Drug Discovery for Mycobacterium Tuberculosis: A Synergistic Approach using QSAR and Machine Learning

Lasyapriya Bharadwaj K[1], Shashank R[1], Vemula Yashodha[1], K Tappan Chengappa[1], S Lalitha[1]

[1]Department of Electronics and Communication Engineering
Amrita School of Engineering, Bengaluru
Amrita Vishwa Vidyapeetham, India
klasya2005@gmail.com, shashankr8724@gmail.com, yashodhavemula09@gmail.com, tapan.chengappa123@gmail.com, s_lalitha@blr.amrita.edu

Abstract. There are significant challenges faced in the field of drug discovery, the accurate prediction of bioactive molecules' interaction with a respective biological target is one of them. Latent Tuberculosis Infection (LTBI) affects approximately 1.7 billion people. Individuals with LTBI have a 5-10% lifetime risk of developing active TB, and current treatments for LTBI show 60-90% efficacy. With drug resistance increasing, the WHO recommends systematic testing and treatment of LTBI in high-risk populations. Dihydrofolate reductase (DHFR) is a key enzyme in Mycobacterium tuberculosis (Mtb) that plays an essential role in DNA replication and re-pair. Inhibiting DHFR disrupts the folate pathway, making it a promising target for antitubercular drug discovery. The proposed research focuses on identifying active inhibitors of the DHFR enzyme using Quantitative Structure-Activity Relationship (QSAR) modeling. The approach includes generating molecular descriptors, developing a QSAR model, and applying various machine learning models to predict the bioactivity of molecules against DHFR using pIC50 values. Unlike previous studies that typically use Random Forest Regressor, this research compares 42 regression models through the Lazy Predict module, selecting the optimal model based on R^2 and RMSE scores. Feature selection was performed using the variance threshold method with a threshold of 0.16 to enhance computational speed and reduce overfitting. Gradient Boosting Regressor, after hyperparameter tuning, yielded an R^2 score of 0.40 and an RMSE of 1.24. This systematic approach improves drug discovery efficiency by accurately predicting potential DHFR inhibitors, offering insights into combating drug-resistant LTBI.

Keywords: QSAR, Mycobacterium Tuberculosis, DHFR inhibitors, Drug Discovery, Machine Learning, Gradient Boosting Regressor.

1 Introduction

Human medical needs are never exhausted, despite humanity's efforts, cures for many diseases remain unknown. With the evolution of humanity, there's an evolution of many diseases too. Drug discovery plays a pivotal role in humanitarian survival and evolution. The discovery of a drug makes it possible to develop treatment that might control the symptoms, slow down the progression of a few diseases, and may also potentially cure some. This process however involves a lot of upscale investment and sometimes time, which always may not lead to accurate results, primarily during the hit identification phase of the therapeutic compounds.[1] Especially at the present time when advanced technologies, such as artificial intelligence or machine learning, continue to evolve, this process can be significantly accelerated both, in terms of time and, therefore, in terms of money as well.

Among these approaches one of the most productive that can be used in this process is the QSAR, which focuses on finding the relationship between bioactivity and the chemical structure and is one of the methods for predicting the potency and selectivity of a drug for a certain target [2]. The very basis of QSAR is in the belief that biologic activity is directly related to chemical structure of the compound. Such structural information is then translated into molecular descriptors, and the QSAR model defines the quantitative associations between these descriptors and distinct biological activities [3][4]. There has been intensive research in finding a potent drug for Tuberculosis (TB). Among many potential targets, DHFR has gained significance due to its potential target for antitubercular disease.

The proposed work makes use of the QSAR approach for a comprehensive understanding and behavior of the DHFR target protein and its ligand which will eventually help in developing and optimizing the modulators for the target protein. Different descriptors of molecules like Molecular Weight (MW), ligand efficiency, and the logarithm of the partition coefficient (LogP) are used to study the physiochemical properties of ligands [5].

Moreover, to evaluate the binding affinity of ligands and target proteins, a regression model known as Gradient Boosting Regressor is applied. This helped in knowing the important characteristics of molecules concerning their binding potency, which was mainly used to classify if the drug is active or inactive.

The main contributions are as follows:

1. Investigating the target proteins that are responsible for Tuberculosis such as DHFR.
2. The analysis of various chemical and biological relationships of the compound using the QSAR technique.
3. Integrating machine learning on the ChEMBL [6][7] dataset to derive valid conclusions on the activities of the receptor.

2 Related Works

Several studies and research have been done in the domain of drug discovery using a technique called QSAR. A study discusses different issues regarding operations and science involving the development process which can result in improving the process of drug discovery [5].

It is important to note that the notion of QSAR was first proactively mentioned. The application of QSAR modeling has led to be useful strategy to develop an issuing predictive model of contaminant transfer across the placenta during development as discussed in the study N.Tahiri (2022) et. Al. S.Kwon et. al. (2019) has used an ensemble-based Machine Learning model to overcome the constraints and deliver some genuine predictive values by QSAR models [2]. One can predict the biological toxicological properties of a novel chemical from only the chemical structure, obviating the long-time molecular docking procedure. This insightful approach has resulted in the application of QSAR being very effective as presented in the article by J.Mao et al. (2021) [8].

