LUNG CANCER DRUG DISCOVERY

Atharva Pandit
atpand@iu.edu
IU Bloomington.

Shubham Patil
shupatil@iu.edu
IU Bloomington

Rahul Ubale
raubale@iu.edu
IU Bloomington

Rahul G.S.
rgomathi@iu.edu
IU Bloomington

ABSTRACT

Lung cancer is the most common cause of cancer related deaths world-wide. Small cell lung cancer (SCLC), the most common type of lung cancer, have both reacted well to chemo-, radiation-, and adjuvant treatments. Surgery to remove the tumor also appeared to be a viable treatment option. However, these treatments had unfavorable side effects, necessitating the development of alternate lung cancer medications such as novel drugs. Activation of Epidermal Growth Factor Receptor (EGFR) protein is a key reason for lung cancer.

Our project aims to compare various regression algorithms such as Random Forest Regressor, Ada-boost Regressor and Bagging Regressor to predict the efficacy of a drug in inhibiting the growth of cancer cells caused by a defect in EGFR protein.

keywords— lung cancer, egfr protein, ic 50, adaboost regressor, random forest regressor, bagging regressor

1. INTRODUCTION

EGFR is protein responsible for helping the cells grow and divide. In case of EFGR-positive lung cancer, a mutation, or a defect in the gene, causes the EGFR to continuously “grow”. This results in uncontrolled cellular proliferation, which is the cause of cancer. Chemotherapy is one of the most effective solutions to cancer, however it comes with its share of side effects such as fatigue, hair loss, and appetite changes. Hence, as an alternative to chemotherapy, different drugs are being tested for their ability to stop the EGFR protein from multiplying.

Every drug is measured in terms of its efficiency in inhibiting the growth of the protein based on certain parameters. In this project, we have used the Inhibitory Concentration (IC 50) value which is a quantitative measure of how much a drug is needed to inhibit a given biological process (in this case multiplying of the EGFR protein) by 50 percent, as a marker for testing the bioactivity of a particular drug. We are using the ChEMBL dataset which is a chemical database of bioactivity molecules to get the canonical notation of molecular formulas of a drug. We then use ‘PubChem’ database to convert the canonical notation to the molecular fingerprint which is nothing but unique binary representation of a molecule. We fed this data into our Machine Learning models and tried to predict the IC50 value for a given chemical. We made a comparison study between 3 machine learning models, Random Forest Regressor, Ada-boost Regressor and Bagging Regressor.

2. METHODOLOGY

The methodology for this project was divided into four parts: -

1. Data Collection
2. Exploratory Data Analysis
3. Molecular Fingerprint Calculation
4. Feature Engineering and Model Building

2.1. Data collection

The process of data collection begins with a target search of ‘EGFR’ molecules within a database known as ‘Chembl’. Since the report focuses on ‘Single Protein’ target types, the molecular id corresponding to humans and the specific type is extracted. The extracted dataset is further mined, wherein the samples with standard type of ‘IC50’ are obtained, to further examine the bioactivity data in length. Subsequently, data pre-processing is done so as to mitigate existing missing values and duplicates and label the compound to its appropriate bioactivity threshold. The resulting curated bioactivity data is used for the subsequent processes.

2.2. Exploratory data analysis

The next step corresponds to performing ‘Exploratory Data Analysis’ on the refined dataset. The dataset passed on from the first process comprises of the molecular id, its matching canonical formula in the form of a ‘SMILE’ and the ‘IC 50’ value which highlights the bioactivity data. The canonical formula is tested with ‘Lipinski’s rule of five’, which is generally the rule of thumb to check if a chemical compound is an active drug in humans. According to this rule,

a) The ‘log p’ value (partition coefficient) must not exceed 5.
b) There must be no more than 5 Hydrogen bond donors.
c) There must be no more than 10 Hydrogen bond acceptors.
d) The molecular mass must not exceed 500 daltons.

In order to align with these descriptors, log of the ‘IC 50’ value is taken and stored as ‘pIC50’. A function is established to test the molecule with the mentioned descriptors using its canonical formula. Finally, EDA is performed on the dataset, with the descriptors along with ‘pIC50’ acting as selecting criterion. It becomes evident that ‘pIC50’ must be the primary measure in dividing active and inactive compounds.

