Treatment of High-Priority Diseases") is an authentic record of work completed as requirements of Summer Training (ART 355) during the period from _____ to _____ in University School of Automation and Robotics under the supervision of Dr. Sushobhan Chowdhury (Assistant Professor, USAR, GGSIPU, EDC)

(Signature of Student)

Parth

Date: _____ **05519051922**

(Signature of Supervisor)

Dr. Sushobhan Chowdhury

Assistant Professor,

USAR, GGSIPU, EDC

Date: _____

Acknowledgement

I would like to extend my sincere appreciation to Dr. Sushobhan Chowdhury for offering me the chance to work on this remarkable project. His expert guidance, encouragement, and profound insights in the field of chemistry were crucial to the success of my internship.

I am immensely grateful for his continuous support, constructive feedback, and thought-provoking discussions, which shaped the trajectory of this work. This internship has deepened my understanding of computational chemistry and the integration of machine learning into drug discovery.

My thanks also go to GGSIPU and the EDC for their invaluable resources and the supportive environment that facilitated the completion of the machine learning model aimed at predicting the activity of chemical compounds. The experience has been highly rewarding, and I look forward to utilising the skills I've acquired in future projects.
I would also like to acknowledge my fellow interns and colleagues, whose cooperation and camaraderie made this project both fruitful and enjoyable.

Thank you all for your contributions and support during this incredible journey..

Sincerely,
Parth
AIDS-B1, 05519011922

About Institution

Guru Gobind Singh Indraprastha University (GGSIPU) was established in 1998 by the Govt. of NCT of Delhi under the provisions of Guru Gobind Singh Indraprastha UniversityAct,1998.The University is recognised by University Grants Commission (UGC), India under section 12B of UGC Act. It has 14 University School of Studies, 11 of them are on the campus in Dwarka and 3 schools, USAR, USDI and USAP are on the East Delhi campus.

The university has been accredited with the prestigious NAAC A++ grade and ranked 74 by the National Institutional Ranking Framework. Recently, it has grabbed a position of 1401 in the QS World Rankings, 2024. These milestones are a testament to our excellence in higher education.

The university aims to facilitate and promote studies, research and innovation in emerging areas of education such as engineering, technology, management studies, medicine, law, arts, humanities and social sciences, design, architecture, etc. The University strives to achieve excellence in these disciplines and connected fields and other matters connected therewith or incidental thereto.

The University School of Automation \& Robotics (USAR), located at Guru Gobind Singh Indraprastha University, East Delhi Campus was established by Govt of NCT of Delhi in 2021 to meet the demands of technology in Industry 4.0. USAR offers B.Tech programs in Artificial Intelligence \& Data Science, Artificial Intelligence \& Machine Learning, Industrial Internet of Things, and Automation \& Robotics.

The curriculum at USAR equips students with specialised skills in the emerging Industry4.0 fields like Artificial Intelligence, Digital Technologies, Data Science, Cloud Computing, Machine Learning, Industrial Internet of Things, Automation, Advanced Robotics, and Smart Manufacturing. The campus fosters a culture of rigorous consistency and discipline to develop all-rounded individuals, poised to tackle the challenges of the professional workspace, and make significant contributions to the dynamic landscape of the fourth industrial revolution.

Table of Contents

Index of Tables

Index of Figures

Chapter-1: Abstract

Drug discovery is a complex and costly endeavour, involving the screening of numerous chemical compounds to identify those effective against specific diseases. In this project, machine learning techniques were applied to predict the activity of chemical compounds against Leishmaniasis(Kala aźhar) cells. Using molecular data from the ChEMBL database, several models were created to classify molecules as either active or inactive. Various molecular fingerprinting methods—Morgan, Atom Pair, RDKit, and MACCS—were utilised to represent the compounds, followed by training different machine learning algorithms, including Random Forest, Logistic Regression, Gradient Boost, KNN, Support Vector Classifier, Decision Tree, ADABoost, Artificial Neural Networks. Ensemble methods, such as Voting Classifiers, were also employed to improve prediction accuracy.

The results indicated that RDKit Fingerprinting was the most effective, particularly after feature selection enhanced computational efficiency. Among the algorithms tested, both Random Forest and Support Vector classifier combined with RDKit Fingerprinting yielded the best classification performance with an accuracy of 0.900, showcasing the strength of ensemble approaches in managing complex drug discovery challenges. The study also included feature importance analysis to identify critical molecular features influencing compound activity, offering valuable insights into structure-activity relationships.

This research contributes to drug development by demonstrating how machine learning can streamline and economise the process of screening drug candidates for diseases like Leishmaniasis(Kala aźar). By optimising fingerprinting techniques and leveraging sophisticated algorithms, the study enhances the efficiency and accuracy of drug discovery efforts.

Chapter-2: Introduction

Drug discovery is a rigorous, costly, and time-consuming process aimed at identifying new pharmaceuticals by screening vast libraries of chemical compounds for their potential efficacy against specific diseases. This intricate process involves several stages, from initial screening and optimisation to clinical trials, each meticulously designed to ensure that drug candidates are both effective and safe. The entire drug discovery journey can span over a decade, with extensive research and testing phases contributing to its length and complexity.Machine learning accelerates drug discovery by automating and optimising key tasks, such as data analysis and compound screening. By quickly analysing vast datasets and identifying patterns, machine learning models predict the efficacy of compounds more efficiently. Techniques like molecular fingerprinting streamline the representation and comparison of chemical structures, while predictive algorithms optimise drug candidates and reduce failures. This results in a significantly shorter timeline and lower costs for developing new pharmaceuticals, making the entire process faster and more cost-effective.

Leishmaniasis is a severe parasitic disease caused by protozoa from the genus Leishmania. It affects an estimated 700,000 to 1 million people annually, primarily in tropical and subtropical regions. The disease manifests in several forms, including cutaneous, mucocutaneous, and visceral leishmaniasis. out of which, Visceral leishmaniasis, also known as kala-azar, is the most dangerous and life-threatening form. If left untreated, it can lead to severe complications such as anemia, liver and spleen enlargement, and can be fatal in 95 percent of cases. Cutaneous leishmaniasis, while less deadly, can cause debilitating skin ulcers and disfigurement, significantly impacting patients' quality of life. Mucocutaneous leishmaniasis affects mucous membranes, leading to severe deformities and complications. The disease poses a significant public health challenge, causing both substantial morbidity and mortality, particularly in impoverished areas with limited access to healthcare.

For leishmaniasis drug discovery, machine learning (ML) leverages four key molecular fingerprints—Morgan, MACCS, RDKit, and Atom Pair. These fingerprints offer varied representations: Morgan details local atom environments, MACCS identifies predefined substructures, RDKit provides flexible features, and Atom Pair highlights spatial atom relationships. ML models use these fingerprints to predict the effectiveness of compounds,

streamline screening processes, and prioritise the most promising candidates. Ensemble techniques, such as Voting Classifiers, enhance prediction accuracy, speeding up the development of new treatments for leishmaniasis.

Various classical and ensemble machine learning models—including Random Forest, Logistic Regression, Gradient Boosting, K-Nearest Neighbours (KNN), Support Vector Classifier, Decision Tree, AdaBoost, and Artificial Neural Networks—were trained and evaluated to assess their effectiveness in predicting compound activity. Additionally, we investigated the use of ensemble techniques, such as Voting Classifiers, to improve the accuracy and reliability of the predictions.

The goal of this project is to advance the application of machine learning in drug discovery, showcasing how ML models can be instrumental in tackling complex diseases like pancreatic cancer. By incorporating computational methods into the drug discovery process, we aim to accelerate the identification of potential therapeutic candidates and reduce the time and cost involved in developing crucial treatments.

Chapter-3: Literature Survey

3.1 Introduction to Leishmaniasis

- **Aetiology**: Leishmaniasis is a parasitic disease caused by protozoan parasites belonging to the genus *Leishmania*. These parasites are transmitted to humans through the bites of infected female *Phlebotomine* sandflies, which act as vectors. The parasites thrive in the sandflies' digestive tract and are injected into the human host when the sandfly feeds on blood.

- **Global Impact**: Each year, approximately 700,000 to 1 million new cases of leishmaniasis are reported globally, with over 20,000 to 30,000 deaths annually. The disease predominantly affects tropical and subtropical regions, including parts of Africa, Asia, the Middle East, Latin America, and southern Europe. It is closely linked to poverty, malnutrition, displacement, and poor living conditions, and is considered a neglected tropical disease (NTD).

3.2 Forms of Leishmaniasis

- **Cutaneous Leishmaniasis (CL)**: This is the most common form, characterised by localised skin lesions or ulcers at the site of the sandfly bite. While not usually fatal, it can cause significant scarring and disfigurement, affecting quality of life and mental health. There are more than 1 million cases of CL each year, and it is more frequently seen in rural areas.

- **Mucocutaneous Leishmaniasis (MCL)**: This rare form of leishmaniasis affects the mucous membranes, particularly in the nose, mouth, and throat, often leading to disfiguring and destructive damage to the soft tissues. MCL usually develops as a complication of untreated cutaneous leishmaniasis and can result in serious morbidity, including facial deformities.

- **Visceral Leishmaniasis (VL, also known as Kala-Azar)**: The most severe and potentially fatal form, VL affects internal organs such as the liver, spleen, and bone marrow. Symptoms include prolonged fever, weight loss, fatigue, Anemia, and enlargement of the liver and spleen (hepatosplenomegaly). If left untreated, visceral leishmaniasis is almost always fatal. VL is responsible for the majority of deaths

associated with leishmaniasis, primarily affecting areas such as East Africa, South Asia, and Brazil.

3.3 Epidemiology and Distribution

- **Geographical Distribution**: Leishmaniasis is endemic in over 90 countries across various regions, with the highest burden seen in countries such as Brazil, India, Sudan, and Ethiopia. The distribution of the disease is influenced by the ecology of *Phlebotomine* sandflies, which thrive in warm, humid environments with abundant organic material. This leads to high incidence rates in rural and forested areas.

- **At-Risk Populations**: The disease disproportionately affects impoverished populations living in remote areas with limited access to healthcare and vector control programs. Conflict zones and regions with displaced populations are also at greater risk. Malnutrition and immunosuppression, including HIV infection, are significant risk factors for developing severe forms of the disease.

3.4 Diagnosis

- **Clinical Diagnosis**: Initial diagnosis often involves a detailed assessment of clinical symptoms, patient history, and travel to endemic regions. Signs of skin lesions or systemic symptoms such as fever and splenomegaly are key indicators.

- **Laboratory Diagnosis**:

    - *Parasitological Methods*: These include direct microscopy to detect the *Leishmania* parasite in tissue samples from skin lesions or lymph nodes. Culture of the parasite from clinical samples is also a common method.
    - *Molecular Techniques*: Polymerase Chain Reaction (PCR) is used to amplify *Leishmania* DNA from clinical samples, providing a more sensitive and specific diagnosis.
    - *Serological Tests*: These detect antibodies against *Leishmania* antigens, especially useful in diagnosing visceral leishmaniasis. Tests like the rK39 dipstick test are used for rapid diagnosis.