Virtual screening is common for QSAR predictions, among them, 3D-QSAR models have been successfully applied to the prediction of ligand interaction for further help in the design of potent anti-tubercular agents [9][10]. Feature-based interactions acknowledge the chemical components of the drug and that of the target to determine the feature vectors [11]. These models help in the screening of synthetic candidates, which are then subjected to molecular docking. This docking is done to screen potential drug molecules that might combat MDR-TB. The objective of this study was to build the QSAR model which would help to quantify the structure of 2,4-diquinoline derivatives and estimate their individual activities against Mycobacterium tuberculosis more effectively. The paper describes preparation of twenty-four derivatives of 2/4-quinolinecarbaldehyde ring-substituted. All these compounds were synthesized in one to two steps at good yields and several of these compounds have their promising inhibitory activity against Mycobacterium tuberculosis. Among the synthesized derivatives, anti-tuberculosis activities have been exhibited by compounds such as 4a, 7c, and 8a, reaching an inhibition of 99% within very low concentrations of 3.125 μg/mL against the drug sensitive M. tuberculosis H37Rv strain. The study adopted 3D-QSAR analysis to elucidate the structure-activity relationship of the compounds [13].

The QSAR model classified the molecules correctly with their activity, giving them much value in predicting new derivatives' efficacy [14]. M.Mizera et. al. (2021) aimed at developing an ML approach that could effectively identify new active compounds within the glucagon receptor family [15]. Leaning in immersive technologies, K. Stergiou et. al. (2022) applied epidemic forecasting and aids to speed up drug and antibiotic discovery processes [16].

Conclusion thereby, with the applications of QSAR models, particularly 3D-QSAR, an effective predictive approach during drug discovery significantly aided in identifying potential therapeutic agents, as exemplified by anti-tubercular. Cutting-edge developments in machine learning and computational approaches further improved the accuracy and efficiency of the models, accelerating the drug development process.

3 Methodology

In this section, the development of a QSAR based machine learning model specifically designed to find the pIC50 values based on the chemical structure and to classify them as active/inactive inhibitor of the DHFR protein is outlined. The primary aim is to predict pIC50 values based on the molecular fingerprints generated by the PaDEL descriptor [17]. The upcoming subsections provide a detailed account of the data collection steps, data pre-processing, model creation and evaluation metrics employed throughout the proposed work.

3.1 Flow of the proposed workflow

The approach taken to implement the QSAR based machine learning model for drug discovery is illustrated in Fig. 1, visualized in the form of a block diagram.

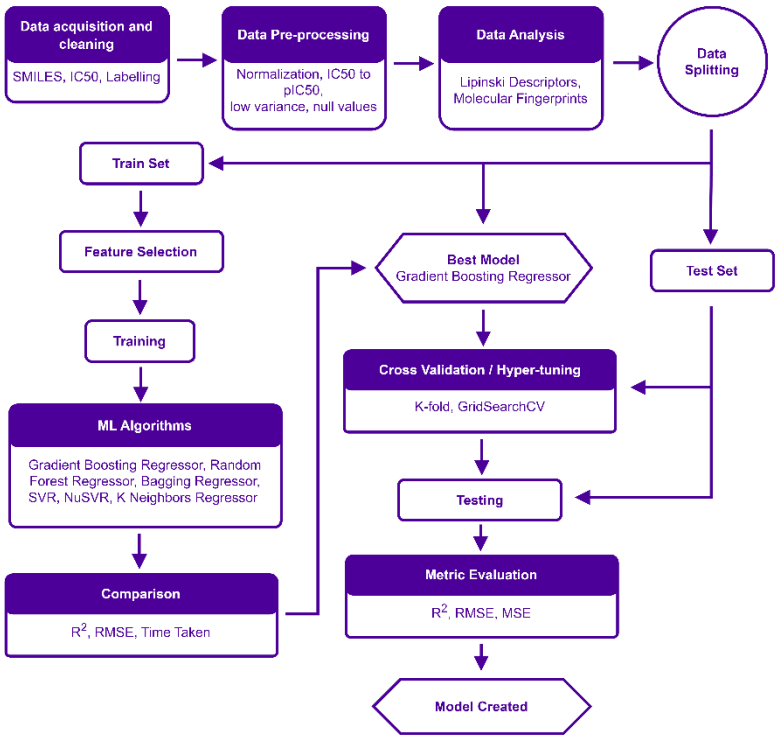

Fig. 1. Workflow diagram

3.2 Dataset Acquisition and Cleaning

The dataset can be found in the well-organized and maintained ChEMBL [7] bioactivity database which includes bioactive molecules for specific target proteins. ChEMBL helps bridge the gap between genome information and the creation of new drugs.

It contains data on the interactions of small molecules with their protein targets, the effects of these substances on individual cells and entire organisms, and data on absorption, distribution, metabolism, excretion, and toxicity (ADMET). Information on bioactivity (such binding constants and pharmacology) is all stored in ChEMBL. To demonstrate connections between molecular targets and published assays, the bioactivity data is labeled.