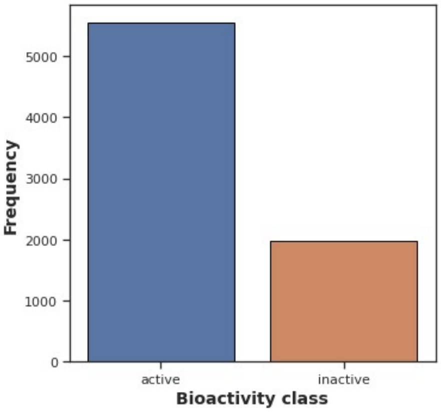

Fig 2a. Frequency comparison of active and inactive compounds

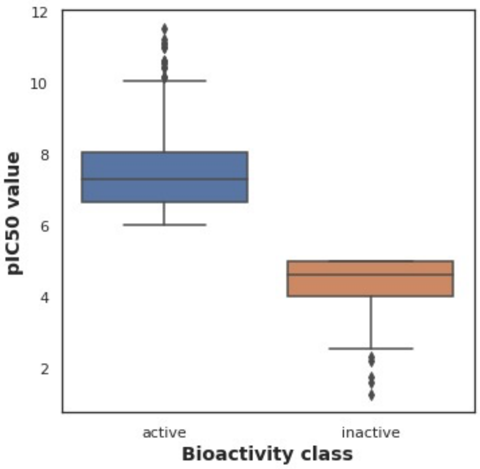

Fig 2b. Bioactivity class vs ‘pIC50’ value

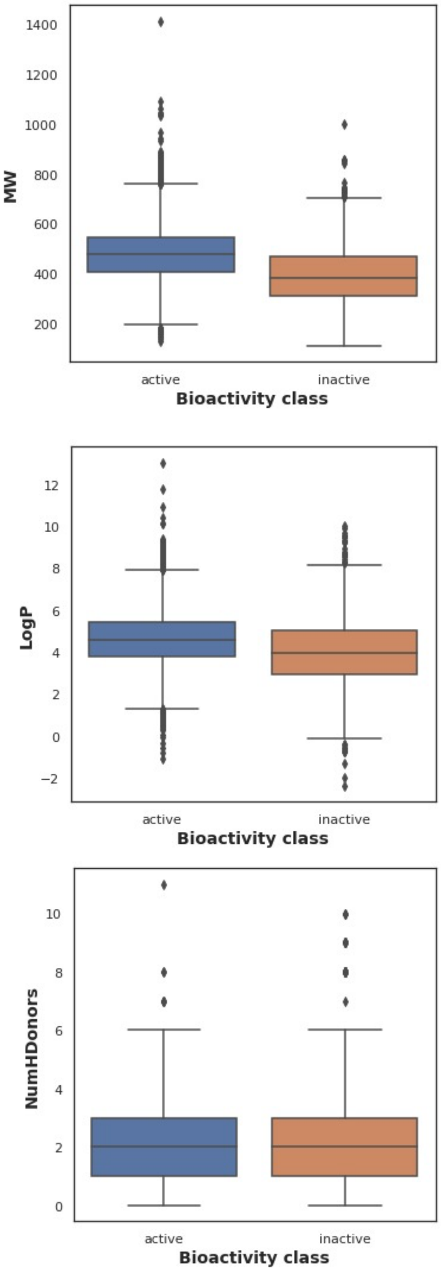

Fig 2c. Bioactivity class vs other descriptors

2.3. Molecular fingerprint calculation

The next phase involves the conversion of the canonical formula into its corresponding molecular fingerprint with the help of the other descriptors. This is done with the aid of ‘PaDEL-Descriptor’, a software that is specifically designed to calculate molecular fingerprints and descriptors. The software removes any unnecessary salt or irrelevant organic compounds from the molecule and standardizes the data, before providing molecular fingerprints that comprise a 0 or 1 value in each column. Each column corresponds to a ‘PubchemFPx’ value, wherein PubchemFP distinguishes different molecules and x ranges from 0 to 880. The files required ‘padel.zip’ and ‘padel.sh’ were downloaded from the original ‘Chembl’ dataset and uploaded to Google Collaboratory for usage. The final dataset to be operated upon contains the molecular id, the molecular fingerprint of 881 columns and the ‘pIC50’ value.

2.4. Feature engineering and model analysis

2.4.1 Random Forest Regressor

Random forest is a type of supervised learning algorithm that uses ensemble methods to solve regression problems. The algorithm operates by constructing a multitude of decision trees at training time and outputting the mean/mode of prediction of the individual trees. Within a random forest, there is no interaction between the individual trees. A random forest acts as an estimator algorithm that aggregates the result of many decision trees and then outputs the most optimal result.