3.5 Treatment and Management

- **Conventional Treatments**:
    - *Antimonial Drugs*: Sodium stibogluconate and meglumine antimoniate are commonly used, although they have significant toxicity.
    - *Amphotericin B*: Especially in its liposomal form, is a more effective but costly treatment option, used primarily for visceral leishmaniasis.
    - *Miltefosine*: The only oral treatment for leishmaniasis, effective for both cutaneous and visceral forms, though emerging drug resistance is a concern.
- **Emerging Treatments**: Research into alternative therapies includes newer drugs like paromomycin and combination therapies to address drug resistance and improve efficacy. Vaccine development is ongoing, with several candidates in preclinical and clinical trials.

3.6 Challenges in Leishmaniasis Control

- **Predictive Modelling**: Machine learning (ML) models are increasingly used to predict the activity of compounds and potential drug candidates for leishmaniasis treatment. These models analyse chemical properties to identify promising molecules for further testing.

- **Molecular Fingerprinting**: Techniques like Morgan, MACCS, RDKit, and Atom Pair fingerprints allow ML algorithms to map out chemical structures, aiding drug discovery by comparing new compounds to known active molecules.

- **Ensemble Techniques**: Combining multiple models through ensemble methods like Voting Classifiers improves prediction accuracy and reliability. This approach helps prioritise potential drug candidates for experimental validation, speeding up the drug discovery process.

3.7 Recent Research and Developments

- **Genomic Studies**: Advances in sequencing the *Leishmania* genome have identified novel drug targets and resistance mechanisms. Genome-wide association studies

(GWAS) are helping researchers understand how the parasite evolves and adapts to treatment.

- **Immunological Research**: Studies on the immune response to *Leishmania* infection have revealed important insights into host-pathogen interactions, paving the way for vaccine development. Several experimental vaccines are in the pipeline, focusing on inducing long-lasting immunity.

- **Field Studies**: Ecological research on the behavior and habitat preferences of *Phlebotomine* sandflies has led to improved vector control strategies. Understanding their breeding sites and feeding habits is critical for developing effective interventions.

3.8 Future Directions

- **Innovative Therapies**: New drugs and combination therapies are being developed to address the challenges of drug resistance. Targeted therapies that inhibit specific *Leishmania* enzymes or pathways are under investigation, offering a more tailored approach to treatment.

- **Enhanced Diagnostics**: The development of rapid, point-of-care diagnostic tests with higher sensitivity and specificity is critical, especially for use in resource-limited settings. Innovations in portable diagnostic devices are expected to improve early detection.

- **Integrated Control Programs**: Successful control of leishmaniasis requires a multi-faceted approach, combining vector control, effective treatment, and public health education. Community-based programs that involve the local population in surveillance and prevention efforts are key to reducing transmission.

Chapter-4: Typical Timeline for Determining Drug Activity

The process of determining drug activity spans multiple stages of research and development, each requiring varying lengths of time based on complexity, resources, and regulatory requirements. Below is a comprehensive overview of the typical timeline for drug discovery and development:

4.1 Initial Screening

- **Purpose**: This is the first step in identifying potential drug candidates. High-throughput screening (HTS) is conducted to rapidly test thousands to millions of chemical compounds against a specific biological target, such as the *Leishmania* parasite or its molecular components. The goal is to identify compounds that demonstrate inhibitory activity.
- **Duration**: 2 to 6 months.
- **Techniques**: Robotic systems and automated assays are used to screen large chemical libraries, and hits are ranked based on their potency, selectivity, and drug-likeness.

4.2 Hit Identification

- **Purpose**: In this phase, the most promising compounds identified during initial screening are selected as "hits." These compounds show sufficient activity against the biological target and have the potential to be developed further.
- **Duration**: 1 to 3 weeks.
- **Techniques**: In vitro biochemical and cell-based assays are performed to measure the activity of the compounds, often in comparison to known reference drugs.

4.3 Hit Validation

- **Purpose**: The next step is to confirm the activity of the hit compounds through secondary testing. This ensures that the observed activity is reproducible and not an artifact of the initial screening conditions. Hits are also tested for specificity to the target and evaluated for off-target effects.
- **Duration**: 2 to 6 weeks.
- **Techniques**: Secondary assays, such as dose-response studies, are used to validate the hits and confirm their potential for further optimization.

4.4 Lead Optimization

- **Purpose**: Lead optimization involves improving the chemical structure of the validated hit compounds to enhance their potency, selectivity, and pharmacokinetic properties. The goal is to refine the leads into drug candidates that are effective and safe for further testing.
- **Duration**: 6 months to 3 years.
- **Techniques**: Iterative chemical modifications are made, guided by structure-activity relationship (SAR) studies. Computational techniques like molecular modeling and quantitative structure-activity relationships (QSAR) are used to predict the impact of modifications on the compound's activity.

4.5 Preclinical Testing

- **Purpose**: Before advancing to human trials, the optimized leads undergo preclinical testing in animal models. This step assesses the safety, efficacy, and pharmacokinetics of the compounds, ensuring that they have acceptable profiles for progression to clinical studies.
- **Duration**: 6 months to 2 years.
- **Techniques**: Studies include toxicology assessments, pharmacokinetic profiling, and efficacy testing in disease-relevant animal models, often utilizing rodents or non-human primates.

4.6 IND Application Preparation

- **Purpose**: The Investigational New Drug (IND) application is prepared to request approval from regulatory agencies (e.g., FDA, EMA) to begin clinical trials in humans. The application compiles all preclinical data, along with detailed plans for the proposed clinical trials.
- **Duration**: 3 to 6 months.
- **Techniques**: Documentation includes reports on the compound’s safety, pharmacology, manufacturing process, and the design of the clinical trial protocol.

4.7 Clinical Trials

- **Phase I (6 to 12 months)**:

    - **Purpose**: This phase primarily assesses the safety and tolerability of the drug in a small group (20–100) of healthy volunteers or patients. The trial also determines the appropriate dosage and identifies potential side effects.
    - **Techniques**: The drug is administered in escalating doses to observe its effects on the human body.
- **Phase II (1 to 2 years)**:
    - **Purpose**: After Phase I demonstrates safety, Phase II trials evaluate the drug's efficacy in patients with the disease (100–300 participants). These trials further assess safety and monitor for short-term side effects.
    - **Techniques**: Randomized controlled trials are often used to compare the drug's effect against a placebo or standard treatment.
- **Phase III (2 to 4 years)**:
    - **Purpose**: In this phase, the drug's efficacy is confirmed in a larger population (1,000–3,000 patients). The trial also monitors long-term side effects and compares the drug to existing treatments.
    - **Techniques**: This phase involves multi-center trials, often conducted in different regions or countries, to gather diverse patient data.

4.8 Regulatory Review and Approval

- **Purpose**: Following successful clinical trials, a New Drug Application (NDA) or Marketing Authorization Application (MAA) is submitted to regulatory authorities for market approval. The regulatory agency reviews the data from all phases of development to ensure the drug is safe, effective, and manufactured to high standards.
- **Duration**: 6 months to 2 years.
- **Techniques**: The review process involves evaluation of clinical trial data, manufacturing protocols, and proposed labeling.

4.9 Post-Marketing Surveillance

- **Purpose**: Once a drug is approved and available on the market, post-marketing surveillance (Phase IV) is conducted to monitor its safety and effectiveness in the general population. This stage helps detect any rare or long-term side effects that may not have been apparent during clinical trials.
- **Duration**: Indefinite.

- **Techniques**: Pharmacovigilance systems are used to track adverse drug reactions (ADRs) and collect real-world data from healthcare providers and patients.

4.10 Total Duration of the Procedure

- The entire process, from the discovery of a compound to its approval for clinical use, typically takes **6 to 20 years**. Factors such as the complexity of the disease, regulatory hurdles, and the success of clinical trials can greatly influence the overall timeline.

Chapter-5: Problem Statement

Development of a Machine Learning Model for Accelerating Drug Discovery in the Treatment of a High-Priority Disease Called Leishmaniasis

Leishmaniasis, a neglected tropical disease affecting millions globally, lacks rapid and cost-effective treatments, particularly for its severe forms like visceral leishmaniasis. Current drug discovery methods for leishmaniasis are time-consuming, resource-intensive, and challenged by the emergence of drug resistance. There is an urgent need to streamline the drug discovery process to identify effective compounds more quickly.

Key Challenges:

- The traditional drug development timeline can take 6 to 20 years, with significant financial and logistical barriers.
- Identifying potential drug candidates with high efficacy and low toxicity remains a complex and iterative process.
- Emerging resistance to existing treatments highlights the need for new compounds that can target drug-resistant strains of *Leishmania*.

Proposed Solution: To address these challenges, a machine learning (ML)-driven approach is proposed to accelerate the identification and optimization of drug candidates for leishmaniasis. The model will leverage large datasets of known compounds and their biological activity to:

- **Predict compound activity** against *Leishmania* based on molecular fingerprints and chemical properties.
- **Prioritize compounds** with favorable drug-like properties for experimental testing.
- **Optimize lead compounds** through iterative model training, using ensemble techniques to enhance prediction accuracy.

By integrating ML models such as Random Forest, Support Vector Machines (SVM), and ensemble methods like Voting Classifiers, this approach aims to reduce the time required for hit identification, validation, and lead optimization. The ultimate goal is to accelerate the discovery of safe and effective treatments for leishmaniasis, potentially shortening the drug development timeline from years to months.

Chapter-6: Description of various training Modules

The training modules for this project focus on equipping researchers and practitioners with the necessary skills and knowledge to develop machine learning models for drug discovery, specifically targeting Leishmaniasis. These modules are structured to guide participants through the key steps of the project, from data acquisition to model evaluation, with a focus on practical application in drug discovery.

1. Module 1: Introduction to Drug Discovery and Machine Learning

- **Objective**: Provide a foundational understanding of the drug discovery process and the role of machine learning in accelerating the identification of potential drug candidates.
- **Content**:
  - Overview of **Leishmaniasis** and its global impact.
  - Stages of drug discovery and development (e.g., **hit identification**, **lead optimization**, **preclinical testing**).
  - Introduction to machine learning concepts and their application in drug discovery.
  - Case studies of successful machine learning applications in pharmaceuticals.
- **Hands-On**: Introduction to Python and essential libraries for machine learning (e.g., **Pandas**, **Scikit-learn**, **RDKit**).

2. Module 2: Data Acquisition and Preprocessing

- **Objective**: Teach participants how to collect, preprocess, and clean biological datasets for machine learning tasks in drug discovery.
- **Content**:
  - Data collection from sources like **ChEMBL**, **PubChem**, or other bioactivity databases.
  - Understanding the structure of bioactivity data (e.g., **SMILES strings**, **IC50/ EC50 values**).
  - Techniques for handling missing data, duplicate removal, and outlier detection.