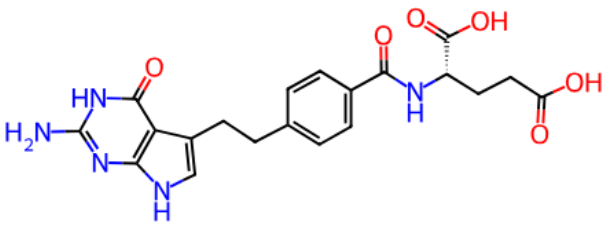

Fig. 2. Chemical Structure of a DHFR inhibitor for Mycobacterium Tuberculosis (CHEMBL225072 PEMETREXED) [7]

The database was accessed through the python library chembl-webresource-client [6] via the python's package manager called pip. A new target client was created and the target DHFR was passed to the target client that is created using the chembl-webresource-client. The target client returned 35 target proteins in which the target of interest was "Homo sapiens" with the chembl_id CHEMBL202. A new activity client is created using the chembl-webresource-client and the target chembl_id CHEMBL202 is passed as the filter query to get the bioactivity data of the target protein.

The main features selected from the bioactivity data are "molecule_chembl_id", "canonical_smiles" and "standard_value". The "canonical smiles" column in the dataframe represent the ligands.

The dataset that was obtained was pre-processed first to handle the missing values in the "canonical_smiles" and "standard_value" column. Then the compounds are labelled as being "active" (< 1000 nM), "inactive" (> 10000 nM) or "intermediate" (1000 nM - 10000 nM) in IC50 units. It was set to a column named 'bioactvity_class' that would be used for further analysis. The curated data frame consisted of the additional 'bioactivity_class' column.

The dataset is finally saved as a .csv (comma separated values) file named dhfr_bioactivity_data.csv for future easy and efficient access. The first 4 rows of the dataframe are shown in Fig. 3.

	molecule_chembl_id	canonical_smiles	standard_value	bioactivity_class
0	CHEMBL25817	CCc1cc(Cc2cnc(N)nc2N)cc(CC)c1O	120.0	active
1	CHEMBL277176	CC(C)(C)c1cc(Cc2cnc(N)nc2N)cc(C(C)(C)C)c1O	61.0	active
2	CHEMBL279455	CCCc1cc(Cc2cnc(N)nc2N)cc(CC)c1OC	44.0	active
3	CHEMBL23338	C/C=C/c1cc(Cc2cnc(N)nc2N)cc(OC)c1OC	93.0	active
4	CHEMBL36083	CN(Cc1cnc2nc(N)nc(N)c2n1)c1ccc(C(=O)NC(CCCN)C(...	2.5	active

Fig. 3. The first 4 rows of the dataset

3.3 Data Preprocessing

The "standard_value" is normalized to improve the computation speed and interpretability of the models, enhancing the stability of numerical computations. A threshold cap of 1×10^8 is applied as any IC50 value greater than 1×10^8 results in a negative pIC50 value.

The "standard_value" column represents the potency of the compounds and is a measure of their biological activity against the DHFR receptor. It is derived from the ChEMBL database and is typically expressed as negative logarithm of the molar concentration required for 50% inhibition (IC50), resulting in units such as nM (nanomolar) or pIC50. The lower the standard_value, the higher the potency or binding affinity of the compound to the DHFR receptor.

The pIC50 represents the negative logarithm (of base 10) of the IC50 value where p stands for negative logarithm. This helps us in easily converting the two values i.e. pIC50 and IC50 to study and analyze a large concentration range. The equation (1) below displays the interconversion of IC50 to pIC50:

$$\text{pIC50} = -\log10(\text{IC50}) \tag{1}$$

3.4 Data Analysis

Descriptor calculation and exploratory analysis performed. The next step performed was extracting the SMILES string with the longest length from a column named canonical_smiles in a DataFrame (df) and storing these longest SMILES strings in a new pandas Series called smiles.

The first objective was to predict the potency of small molecules to the DHFR receptor. The "canonical_smiles" column is originally the ligand that consists of the SMILES strings. The Lipinski Descriptors are computed using the "canonical_smiles" with the help of RDKit using the python's package manager called pip. The molecular descriptors "MW", "LogP", "HBA" (Hydrogen Bond Acceptors) and "HBD" (Hydrogen Bond Donors) are computed. The final dataframe was prepared by joining the raw dataframe along with the curated one using "molecule_chembl_id" as the dataframe since the "standard_value" column was used as the target column. The molecules with the bioactivity_class "intermediate" are dropped.