We are using sklearn module to train our Random Forest Regressor Model. We can select number of parameters for our model. We are using parameters such as **n_estimators = 200** which represents the number of trees, **max_depth = 100** which sets the maximum possible depth of each tree.

2.4.2. Ada-Boost Regressor

An Ada-Boost is a meta estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction. AdaBoost decreases the bias of the decision tree and not variance.

We are using sklearn module to train our AdaBoost Regressor Model. We are using **n_estimators = 100** and **random_state = 123**. We are also using **learning_rate = 0.1**. Learning rate is a weight applied to each regressor at each boosting iteration.

2.4.3. Bagging Regressor

It is also known as Bootstrap Aggressor. A Bagging regressor is an ensemble meta-estimator that fits base regressors each on random subsets of the original dataset and then aggregate their individual predictions either by voting or by averaging to form a final prediction. It is generally used to reduce the variance of decision tree by introducing randomization into its construction procedure and then making an ensemble out of it.

We are using sklearn module to train our Bagging Regressor Model. Here also we can select number of parameters for our model. We are using **n_estimators =100** and **base_estimator** as **Decision Tree Regressor** which is default one and **random_state as 123**. Random state controls the random resampling of the original dataset. If the base estimator accepts a random_state attribute, a different seed is generated for each instance in the ensemble.

3. RESULTS

To evaluate our regression models, we are using Scoring Parameter as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

Following are the RMSE and MAE values we obtained for each model.

	Random Forest	Bagging	AdaBoost
RMSE	0.634	1.048	1.305
MAE	0.409	0.736	1.060

We have also plotted the Graphs of **Experimental pIC50** Values against **Predicted pIC50** values for every model.