    - Unit standardization for bioactivity data (e.g., converting various units to a uniform measurement like nM).
    - Introduction to **SMILES** and how molecular structures are represented.
- **Hands-On**: Preprocessing a real dataset, including **duplicate removal**, **null handling**, and **feature selection** using Python. Introduction to **MinMax Scaler** and **KNN imputation**.

3. Module 3: Feature Engineering and Fingerprinting

- **Objective**: Teach the generation of molecular features (fingerprints) for input into machine learning models.
- **Content**:
    - Introduction to molecular fingerprints: **RDKit**, **Morgan**, **MACCS**, and **Atom Pair fingerprints**.
    - Feature engineering strategies for drug discovery (e.g., extracting descriptors like molecular weight, hydrogen bond donors/acceptors).
    - How to use molecular fingerprints to represent chemical structures for machine learning.
    - Comparison of different fingerprinting methods for machine learning tasks.
- **Hands-On**: Generate molecular fingerprints from a dataset using **RDKit** and visualize the resulting feature vectors.

4. Module 4: Machine Learning Model Training

- **Objective**: Equip participants with the skills to build, train, and optimize machine learning models for predicting compound activity.
- **Content**:
    - Supervised learning techniques (classification) for predicting drug activity.
    - Overview of key machine learning algorithms: **Random Forest**, **Support Vector Machines (SVM)**, **Gradient Boosting**, and **Voting Classifiers**.
    - Hyperparameter tuning and model optimization techniques.
    - Cross-validation techniques to evaluate model performance.

- **Hands-On**: Train various classifiers (e.g., **Random Forest**, **SVM**) on the preprocessed dataset and tune hyperparameters using **GridSearchCV** or **RandomizedSearchCV**.

5. Module 5: Model Evaluation and Validation

- **Objective**: Teach how to evaluate machine learning models and ensure that they generalize well to unseen data.
- **Content**:
    - Model evaluation metrics: **Accuracy**, **Precision**, **Recall**, **F1-Score**, **ROC-AUC**.
    - Understanding and handling class imbalance (e.g., using **SMOTE** or **undersampling** techniques).
    - Overfitting vs. underfitting: Techniques to avoid overfitting (e.g., **cross-validation**, **regularization**).
    - Strategies for evaluating the generalization of models to unseen data.
- **Hands-On**: Evaluate the trained models using various metrics, generate **confusion matrices**, and visualize the **ROC curves**.

6. Module 6: Model Deployment and Interpretation

- **Objective**: Learn how to interpret the results of machine learning models and deploy them for real-world applications.
- **Content**:
    - Interpreting the model results: Feature importance analysis using Random Forest and decision trees.
    - Understanding the significance of molecular descriptors in predicting drug activity.
    - Introduction to model deployment tools (e.g., **Flask**, **Streamlit**).
    - Ethical considerations and limitations of using machine learning in drug discovery.
- **Hands-On**: Deploy a simple web app using **Streamlit** or **Flask** to predict drug activity based on SMILES input and model output.

7. Module 7: Advanced Topics (Optional)

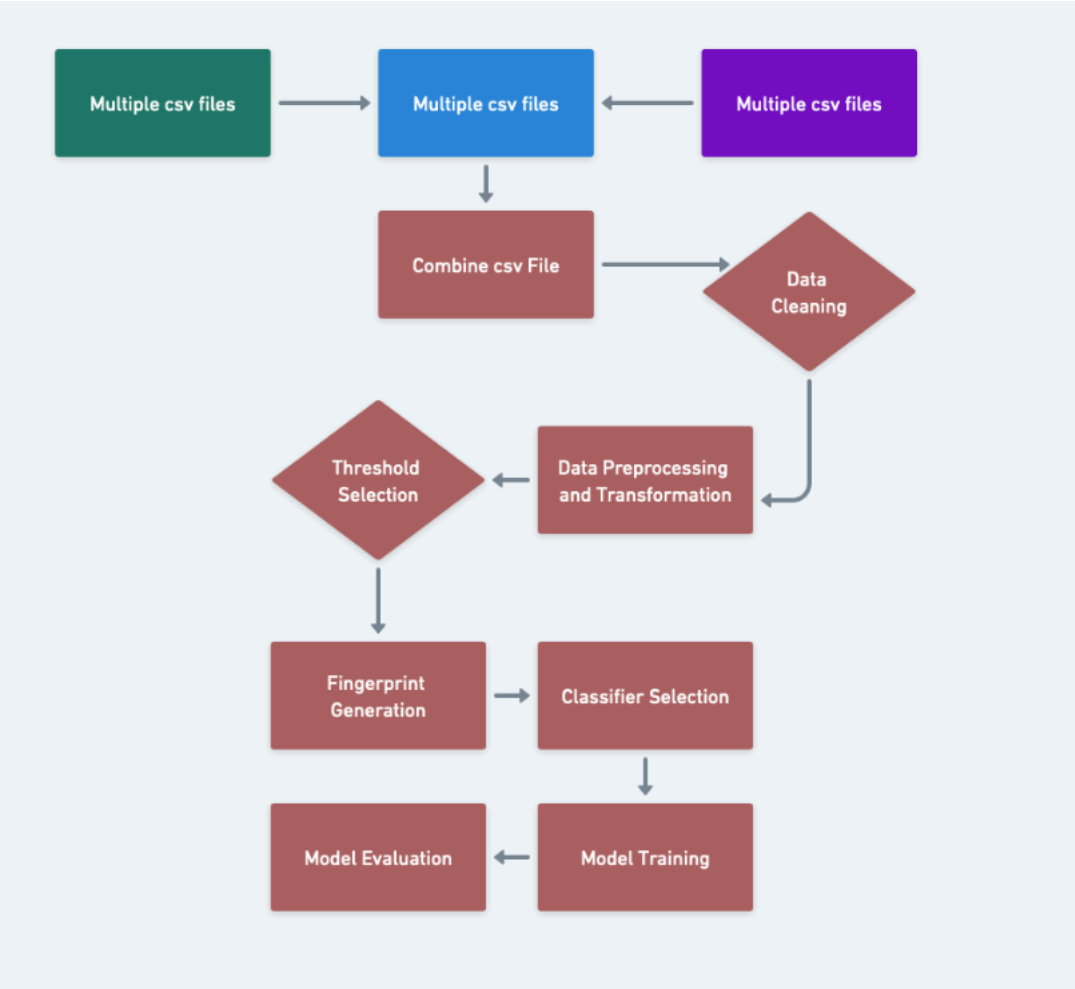
- **Objective**: Introduce advanced topics for those interested in further improving their models and exploring deeper areas of drug discovery.
- **Content**:
  - Deep learning models for drug discovery (e.g., **Graph Neural Networks (GNN)**, **Autoencoders**).
  - Ensemble learning techniques for improving model accuracy and robustness.
  - Advanced molecular descriptors and cheminformatics tools.
  - Techniques for feature selection and dimensionality reduction (e.g., **PCA**, **t-SNE**).
- **Hands-On**: Implement advanced models such as **GNN** for molecular property prediction, or use **ensemble methods** to improve model performance.

Chapter-7: Methodology Adopted

7.1 Design of Experiment/ Flow Chart

7.1.1 Dataset Collection and Preprocessing

The dataset used for the project was sourced from the **ChEMBL Database**, a large-scale bioactivity database that provides information on the biological activities of small molecules. The data was originally available as multiple compressed comma-separated value (CSV) files,



which were downloaded and subsequently merged into a single unified database for analysis. Below is a detailed breakdown of the data collection and preprocessing steps:

Data Collection

- **Original Dataset**: The initial dataset comprised **278,045 data points**, representing a broad range of bioactivity data for small molecules. These data points included molecular structures represented as **SMILES (Simplified Molecular Input Line Entry**

System) strings, bioactivity measurements, and additional metadata such as experimental conditions and molecule properties.

Data Preprocessing

The preprocessing of this dataset was a multi-step process aimed at cleaning and refining the data for machine learning model development. The following steps were taken:

1. **Duplicate Removal**:
    - **Duplicates based on SMILES**: Out of the 278,045 original entries, **183,610 entries were identified as duplicates** on the basis of the SMILES signatures. Duplicate entries occur when the same molecule is reported multiple times across different experiments or studies. Removing these duplicates left a refined dataset of **94,435 data points**.
2. **Further Data Refinement**:
    - Additional preprocessing steps were undertaken to remove irrelevant or redundant data, resulting in a final dataset of **10,418 rows**. This significant reduction in data size was necessary to focus on the most reliable and relevant data points for the project.
3. **Bioactivity Filtering**:
    - The dataset contained dozens of types of biological activity measurements. However, for the purposes of the study, only the most commonly reported bioactivity types were retained:
        - **IC50**: Concentration required to inhibit 50% of the biological activity.
        - **Percent Effect**: Percentage of inhibition or effect observed at a specific concentration.
        - **Inhibition**: A measure of the extent to which a molecule inhibits the

Figure 1: Workflow presented as a flowchart.

biological target.
        - **EC50**: The concentration required to achieve 50% of the maximal effect.
        - **IC90**: Concentration required to inhibit 90% of the biological activity.

- Other less commonly reported activity types were discarded to streamline the dataset and avoid sparsity.

4. **Unit Standardization**:

   - To ensure consistency across bioactivity measurements, a uniform unit of **nM (nanomolar)** was chosen for all values. Any measurements reported in other units (e.g., mM) were **converted to nM**, while entries that could not be converted were discarded. This standardization step was crucial for ensuring uniformity in the data and allowing for accurate comparison between molecules.

5. **Classification of Molecules**:

   - The molecules in the dataset were classified as either **Active (2)** or **Inactive (1)** based on a threshold value of **10,000 nM**. Molecules with activity values below this threshold were labeled as **Active**, while those with values above the threshold were considered **Inactive**. This binary classification was used to prepare the dataset for supervised machine learning tasks.

6. **Column Selection**:

   - Out of the **47 columns** initially present in the dataset, only **4 key columns** were retained:
     - **SMILES**: The molecular structure in SMILES format, representing the chemical composition of each molecule.
     - **Standard Type**: The type of bioactivity measurement (e.g., IC50, EC50).
     - **Standard Value**: The numerical value of the bioactivity measurement.
     - **Standard Units**: The unit of measurement, standardized to nM.
   - The remaining columns, which included additional experimental metadata and molecule properties, were discarded to focus on the core data required for the machine learning model.

7. **Handling Missing Values**:

   - **Null values** present in the dataset were either:
     - **Removed**: If the missing data was deemed non-critical or the entries were sparse.
     - **Imputed using K-Nearest Neighbors (KNN) Imputer**: For cases where imputation was feasible, the KNN imputation method was used to

estimate missing values based on the similarity between data points. This ensured that the dataset remained as complete as possible without introducing significant bias.