A molecule.smi file is created with the data of "canonical_smiles", "molecule_chembl_id" columns, this file is then used to compute the molecular fingerprints of the molecules. This is done with the help of PaDEL descriptor. PaDEL-Descriptor

is another stand-alone software application that has its unique characteristic in the sense that it provides a fast and accurate method for the computation of molecular descriptors and fingerprints. The fingerprint data is then attached to the dataset, the features HBD and HBA are dropped due to high correlation. The final dataset is generated and saved as *dhfr_bioactivity_data_pIC50_fingerprints.csv*. The first 4 rows of the dataframe after the calculation of molecular descriptors are shown in Fig. 4.

	molecule_chembl_id	canonical_smiles	bioactivity_class	MW	logP	HBD	HBA	pIC50
0	CHEMBL25817	CCc1cc(Cc2cnc(N)nc2N)cc(CC)c1O	active	272.352	2.0622	3	5	6.920819
1	CHEMBL277176	CC(C)(C)c1cc(Cc2cnc(N)nc2N)cc(C(C)(C)C)c1O	active	328.460	3.5324	3	5	7.214670
2	CHEMBL279455	CCCc1cc(Cc2cnc(N)nc2N)cc(CC)c1OC	active	300.406	2.7553	2	5	7.356547
3	CHEMBL23338	C/C=C/c1cc(Cc2cnc(N)nc2N)cc(OC)c1OC	active	300.362	2.2821	2	6	7.031517
4	CHEMBL36083	CN(Cc1cnc2nc(N)nc(N)c2n1)c1ccc(C(=O)NC(CCCN)C(...	active	439.480	0.1425	5	10	8.602060

Fig. 4. The first 4 rows of the data frame after calculation of the Lipinski descriptors

The illustration shown in Fig. 5 visualizes the molecular fingerprint calculations for different ligands and their docking point. The PaDEL descriptor software can calculate 797 descriptors and 10 types of fingerprints.

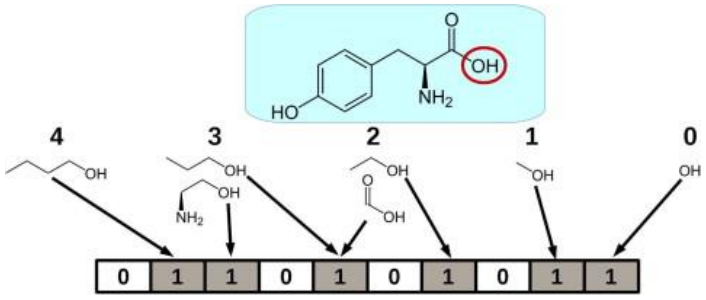

Fig. 5. Simple illustration of molecular fingerprints for different ligands and docking points [18]

3.5 Model Selection

The next step was model selection, feature selection was done on the dataset, to avoid overfitting and increasing computation speed by dimensionality reduction. The low variance threshold method was used in this case with a threshold of 0.16 (0.8*(1-0.8)). All the low variance features with more than 80% repetition are removed i.e. a feature which contains only 0's 80% of the time or only 1's 80% of the time are retained.

The resultant dataframe consisted of 988 rows with the columns "MW", "LogP", "standard_value" and 123 fingerprint descriptors. 80% of the dataset is used for training and the rest for testing purposes, with a random state of 42. The data was passed to a Lazy Regressor from Lazy Predict by a pip install. Data is input in the command line by installing a package called Lazy Predict, which already comes with installation of pip. Essentially, Lazy Predict can create several baseline models within mere lines

of code by simplifying how one creates baseline models to make quick model comparisons without parameter tuning.

It was found that Gradient Boosting Regressor had the highest R^2 Score with 0.41 and had the lowest RMSE of 1.23. A new regression model was built with the best hyper-tuning parameters using K-Fold Cross Validation and Hyper-tuning techniques such as Grid Search CV. The regression model was trained with respect to the training data and the corresponding predictions were made.

Gradient boosting is a machine learning technique which builds on boosting in functional space, with the target being pseudo-residuals instead of residuals as it is in traditional boosting techniques. It provides a prediction model in the form of a combination of weak prediction models, meaning that it allows for models that make little assumptions on the data basis, which often are very simple decision trees. When the decision tree is the weak learner, the resultant algorithm is known as gradient-boosted trees and this algorithm normally has better accuracy than random forest. When compared with other boosting methods, the gradient-boosted trees model is observed to be built in stages, while generalizing the other methods by allowing optimization of an arbitrary differentiable loss function [1][2].

4 Result and Analysis

This section discusses in detail the result obtained for the proposed work and exploratory data analysis of the pre-processed dataset which is created to create the Gradient Boosting Regressor model.

From the obtained results, it is understood that the Gradient Boosting Regressor model performed modestly in predicting the ligand binding affinity to the DHFR receptor. The RMSE indicates that there is room for improvement in reducing the prediction errors, while the R^2 score implies that the model describes approximately 40.08% of the variance in the binding affinity values. It might be necessary to go a step further and perform the secondary level analysis and data modelling which would make the model more refined and closer to real characteristics in the data. The methodology proposed has demonstrated an effective way of applying ML techniques to predict the activity of DHFR using molecular descriptor and provided a satisfactory level of predictive performance. After applying the K-Cross Validation and Grid Search CV for hyper-tuning the model, the resultant R^2 value is 0.40085278761877163.