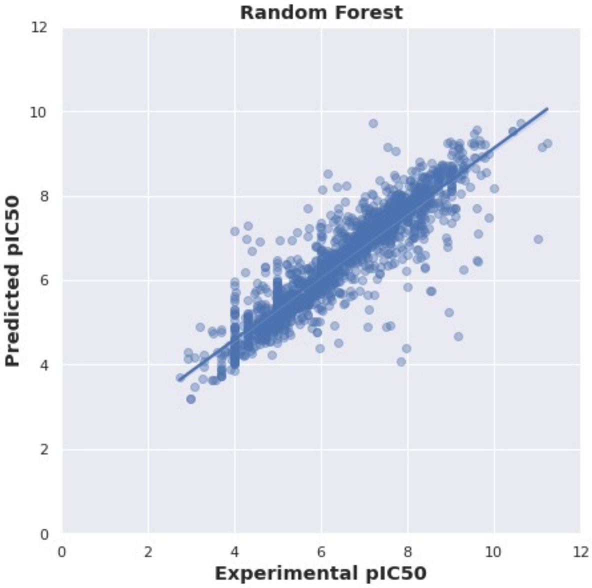

Fig 3a. Random Forest Graph

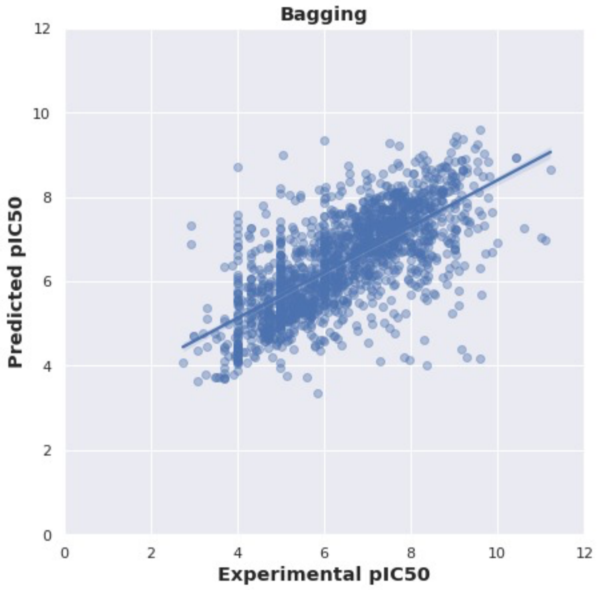

Fig 3b. Graph for Bagging Regressor

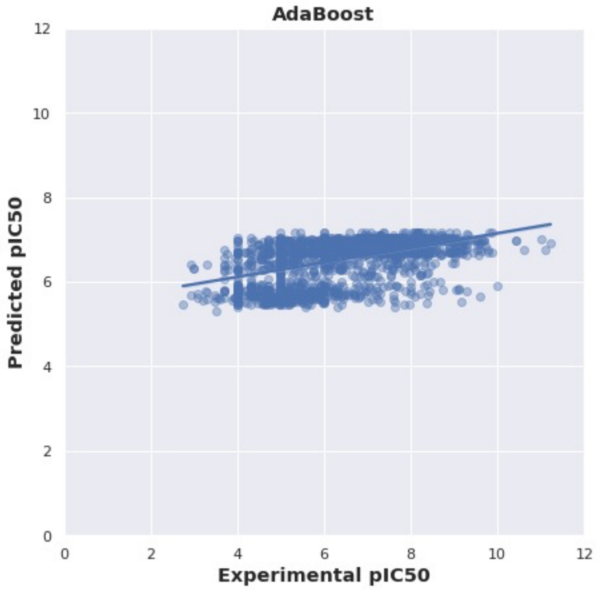

Fig 3c. Graph for Ada-Boost Regressor

4. DISCUSSION

By comparing the RMSE and MAE scores obtained from three models we can clearly see that the Random Forest Regressor is best performing regressor with lowest MAE and RMSE values i.e., 0.634 and 0.409 respectively. Bagging Regressor is second best performing regressor followed by AdaBoost Regressor with highest MAE (1.06) and RMSE (1.30) values.

We can also observe the performance of these different regressor models by analyzing the graphs that we have plotted for experimental vs predicted pIC50 values. The regression line in graph of Random Forest Regressor is almost passing through origin which implies that there is very small difference in Experimental and Predicted Values. Hence Random Forest Regressor is performing best in these three models. If we observe the graph of Bagging regressor we can see that the regression line is bit far from the origin hence there will be some difference in Experimental and Predicted Values. In the graph of AdaBoost regressor the regression line is very far from origin hence there will be significant difference in predicted and experimental values, hence the Error values (RMSE and MSE) for AdaBoost regressor are highest.

Since our Target variable(pIC50) values ranges from 0-10, the Root Mean Squared Error and Mean Absolute Error values i.e.,0.634 and 0.409 respectively are acceptable. Hence, we can conclude that, using Random Forest Regressor Model we predicted the pIC50 values that are very close to the experimental pIC50 values.

12. REFERENCES

[1] R. Qureshi, B. Zou, T. Alam, J. Wu, V. Lee and H. Yan, "Computational Methods for the Analysis and Prediction of EGFR-mutated Lung Cancer Drug Resistance: Recent Advances in Drug Design, Challenges and Future Prospects," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, doi: 10.1109/TCBB.2022.3141697.

[2] A. Ghosh and H. Yan, "Stability Investigation Using Hydrogen Bonds for Different Mutations and Drug Resistance in Non-Small Cell Lung Cancer Patients," 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), 2019, pp. 122-126, doi: 10.1109/BIBE.2019.00030.

[3] L. Wang, S. Wang, H. Yu, Y. Zhu, W. Li and J. Tian, "A Quarter-split Domain-adaptive Network for EGFR Gene Mutation Prediction in Lung Cancer by Standardizing Heterogeneous CT image," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021, pp. 3646-3649, doi: 10.1109/EMBC46164.2021.9630395.

[4] H. Motohashi, T. Teraoka, S. Aoki and H. Ohwada, "Regression Models and Ranking Method for p53 Inhibitor Candidates Using Machine Learning," 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 708-712, doi: 10.1109/BIBM.2018.8621142.

[5] B. Duan, B. Zou, D. D. Wang, H. Yan and L. Han, "Computational Evaluation of EGFR Dynamic Characteristics in Mutation-Induced Drug Resistance Prediction," 2015 IEEE International Conference on Systems, Man, and Cybernetics, 2015, pp. 2299-2304, doi: 10.1109/SMC.2015.402.

[6] M. Malafaia, T. Pereira, F. Silva, J. Morgado, A. Cunha and H. P. Oliveira, "Ensemble Strategies for EGFR Mutation Status Prediction in Lung Cancer," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021, pp. 3285-3288, doi: 10.1109/EMBC46164.2021.9629755.