8. **Feature Scaling**:

   - To normalize the data and ensure that all features had comparable scales, the **MinMax Scaler** was applied. This scaling method transformed the bioactivity values into a range between 0 and 1, which is crucial for machine learning models that rely on gradient-based optimization techniques, such as neural networks or support vector machines (SVM). Scaling ensures that the model treats all features equally during training.

7.1.2 Threshold Selection

A threshold value of 10000 nM was selected as per domain expert's advice. Using, thi

Status	Count
Inactive	5774
Active	4226

Table 1: No. Of Active and Inactive Compounds

threshold , values >10000nM were termed as Class 1 i.e. "Inactive" whereas values <10000nM were termed as Class 2 i.e. "Active" , whereas values >1000nM and <10000nM were termed as "Intermediate".

The intermediate class was then removed to minimise overlapping of data and properly defining the two important active and Inactive classes.

7.1.3 Molecular Fingerprints

Four methods for molecular fingerprinting were utilised to develop robust models, all generated using the RDKit library. The chosen fingerprinting methods include RDKit (2048 bits), Atom Pairs (2048 bits), MACCS keys (166 bits), and Morgan fingerprints (2048 bits). These methods encode molecular structures by considering atoms and their neighbouring atoms, activating bits based on these structural features. While the initial bit representations vary in size, they are generally standardised to a 2048-bit format through hashing for uniformity. The 4 types of fingerprints are described down below:

- Morgan Fingerprint: Morgan Fingerprints, also known as circular fingerprints, encode molecular structures based on the atom's environment in concentric circles. They are widely used in drug discovery for their ability to capture detailed structural information and facilitate the classification of compounds.
- MACCS Fingerprint: MACCS Fingerprints are a type of fixed-length bit vector that encode the presence or absence of predefined substructures in molecules. They provide a straightforward representation of molecular features and are valuable for quick and efficient compound screening.
- RDKit Fingerprint: RDKit Fingerprints are generated using the RDKit library and capture various aspects of molecular structure, including atom pairs and their relationships. They are versatile and can be customised to include different types of structural information for enhanced predictive performance.
- Atom Pair Fingerprints: Atom Pair Fingerprints encode the relationship between pairs of atoms in a molecule. They are particularly useful for capturing the spatial arrangement of atoms, which can be crucial for understanding molecular interactions and activities.

7.1.4 Principal Component Analysis of Fingerprint Data

Principal Component Analysis (PCA) was applied to molecular fingerprint data to reduce its dimensionality while retaining the most critical information. Given the high-dimensional nature

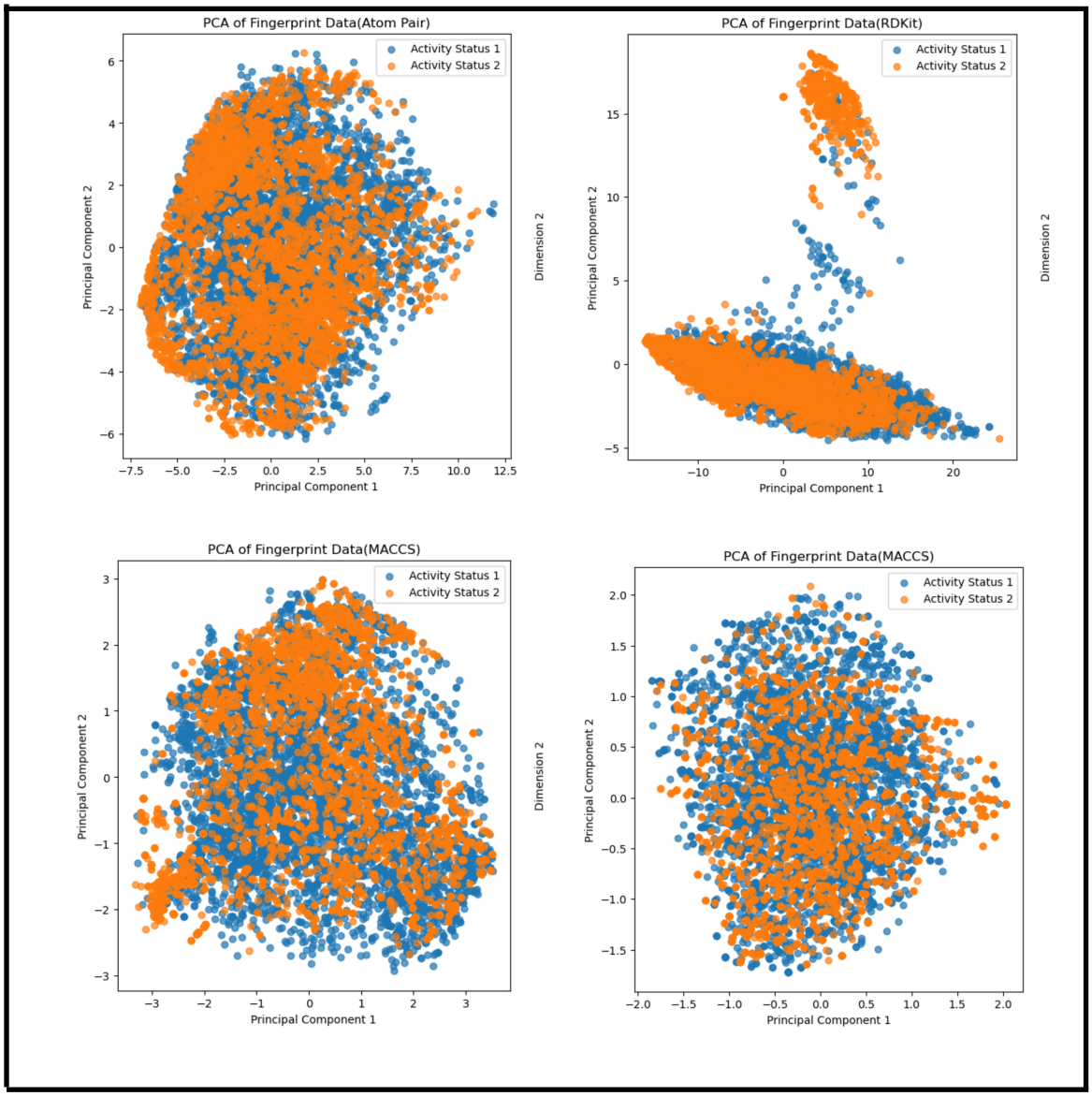


Figure 2: PCA Plots of the Four Fingerprints

of fingerprint data, PCA helps simplify the dataset by transforming it into a smaller set of principal components that capture the majority of variance in the original features. This not only enhances computational efficiency but also helps visualise patterns in the data, such as clusters of similar molecules. By reducing redundancy and noise in the fingerprint features,

PCA enables better insights into molecular relationships and improves the performance of downstream machine learning models.

7.1.5 t-distributed Stochastic Neighbour Embedding of Fingerprint Data

t-SNE (t-distributed Stochastic Neighbour Embedding) was used on molecular fingerprint data to visualise high-dimensional relationships in a lower-dimensional space, typically 2D or 3D. Unlike PCA, which focuses on capturing global variance, t-SNE excels at preserving local

similarities, making it ideal for exploring complex structures in fingerprint data. By applying t-SNE, researchers could identify clusters of molecules with similar fingerprints, revealing

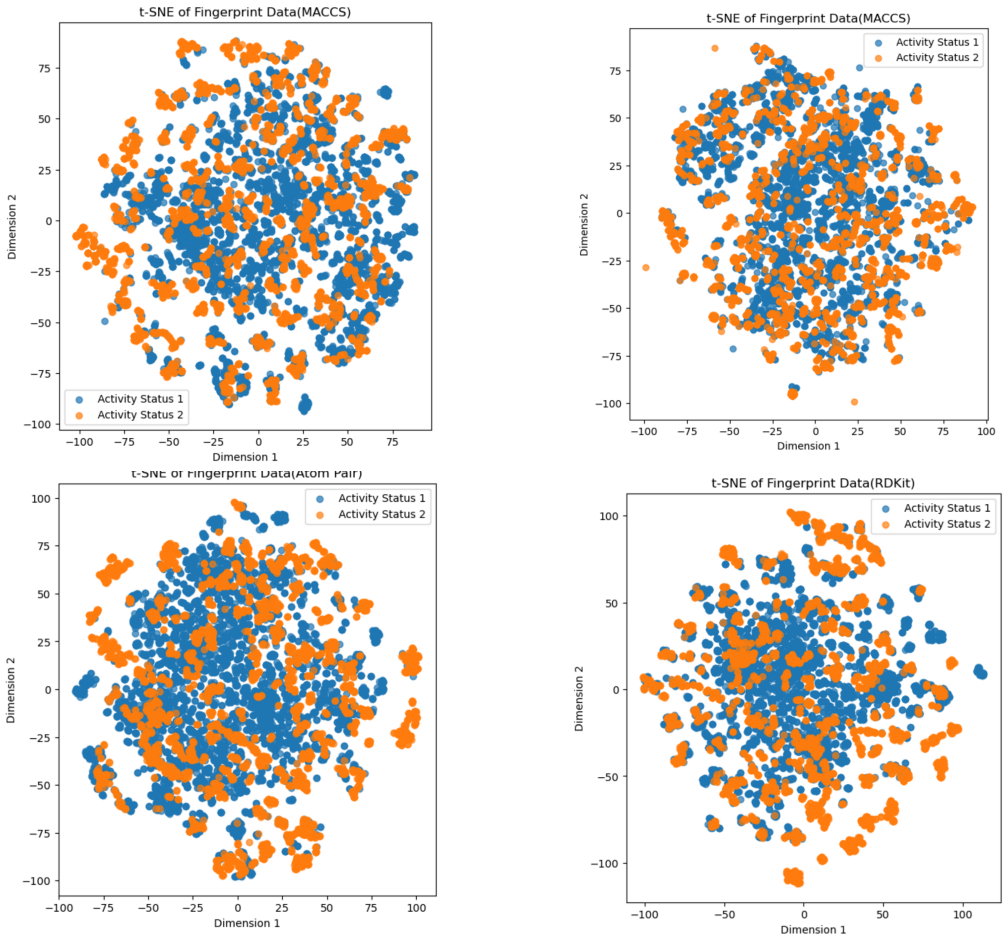


Figure 3: t-SNE plot of the four fingerprints

underlying patterns such as chemical properties or

functional similarities. This technique is particularly useful for gaining intuitive, visual insights into the molecular data, aiding in tasks like identifying lead compounds or understanding molecular diversity.

PCA	t-SNE
Dimensionality reduction focusing on capturing maximum variance	Dimensionality reduction for visualizing high-dimensional data
Global structure (variance across the entire dataset)	Local structure (similarities between neighbouring data points)
Linear transformation	Non-linear transformation

PCA	t-SNE
Uncorrelated principal components	Low-dimensional embedding (usually 2D or 3D) for visualization
Components can be interpreted based on the original features	Embeddings are difficult to interpret directly
Relatively low; scales well to large datasets	Computationally expensive, especially for large datasets

Table 1: PCA vs t-SNE

7.1.6 Performance Evaluation

The performance of the models was assessed and compared primarily using two metrics: Total Accuracy and Recall Score for inactive compounds. The choice of Recall as a key metric was made to minimise the number of potentially active molecules incorrectly classified as inactive. The definitions for these metrics are as follows:

1. **Total Accuracy:** It is the number of correct predictions out of all the predictions made.