4.1 Exploratory Data Analysis

The heatmap shown in Fig. 6 (a) is used to study the correlation between MW, LogP, HBD, HBA and pIC50. The features are well related to the target feature pIC50, with MW, HBD and HBA having a positive correlation and LogP having a negative correlation, negative correlation here signifies that increase in LogP decreases pIC50. Let's discuss the interpretation of the graph:

The x-axis and y-axis represent the features (MW, LogP, HBD, HBA and pIC50) for which the correlation has been computed. The gradient legend that is positioned on the right side represents the intensity of the correlation. A combination of heat maps is used, where the heat map illustrates the correlation as follows: a dark red color is assigned to the higher correlation, positive correlation; a dark blue color is given to represent the lower correlation, negative correlation. The features HBA and HBD are dropped as these features are highly correlated ($r > 0.6$ or $r < -0.6$), Corr (HBD, HBA) = 0.69, Corr (HBA, MW) = 0.71 and Corr (HBD, LogP) = -0.63.

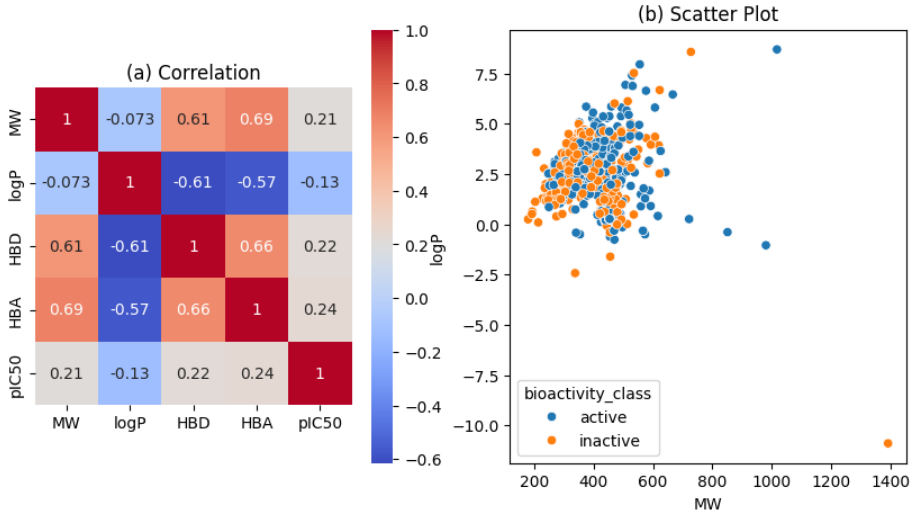

Fig. 6. (a) Correlation matrix of MW, LogP, HBD, HBA and pIC50 **(b)** Scatterplot analysis of LogP vs MW

The scatter plot shown in Fig. 6 (b) is used to study the relation between MW, LogP visually along with pIC50 values of compounds helping us to find trends in the variables and analyze the results to get insights of the properties and activities of the compounds against DHFR receptor.

hue='bioactivity_class' assigns different colors to the data points based on the 'class' column. This helps to distinguish different classes or categories of compounds.

The LogP vs MW graph as shown in Fig. 6 (b) is a scatter plot that visualizes the relationship between the LogP and MW of compounds.

It can be observed that despite molecules having similar MW and LogP values they are not well separated and compounds having the similar structure have similar MW and LogP values, this leads to the inference that the compounds can be classified as active or inactive inhibitors only based on the pIC50 value.

The bioactivity class bar graph as shown in Fig. 7 (a) is plotted with frequency displaying the active and inactive samples. There are 625 rows classified with a bioactivity class of "active" and 363 rows classified with a bioactivity class of "inactive", summing up to a total of 988 rows.

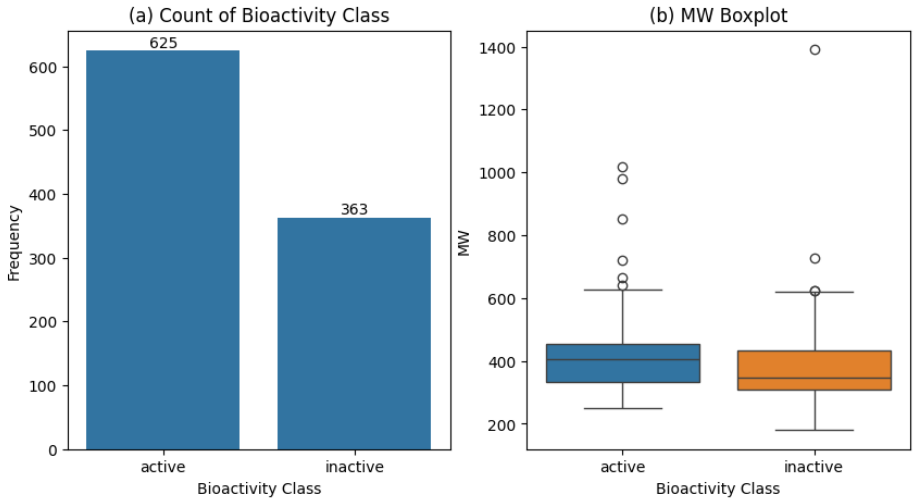

Fig. 7. (a) Frequency vs Bioactivity bar graph **(b)** Boxplot of the number of active and inactive categories for MW

The bioactivity class boxplot is plotted with MW as shown in Fig. 7 (b) with respect to the "bioactivity_class". The boxplot gives a visual representation of the outliers present in the dataset. The boxplot has a median of 406.339, lower whisker of 154.396 and upper whisker at 634.477 for the "active" bioactivity_class and a median at 347.378, lower whisker at 118.964 and upper whisker at 621.377 for the "active" bioactivity_class. There are 6 outliers with "active" bioactivity_class and 3 outliers with "inactive" bioactivity_class.

From the boxplot shown in the Fig. 7 (b), it can be observed that both the "active" and "inactive" bioactivity_class have nearly similar MW, this signifies that the drugs which are used for DHFR inhibition have same number of chemical atoms. Thus, this feature alone can't be used to classify drugs as "active" or "inactive".

The bioactivity class box plot is plotted with LogP as shown in Fig. 8 (a) displaying the active and inactive samples. The boxplot gives a visual representation of the outliers present in the dataset. The boxplot has a median at 2.107, lower whisker at -1.780 and upper whisker at 5.982 for the "active" bioactivity_class and a median at 2.562, lower whisker at 1.833 and upper whisker at 6.575 for the "active" bioactivity_class. There are 7 outliers with "active" bioactivity_class and 5 outliers with "inactive" bioactivity_class.

From the boxplot shown in the Fig. 8 (a), it can be observed that both the "active" and "inactive" bioactivity_class have nearly similar LogP, this signifies that the drugs which are used for DHFR inhibition have same solubility in different solvents and the drug's ability to cross cell membranes. Thus, this feature alone can't be used to classify drugs as "active" or "inactive".

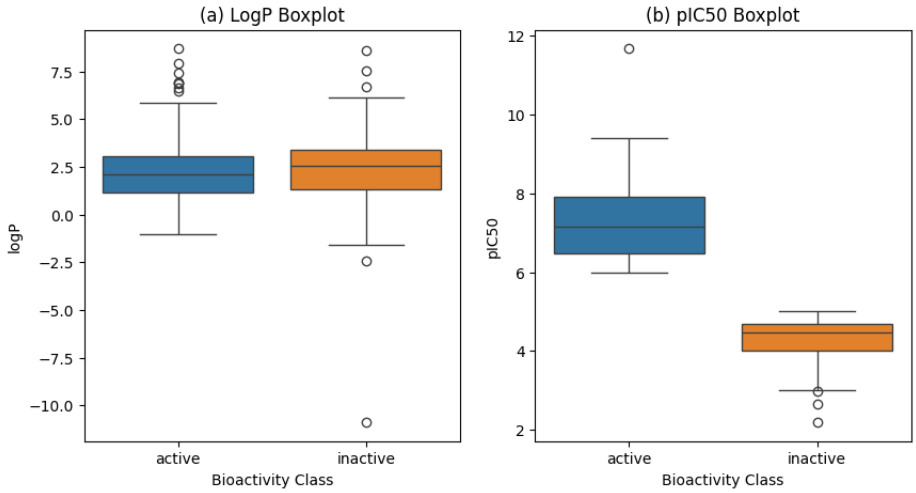

Fig. 8. (a) Boxplot of the number of active and inactive categories for LogP **(b)** Boxplot of the number of active and inactive categories for pIC50

The bioactivity class boxplot is plotted with pIC50 value as shown above in Fig. 8 (b) displaying the active and inactive samples. The boxplot gives a visual representation of the outliers present in the dataset. The boxplot has a median at 7.142, lower whisker at 4.290 and upper whisker at 10.099 for the "active" bioactivity_class and a median at 4.481, lower whisker at 2.983 and upper whisker at 5.694 for the "active" bioactivity_class. There are 1 outlier with "active" bioactivity_class and 3 outliers with "inactive" bioactivity_class.

From the boxplot shown above in the Fig. 8 (b), it can be observed that both the "active" and "inactive" bioactivity_class don’t have similar pIC50 values, this signifies that the drugs which are used for DHFR inhibition have different pIC50 values for different bioactivity_class. This is a target feature.

The dataset is preprocessed, and relevant molecular descriptors are calculated using RDKit paticularly LogP and MW. Additionally, ligand efficiency information is extracted from dictionary-like strings and transformed into numerical features. Categorical columns (canonical smiles) are converted to numerical representations (molecular fingerprints) using the PaDEL Descriptor.

The features MW, LogP and other 124 molecular fingerprint descriptors are used to compute the pCI50 values using a regressor model. Here MW is the measure of the sum of the atomic weights of all the atoms in a molecule. Generally, MW provides information about the size and complexity of the compound. The LogP is the compound’s hydrophobicity measure, indicating its tendency to dissolve in lipids (hydrophobic environments) versus water (hydrophilic environments). The logarithmic scale is used to compress the wide range of possible values into a more manageable range.