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

2. **Recall Score:** Proportions of actual positives identified correctly.

$$\text{Recall} = \frac{TP}{TP + TN}$$

7.2 Hardware and Software Used

7.2.1 Hardware Used

- CPU and GPU: Apple M1 Chip
- RAM: 8 GB
- Storage: 256 GB SSD memory

7.2.2 Software Used

- Operating System: macOS Sonoma 14.6.1
- Code Editor : Visual Studio Code v1.93.1
- Programming Language: Python 3.12

- Jupyter Notebook
- LaTEX (For Project Report)

Python Libraries Used: Numpy , Pandas, Matplotlib, SeaBorn, RDKit, SciKit-Learn, XGBoost, tqdm.

7.3 Data Flow Diagram & Algorithms Used

Algorithms Used

Seven classical algorithms namely Random Forest, Logistic Regression, Gradient Boosting, K-Nearest Neighbours (KNN), Support Vector Classifier, Decision Tree, AdaBoost, one ensemble model VotingClassifier and ANN algorithms(MLP Classifier) were used to train and test a total of 36 models based on the four types of fingerprints mentioned in subsection Molecular Fingerprints.

Seven classical algorithms namely Random Forest, Logistic Regression, Gradient Boosting, K-Nearest Neighbours (KNN), Support Vector Classifier, Decision Tree, AdaBoost, Voting Classifier, ANN models were constructed using SciKit-Learn python package. All the models were trained on Apple M1 chip. The models have been briefly discussed below.

7.3.1 Random Forest

An ensemble learning technique that constructs a multitude of decision trees during training and merges their outputs to improve accuracy and control overfitting. Each tree is built using a random subset of features and data points. This method is robust to noise and effective for both classification and regression tasks.

7.3.2 Logistic Regression

A statistical method for binary classification that uses a logistic function to model the probability of a binary outcome based on one or more predictor variables. It estimates the relationship between the dependent variable and independent variables using a linear equation. Despite its simplicity, it’s quite effective for linearly separable data

7.3.3 Gradient Boosting

An iterative ensemble technique that builds models sequentially, with each model correcting the errors of its predecessor. It combines weak learners, usually decision trees, to form a strong

predictive model. Gradient boosting is highly flexible and can be fine-tuned for high performance but is prone to overfitting if not managed carefully.

7.3.4 K-Nearest Neighbours

A non-parametric classification algorithm that assigns a class to a sample based on the majority vote of its K nearest neighbours in the feature space. The algorithm is simple and intuitive, but its performance can degrade with large datasets and high-dimensional spaces. KNN is computationally expensive during prediction as it requires calculating distances between the sample and all other data points.

7.3.5 Decision Tree

A model that makes decisions by splitting data into subsets based on feature values, forming a tree structure. Each branch represents a decision rule, and each leaf node represents an outcome. Decision trees are easy to interpret but can be prone to overfitting, especially with complex trees.

7.3.6 AdaBoost

An ensemble learning method that combines multiple weak classifiers to create a strong classifier. It sequentially adjusts the weights of misclassified instances so that subsequent classifiers focus more on difficult cases. AdaBoost improves the performance of weak models and is less prone to overfitting compared to other boosting methods.

7.3.7 VotingClassifier

An ensemble technique that combines predictions from multiple classifiers by aggregating their votes. It can be used with different types of classifiers to leverage their individual strengths. Voting can be done using majority voting or averaging predicted probabilities, providing robustness to model variance.

7.3.8 Artificial Neural Network (MLP Classifier)

A type of deep learning model that consists of multiple layers of neurone, with each layer learning to transform the data in a way that captures complex patterns. MLPs are capable of modelling intricate relationships but require significant computational resources and careful

tuning of hyper-parameters. They are particularly useful for tasks like image and speech recognition.

Optimisation

7.3.8 Handling Dataset Imbalances

SMOTE (Synthetic Minority Over-sampling Technique) a popular technique for handling imbalanced datasets, especially in classification problems was used to optimize the model. It generates synthetic examples of the minority class to balance the dataset.

How SMOTE Works:

- Identify Minority Class Samples: SMOTE focuses on the minority class in an imbalanced dataset.
- Select K Nearest Neighbours: For each sample in the minority class, it finds the k-nearest neighbours (typically, k=5).
- Generate Synthetic Data: SMOTE randomly selects one of the k-nearest neighbours and generates a new synthetic sample along the line segment connecting the two samples.

7.3.9 Selection of Fingerprints

Among the fingerprinting techniques, RDKit Fingerprints were selected for further analysis and compression due to their superior accuracy and performance in capturing complex substructures within molecules. This was determined by the models' overall accuracy and reliability in identifying active compounds. To validate the precision of this compression approach, cross-validation was performed.

7.4 Characterisations/Snapshots of results obtained

7.5 Model training

The models have been split on a 80/20 split for training and testing respectively. The below figure shows the performance of single models under given fingerprint.

As can be interpreted from the graphs, Morgan Fingerprinting method underperformed against other fingerprinting methods.

7.6 Ensemble Models

After modelling and assessing individual models for each fingerprint, I moved on to evaluating ensemble models using the VotingClassifier to reduce bias and enhance performance. The experiments involved testing various combinations of the top-performing individual models, initially prioritising recall and then fine-tuning the combinations for the best accuracy within the VotingClassifier model. The results of ensemble models are shown below:

7.6.1 Morgan Fingerprinting

I evaluated the performance of several machine learning models trained on RDKit fingerprints using four metrics: Accuracy, Precision, Recall, and F1 Score. The results showed consistently high accuracy across most models, with the MLP Classifier achieving the highest accuracy of 0.900. However, recall scores tended to be lower. RDKit-tuned models showed improved recall, with the highest reaching 0.830, though this improvement often resulted in a slight reduction in accuracy.

7.6.2 Atom Pair Fingerprinting

The study evaluates the performance of several machine learning models based on atom-pair fingerprint data using different metrics:Accuracy, Precision, Recall, and F1 Score . When models consistently showed high accuracy, ranging from 0.890 to 0.900, demonstrating their reliability in correct classifications, though their recall varied, indicating potential limitations in identifying all relevant instances. On the other hand, models enhanced the identification of relevant cases, with recall scores reaching up to 0.790, while maintaining solid accuracy up to 0.890. Ensemble models, combining seven classifiers tuned for either accuracy or recall, surpassed individual models by striking a better balance between accuracy (up to 0.888) and recall (up to 0.789).

7.6.3 MACCS Fingerprinting

The study examines the performance of various machine learning models trained on MACCS fingerprint data, applying different evaluation metrics: Accuracy, Precision, Recall, and F1 Score.When models consistently showed high accuracy, ranging from 0.888 to 0.894, demonstrating their strength in making correct classifications, though their recall scores varied,

indicating challenges in identifying all relevant instances. In contrast, when models were better at detecting relevant cases, with recall scores reaching up to 0.820 while maintaining strong accuracy as high as 0.880. Ensemble models, which combined seven classifiers optimised for either accuracy or recall, outperformed individual models by striking a better balance, with accuracy reaching up to 0.892 and recall up to 0.810.

7.6.4 RDKit Fingerprinting

The provided visualisations present a comprehensive comparison of recall and accuracy metrics across various machine learning models using the RDKit fingerprint dataset. The analysis reveals that models consistently achieve high accuracy, with scores around 0.894 to 0.902. Notably, the SVC and RF models exhibit the highest accuracy at 0.901 and 0.902. In contrast, recall values vary more significantly, with scores ranging from 0.720 to 0.820, highlighting the impact of tuning on this metric. the 7-voter classifier, accuracy remains 0.901, while recall values fluctuate, with the ADABoost classifier exhibiting the lowest recall at 0.570.

7.6.5 Model Accuracy

In total, 36 models were trained on 4 Fingerprints. 7 Classifiers, 1 MLP Classifier(ANN), and 1 Ensemble Model of the seven classifiers with the ANN for each of the 4 fingerprints.

The most suitable model for the problem came out as Random forest boasting an average accuracy of 0.900 and

The most important fingerprint was RDKit Fingerprint with its most important feature being RDKit_747.

The Accuracies, Precision, Recall and F1 scores of the various models for the fingerprints are represented in the graphs below.

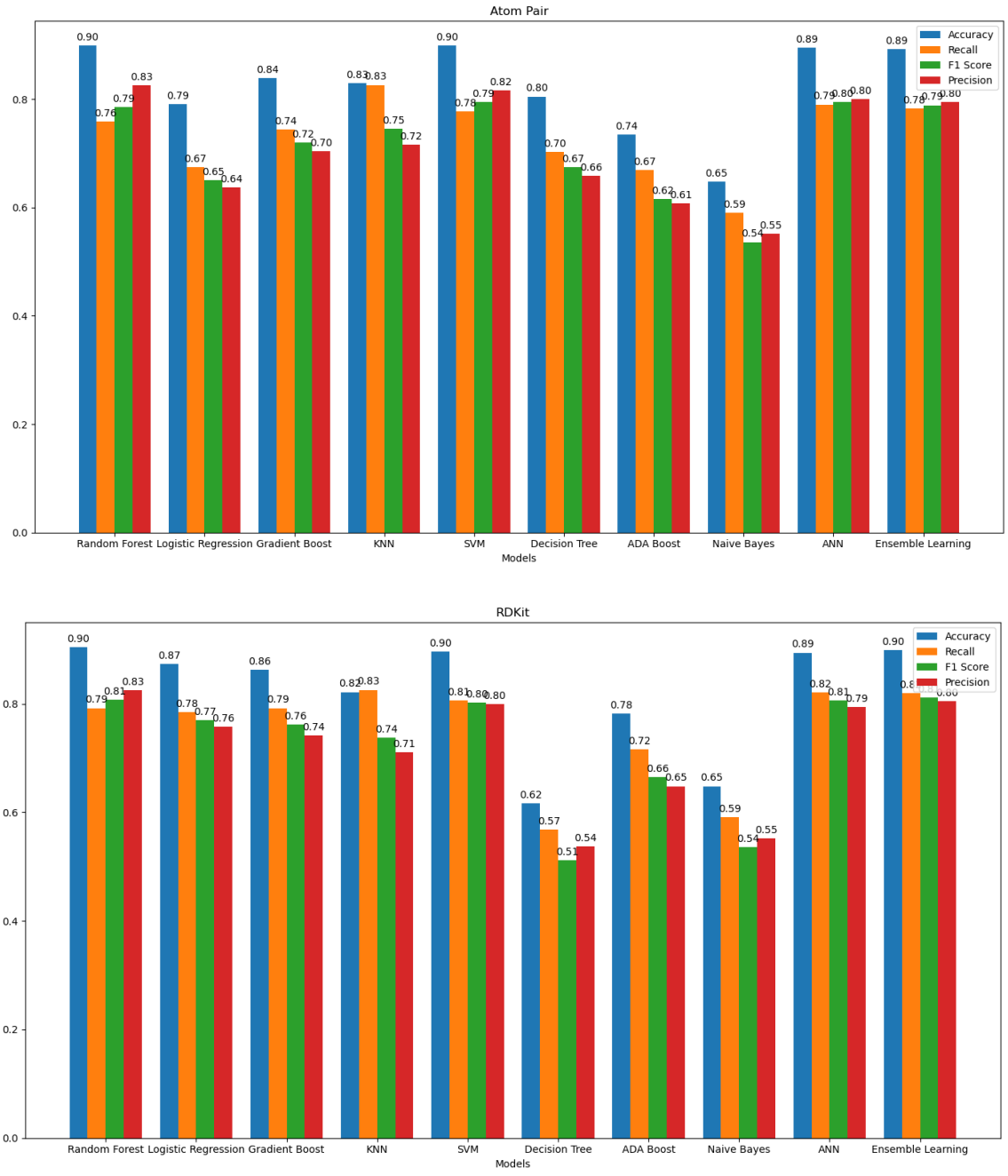


Figure 4: Accuracy, Recall, Precision, F1 Score of Atom Pair and RDKit Fingerprints