4.2 Result

The traditional way of choosing the best model for a machine learning project is quite difficult and time consuming. To avoid this problem, a module called Lazy Predict can be used. A comparison with 42 regression models is done using the Lazy Regressor function which can be imported from the Lazy Predict module.

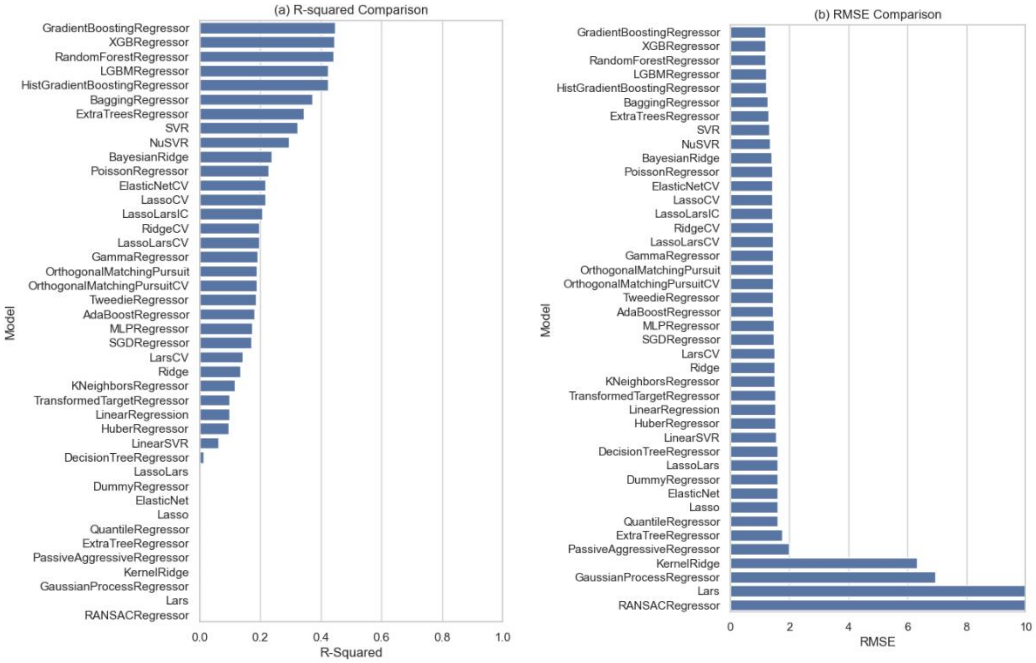

Fig. 9. (a) R2 scores **(b)** RMSE scores of various regression models using the Lazy Predict module

The comparison which is visualized in Fig. 9 (a) shows the R^2 scores of 42 regression models which are compared using the Lazy Regressor function from the Lazy Predict module, from this it can be concluded that the Gradient Boosting Regressor has the highest R^2 score (higher the R^2 score better fitting of data to the model) and is best for the dataset which is created. The comparison which is visualized in Fig. 9 (b) shows the RMSE scores of 42 regression models which are compared using the Lazy Regressor function from the Lazy Predict module, from this it can be concluded that the Gradient Boosting Regressor has the highest RMSE score (lower the RMSE score better fitting of data to the model and low deviation from the actual value) and is best for the dataset which is created.

Hyper-tuning of the Gradient Boosting Regressor is done using K-Fold Cross Validation and Grid Search CV methods and a model is created to predict the pIC50 val-

ues of medicinal drugs which then can be classified as active or inactive inhibitors of DHFR. The results obtained are as such:

1. **R^2 Score**: The R^2 score is the coefficient of determination showing variation between in the target variable (binding affinity) as shown in the model. It varies from 0 to 1, with greater values showing better predictive power. Ranges from 0 to 1 where larger than 0.5 represent model with an excellent ability to predict. In this research, the obtained R^2 score was 0.40085278761877163. An R^2 score of 1 tells us that the values predicted from the model best fits the data while an R^2 score of 0 means the model is unable to capture any form of relation between the predictors and the target variable.
2. **RMSE Score**: According to the number of pIC50 in the test set, the RMSE gives out the average of the square of the differences between the predicted pIC50 values and the actual values. In general, the smaller the RMSE values the better model prediction since RMSE is the total measurement of the quality of predictive technique. In this case, the RMSE that was achieved was 1.2441623224685912.

5 Conclusion and Future Scope

The proposed approach is particularly promising for drug discovery, where it can enhance predictive accuracy by modeling intricate, non-linear relationships which showcased the immense potential of integrating machine learning methodologies in the realm of drug discovery, specifically in predicting pIC50 values. Through a systematic approach, starting from meticulous data preprocessing of bioactivity data to exploratory analysis and finally predictive modeling, the importance of both traditional machine learning and modern ensemble methods has been illuminated.

In this work, QSAR approach has been applied to target the DHFR protein, a critical enzyme in Tuberculosis treatment. By using machine learning techniques, predictive models were developed to identify potential inhibitors of DHFR. The integration of Lazy Predict, an ensemble-based framework, provided a baseline for rapid model comparison, enabling us to select the most promising algorithms for further refinement. Hyper-tuning parameters improved the model's performance, enhancing prediction accuracy and robustness.

This approach demonstrates the potential of computational drug discovery to accelerate the identification of effective compounds, saving time and resources in early-stage drug development. This advancement is expected to open new avenues and provide deeper insights in the field of drug discovery.

The objective is to evaluate different combinations of hypotheses while preserving the ensemble effect. Ensemble learning is where the data is trained with many models, and their results together are combined. Both theoretical and empirical research studies have confirmed that the use of ensemble learning tends to outperform single or individual models in terms of accuracy. By combining multiple weak models (inducers), a more robust ensemble model can be formed.

Additionally, techniques like advanced feature engineering that includes incorporating molecular descriptors and fingerprints from the wider range of chemical data-

bases can improve model accuracy, applying deep learning neural networks capable handling sequential chemical structures can be integrated for more accurate representation and prediction of molecular activities. Furthermore, validation of computational predictions using vitro and in vivo essays to ensure biological relevance and efficacy against Tuberculosis.

References

1. Odugbemi, A. I., Nyirenda, C., Christoffels, A., & Egieyeh, S. A. (2024). Artificial intelligence in antidiabetic drug discovery: The advances in QSAR and the prediction of α-glucosidase inhibitors. *Computational and Structural Biotechnology Journal*, *23*, 2964–2977.
2. Kwon, S., Bae, H., Jo, J., & Yoon, S. (2019). Comprehensive ensemble in QSAR prediction for drug discovery. *BMC Bioinformatics*, *20*(1).
3. Hansch, C., Maloney, P. P., Fujita, T., & Muir, R. M. (1962). Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature*, *194*(4824), 178–180.
4. Achary, P. G. R. (2020). Applications of Quantitative Structure-Activity Relationships (QSAR) based Virtual Screening in Drug Design: A Review. *Mini-Reviews in Medicinal Chemistry*, *20*(14), 1375–1388.
5. Sharma, Vandana & Sarkar, Oshmita & Mishra, Sushruta & Sinha, Satyam. (2023). QSAR Approach for Drug Discovery Targeting the Glucagon Receptor Using Machine Learning. 702-706.
6. Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L., & Overington, J. P. (2015). ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Research*, *43*(W1), W612–W620.
7. Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., De Veij, M., Ioannidis, H., Lopez, D. M., Mosquera, J. F., Magarinos, M. P., Bosc, N., Arcila, R., Kizilören, T., Gaulton, A., Bento, A. P., Adasme, M. F., Monecke, P., Landrum, G. A., & Leach, A. R. (2023). The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, *52*(D1), D1180–D1192.
8. Mohs, R. C., & Greig, N. H. (2017). Drug discovery and development: Role of basic biological research. *Alzheimer S & Dementia Translational Research & Clinical Interventions*, *3*(4), 651–657.
9. Mathew, B., Scotti, M. T., Herrera-Acevedo, C., Dev, S., Rangarajan, T., Kuruniyan, M. S., Naseef, P. P., & Scotti, L. (2021). Development of 2D, 3D-QSAR and pharmacophore modeling of chalcones for the inhibition of monoamine oxidase B. *Combinatorial Chemistry & High Throughput Screening*, *25*(10), 1731–1744.
10. Ani, Riv & Deepa, O.s & Manju, B.R. (2023). Ligand Based Virtual Screening of Molecular Compounds in Drug Discovery Using GCAN Fingerprint and Ensemble Machine Learning Algorithm. Computer Systems Science and Engineering. 47. 3033-3048.
11. Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2020). Artificial intelligence in drug discovery and development. *Drug Discovery Today*, *26*(1), 80–93.
12. N. Sukumar, Harishchander Anandaram and Pratiti Bhadra, "Computational Drug Discovery – A Primer" (Ion Cures Press, 2023). ISBN: 979-8850083663

13. Adeniji, S. E., Uba, S., Uzairu, A., & Arthur, D. E. (2019). A Derived QSAR Model for Predicting Some Compounds as Potent Antagonist against Mycobacterium tuberculosis: A Theoretical Approach. *Advances in Preventive Medicine*, *2019*, 1–18.
14. Nayyar, A., Malde, A., Coutinho, E., & Jain, R. (2006). Synthesis, anti-tuberculosis activity, and 3D-QSAR study of ring-substituted-2/4-quinolinecarbaldehyde derivatives. *Bioorganic & Medicinal Chemistry*, *14*(21), 7302–7310.
15. Mizera, M., & Latek, D. (2021). Ligand-Receptor interactions and machine learning in GCGR and GLP-1R drug discovery. *International Journal of Molecular Sciences*, *22*(8), 4060.
16. Stergiou, K. D., Minopoulos, G. M., Memos, V. A., Stergiou, C. L., Koidou, M. P., & Psannis, K. E. (2022). A Machine Learning-Based model for epidemic forecasting and faster drug discovery. *Applied Sciences*, *12*(21), 10766.
17. Yap, C. W. (2010). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, *32*(7), 1466–1474.
18. Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2014). Molecular fingerprint similarity search in virtual screening. *Methods*, *71*, 58–63.