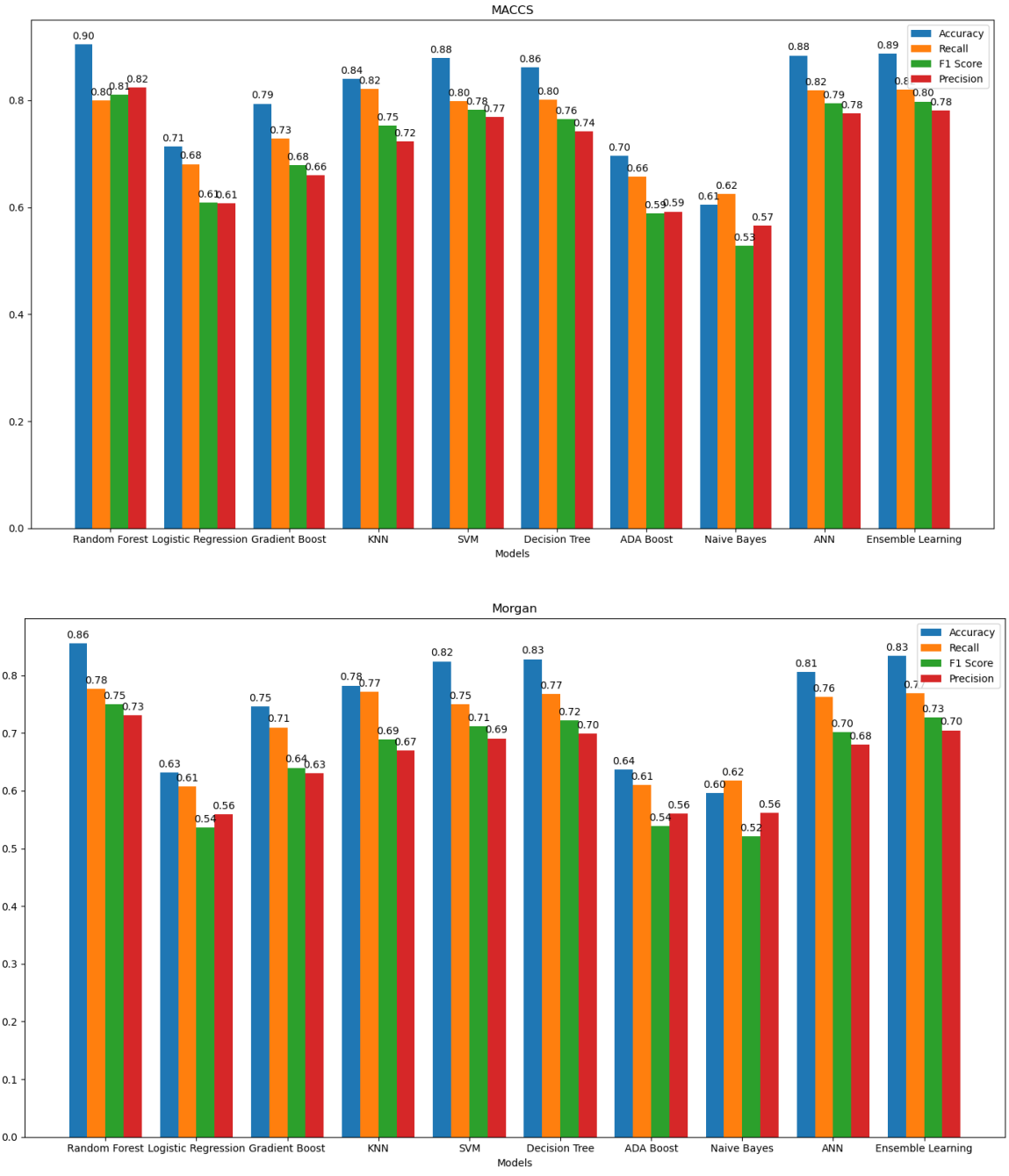


Figure 5: Accuracy, Recall, Precision, F1 Score of MACCS and Morgan Fingerprints

Chapter-8: Results and Discussions

8.1 Important features

8.1.1 Feature importance

molecular fingerprints were used as features to predict outcomes such as drug resistance or disease progression. These fingerprints, which capture the structural and chemical properties of molecules, were crucial in understanding how different compounds or genetic variations affect the parasite's behaviour. By applying feature importance techniques, the project could identify which molecular fingerprints had the greatest impact on the model's predictions. This helped highlight key molecular patterns, guiding further research into drug discovery and the molecular mechanisms underlying Leishmania infection.

Top 30 features for each of the molecular fingerprints were taken and represented in the graphs below:

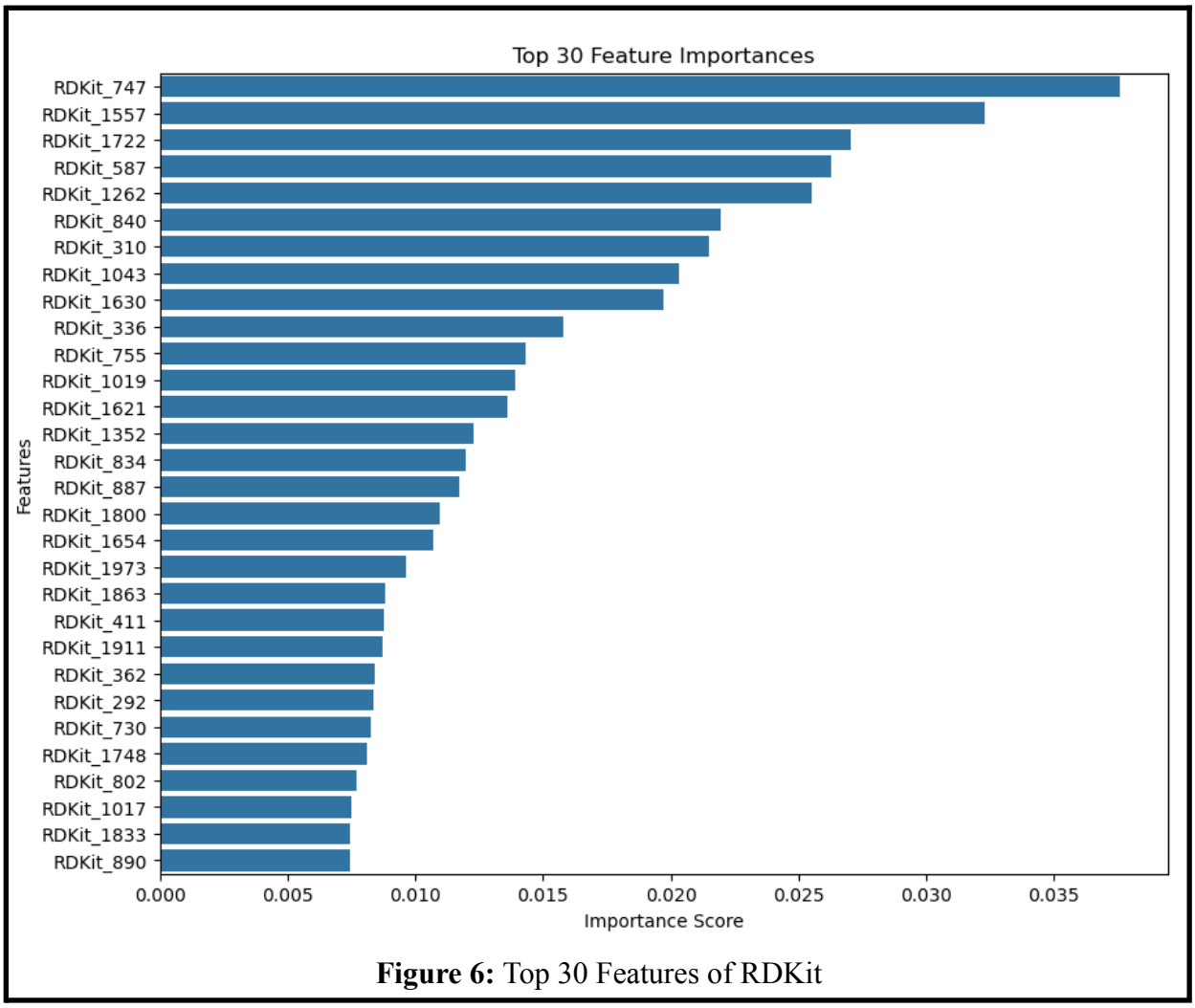


Figure 6: Top 30 Features of RDKit

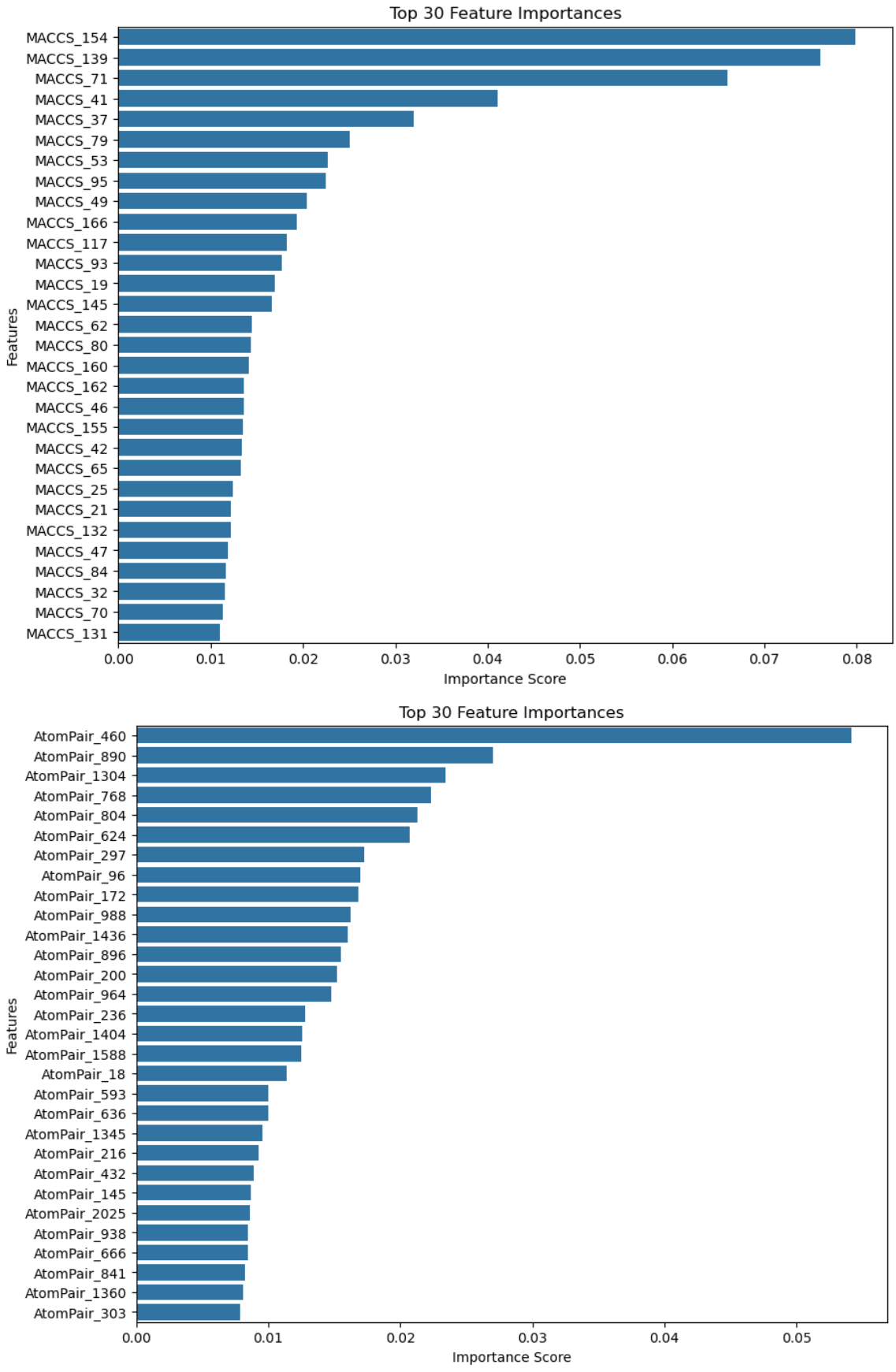
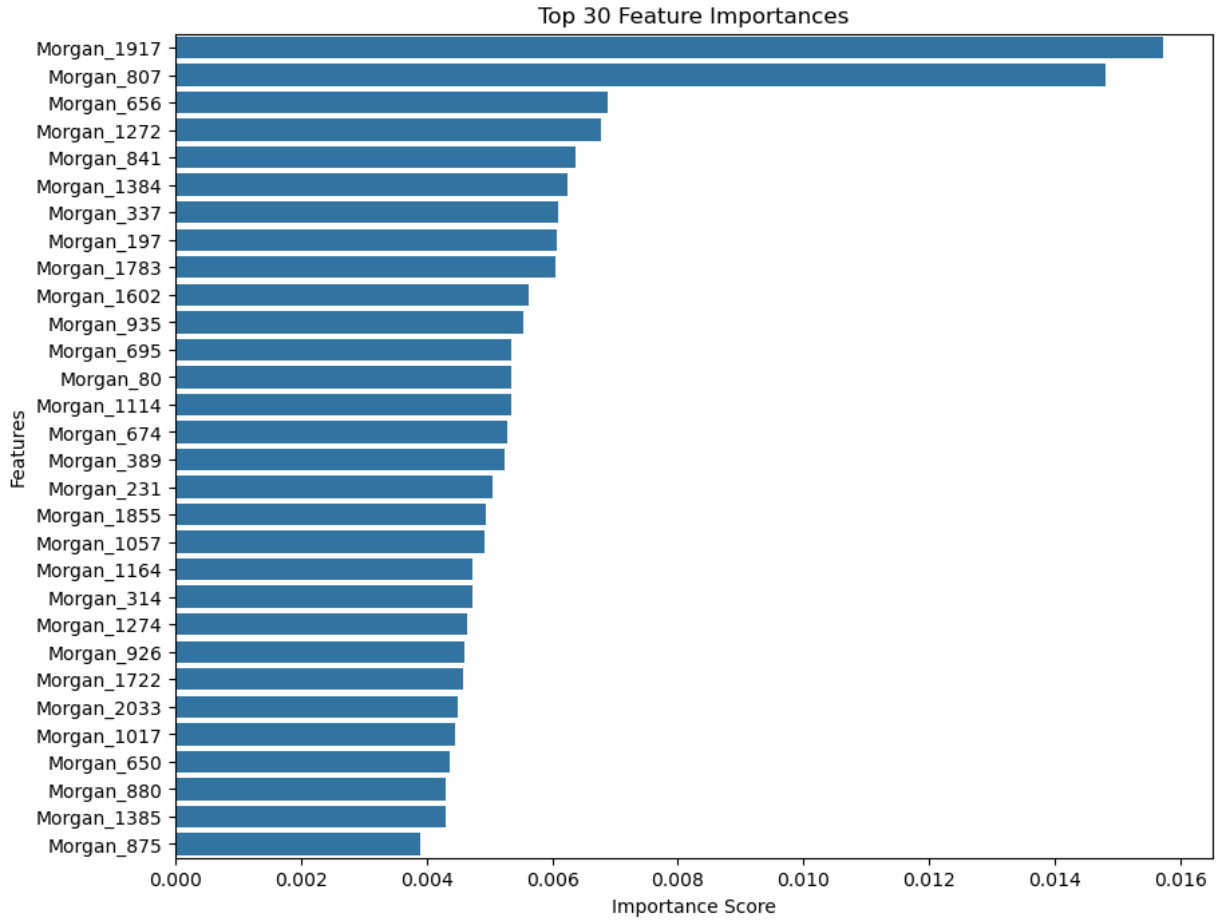


Figure 7: Top 30 Features of RDKit and MACCS



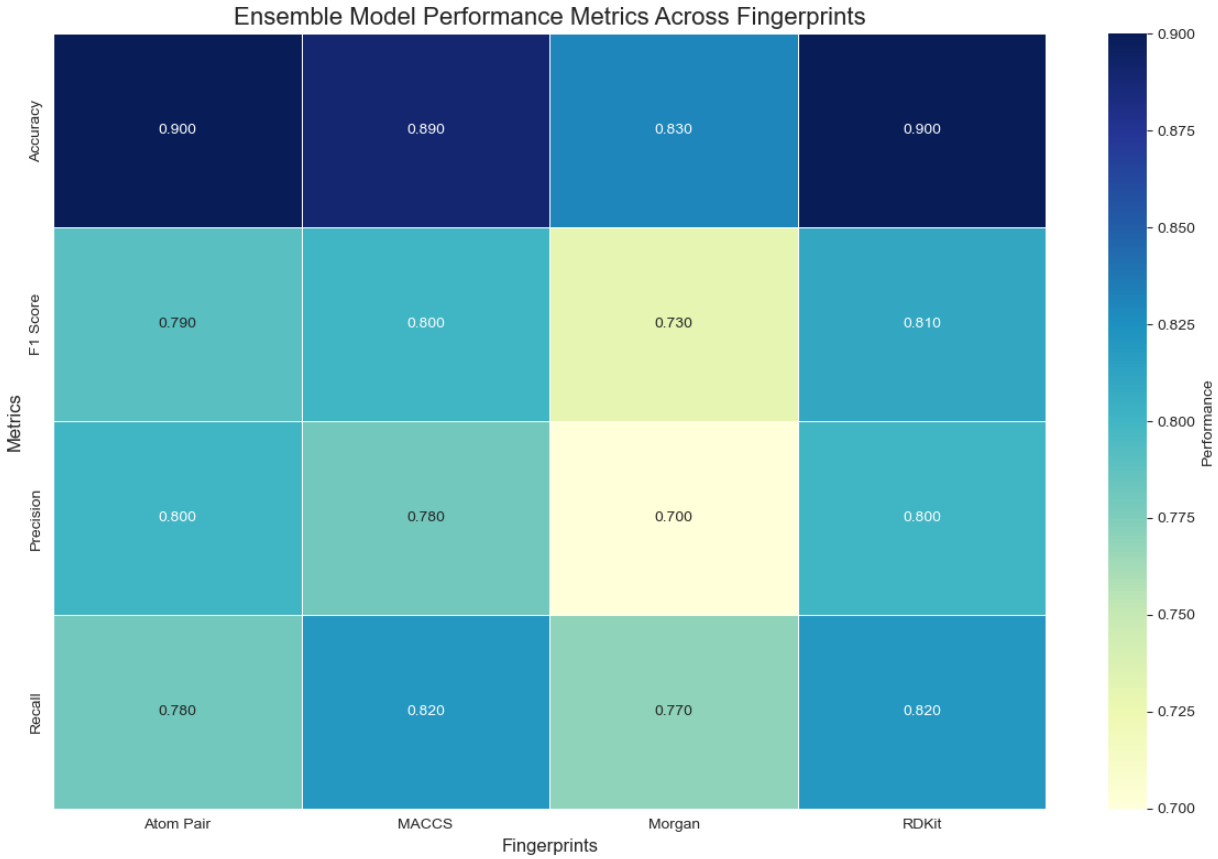


Figure 8: a) Top 30 important features of Morgan **b)**Highest evaluation metrics for each fingerprint

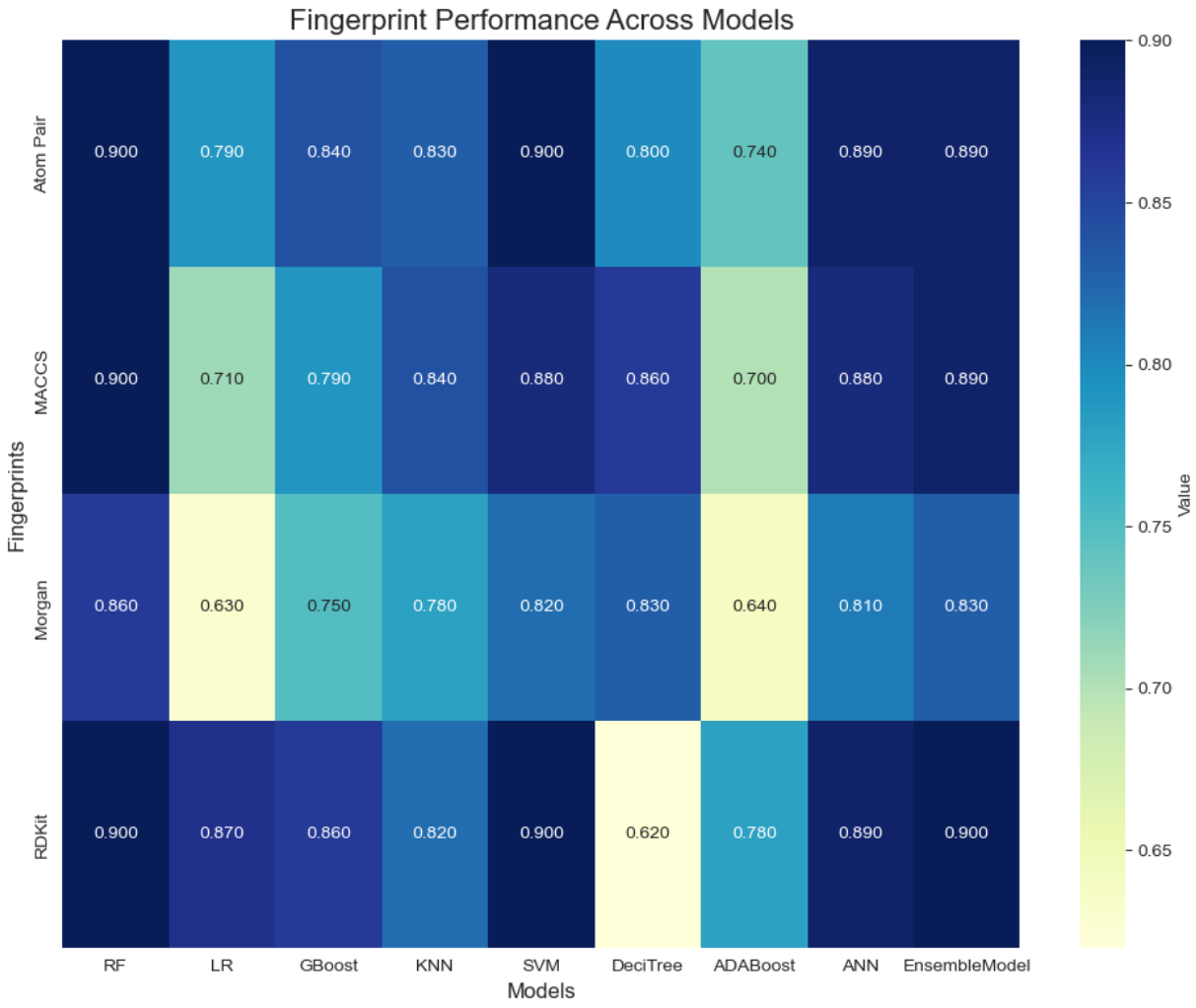


Figure 9 Model Evaluation heatmap

8.1.2 Limitations of the Dataset

While the dataset compiled from the ChEMBL Database provided valuable insights for the machine learning (ML) model in the drug discovery process for **Leishmaniasis**, certain limitations were encountered that posed challenges during model development. These limitations, primarily stemming from the scarcity and quality of available data, are discussed below:

1. Limited Availability of High-Quality Data for Leishmania

- **Obscurity of Research**: Leishmaniasis is a neglected tropical disease, and as a result, fewer research studies have been conducted on this topic compared to other more prevalent diseases such as cancer or malaria. The **ChEMBL Database**, while extensive,

contains only a limited number of high-quality experimental results related to Leishmania.

- **Sparse Data**: The specific focus on *Leishmania* compounds means there were fewer molecules with well-documented and reliable bioactivity data. This constrained the variety of compounds available for the study and, in turn, limited the potential training data for the ML model.
- **Inconsistent Data Reporting**: Many of the available data points for Leishmania-related compounds had missing or inconsistent bioactivity measurements. Certain experimental setups did not report crucial bioactivity metrics such as IC50 or EC50, leading to the exclusion of these data points during preprocessing.

2. Data Imbalance

- **Class Imbalance**: A significant challenge was the **imbalance between active and inactive molecules** in the dataset. Due to the scarcity of experimental data on highly active molecules against Leishmania, there was a higher proportion of inactive molecules, which could potentially bias the ML model towards predicting inactivity.
- **Mitigation**: Techniques such as oversampling of the minority class (active compounds) or undersampling of the majority class (inactive compounds) were explored, but the overall dataset remained relatively small after preprocessing. This imbalance can limit the generalizability of the model, particularly when deployed to predict new active compounds.

3. Data Constraints

- **Noise in the Data**: The data collected from the ChEMBL database often contained noise due to inconsistent experimental conditions, variability in measurement techniques, and reporting errors. While steps were taken to remove duplicates and irrelevant data, some noise remained in the final dataset.
- **Limited Feature Variety**: While dozens of columns were initially available in the raw dataset, most of these features were discarded due to irrelevance or poor data quality, leaving only four essential columns for analysis (SMILES, Standard Type, Standard Value, and Standard Units). This reduction in feature variety limits the complexity and depth of feature engineering that could be used to improve model accuracy.

4. Preprocessing Challenges

- **Standardization of Units**: Although unit conversion was successfully performed for bioactivity measurements (e.g., converting millimolar [mM] to nanomolar [nM]), certain data points could not be converted due to missing or ambiguous unit information. This resulted in the loss of potentially useful data.
- **Missing Data Imputation**: The imputation of missing values, particularly using techniques like the K-Nearest Neighbors (KNN) imputer, may have introduced some level of approximation that could affect the accuracy of the final predictions. Imputation is inherently a process of estimation, and this could lead to inaccuracies in the final model.

5. Mitigation of Data Limitations

Despite the challenges, various approaches were employed to mitigate the limitations:

- **Testing Multiple Fingerprints**: Different molecular fingerprinting techniques were tested, including **Morgan fingerprints**, **MACCS keys**, and **RDKit descriptors**, to capture the molecular structure and properties of compounds in diverse ways. These fingerprints provided rich representations of molecules that helped in improving the model's ability to learn the relationships between chemical structure and bioactivity.
- **Use of Multiple Classifiers**: Several machine learning classifiers were explored, including **Random Forest (RF)**, **Support Vector Machine (SVM)**, and **Gradient Boosting**. The use of **ensemble methods** like **Voting Classifiers** allowed for a combination of predictions from multiple models, enhancing robustness and mitigating the impact of data constraints.
- **Ensemble Methods**: Ensemble methods like **Voting Classifiers** combined the strengths of different models to improve prediction accuracy and reduce overfitting. By leveraging multiple classifiers, the impact of the limited and imbalanced dataset was minimized, resulting in more reliable predictions.

6. Model Performance

- **Accuracy Achieved**: Despite the aforementioned limitations, the highest accuracy achieved by the ML model was **0.900**, which is a promising result given the constraints of the dataset. This accuracy indicates that the model was able to predict the activity of compounds against *Leishmania* with **90% accuracy**, reflecting its potential to assist in the early stages of drug discovery.

Chapter-9: Conclusions

One of the most significant challenges faced during model training was the **limited availability of data points**. This scarcity arose due to the time-consuming nature of drug discovery and the lengthy process involved in determining a drug's effectiveness against *Leishmania*. The limited dataset presented obstacles for building robust machine learning models, especially in terms of generalisation to new, unseen compounds.

Key Insights and Challenges

- **Limited Dataset**: The relatively small number of available data points constrained the model's ability to generalise well to novel compounds. This limitation made it challenging to capture the full complexity of drug-compound interactions. The lack of diverse and comprehensive bioactivity data for Leishmaniasis, compounded by issues such as class imbalance (i.e., more inactive compounds than active ones), impacted the overall performance of the models.

- **Data Constraints**: The quality and quantity of data directly influence the performance of machine learning models. The smaller dataset meant that overfitting was a concern, and the model's predictions might not perform as effectively when exposed to new, untested compounds. This highlighted the **need for larger and more comprehensive datasets** in future research to build more accurate and generalisable models.

- **Mitigation Strategies**: Despite these challenges, the use of **multiple instances of fingerprinting techniques** and **thorough feature testing** helped mitigate some of the limitations. By experimenting with different fingerprint types (such as **RDKit**, **Morgan**, and **MACCS**) and testing different classifiers, it was possible to optimise the performance of the machine learning models given the data constraints.

Optimal Model Selection

After extensive testing of various machine learning classifiers and molecular fingerprints, **Random Forest (RF) with RDKit fingerprints** was determined to be the **best possible solution** for identifying compounds with anti-Leishmania activity. Key reasons for this choice include:

- **RDKit Fingerprint Effectiveness**: RDKit fingerprints provided a detailed and informative representation of molecular structures, enabling the model to better capture the relevant chemical features linked to biological activity. This allowed for more accurate predictions of whether a compound was active against *Leishmania*.

- **Random Forest Classifier**: The Random Forest classifier demonstrated strong performance, likely due to its robustness against overfitting and its ability to handle imbalanced datasets. By aggregating predictions from multiple decision trees, Random Forest provided improved accuracy and reliability, even with limited data points.

- **Performance**: The final model achieved a **highest accuracy of 0.900**, indicating that it was able to predict the bioactivity of compounds with 90% accuracy. This result is promising, particularly in the context of the dataset limitations and serves as a foundation for future improvements.

Future Directions

- **Expanding Dataset**: To further improve model robustness and generalisation, future research should focus on expanding the dataset with more high-quality bioactivity data related to Leishmania. This could be achieved through **collaboration with research institutions**, **data sharing**, or the development of novel experimental techniques that accelerate the drug discovery process.

- **Feature Engineering**: Additional feature engineering, such as incorporating more molecular descriptors, could enhance the model's predictive power. **Advanced imputation techniques** and handling of class imbalance may also further improve the quality of the dataset.

- **Further Model Optimisation**: Exploring more advanced machine learning techniques, such as **deep learning** or **ensemble learning**, could provide additional gains in accuracy, especially as more data becomes available. Integrating **chemoinformatics** tools and domain-specific knowledge will also likely lead to better model outcomes.

Chapter-10: References/Bibliography

- https://www.ebi.ac.uk/chembl
- https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2021.796534/full
- https://ijece.iaescore.com/index.php/IJECE/article/view/31562/16900
- https://www.mdpi.com/2218-273X/11/12/1750
- https://www.rdkit.org/docs/GettingStartedInPython.html
- https://numpy.org/doc/stable/reference/index.html#python-api
- https://pandas.pydata.org/docs/user_guide/index.html
- https://scikit-learn.org/stable/user_guide.html
- https://xgboost.readthedocs.io/en/latest/python/python_api.html#xgboost.XGBClassifier
- https://keras.io/guides/sequential_model/
- https://www.tensorflow.org/guide/keras/training_with_built_in_methods
- https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html