Development of a Machine Learning computational method for classification of agonist or antagonist.

Adithi Srikumar, Deeksha R Kurthkoti, Dr. Prashantha Karunakar

Department of Biotechnology, PES UniversityBangalore, India

Abstract-

Machine learning is a branch of artificial intelligence which involves the ability of computers to help us solve problems and design good algorithms by making decisions, recognizing patterns in data with little to no human input. In today's world Machine Learning is considered as one of the best ways to interpret results because of its high accuracy and efficiency. The classification of agonist and antagonist of a receptor can be quite tedious, time consuming and labor intensive hence, we have developed a Machine Learning model which predicts the agonism and antagonism.

Keywords- Machine Learning, agonist, antagonist, accuracy

I.INTROCDUCTION

In today's world Machine Learning is considered as one of the best ways to interpret results because of its high accuracy and efficiency. Machine Learning along with Artificial Intelligence helps us to find new ways to solve problems and design good algorithms. Classification of any molecule is a very tedious task and it takes months to obtain the results which might be a loss for the researchers, in order to make the task of selecting an appropriate molecule easier we have developed a Machine Learning model which predicts the agonism and antagonism. With the help of these new techniques, we can classify the agonist and antagonist, in biological terms agonist is a molecule which attaches to the target site and in return activates the target and an antagonist is a molecule which attaches to the target site but fails to activate the target and also prevents different molecules to attach to the site. ML has many algorithms such as Linear Regression, Multiple Regression, K Nearest Algorithms, Random Forests and Decision Tree. Since Linear Regression and Multiple Regression are considered as the most reliable ML algorithms it has been used to create the model which will predict the agonism or the antagonism of the particular drug molecule. Machine Learning models are considered to be reliable when the accuracy lies between 70 – 90%. Our final model has an accuracy of 79.7% and this model can be used to detect any agonist or antagonist of a receptor.

II.IMPLEMENTATION

In order to run the algorithm, we need to create a dataset which will help us detect whether the molecule is an agonist or antagonist. Dataset was formed with the help of multiple Bioinformatics databases such as PaDEL and ChEMBL.The dataset makes it well-suited for use in machine learning and other computational methods. It contains about 1448 properties which are obtained after converting them from SMILES format. In machine learning, a dataset is a collection of data that is used to instruct and evaluate a model. The dataset consists of a set of input data, often called features or predictors, and a set of target values, which the model aims to predict. The quality and diversity of the data in a dataset can have a significant impact on the performance of a machine learning model. A well-curated dataset with a wide range of examples and diverse features can help a model learn more effectively and make more accurate predictions. In contrast, a dataset with limited or poorly-selected data can lead to a poorly-performing model. Datasets are an essential part of the machine learning process, and carefully selecting and pre-processing a dataset is an important step in building an effective model.

Some of the algorithms which were were used to create the ML Model :
Using these tools, we were able to create a concise dataset for developing the model. Now coming to the Machine Learning algorithms used to classify the agonist and antagonist we have implemented LR and MR.In ML linear regression is often used to prognosticate the continuous outcome variable within a set of given input features. The end result of Linear Regression is to find the best fit which describes the relationship between the given coefficients.The coefficients learn from the training set using an optimization algorithm, such as gradient descent. Once the

coefficients have learnt, the trained model can be used to make predictions on the new dataset. Linear regression is one of the simplest types of algorithms in machine learning because of its efficiency and can be used on multiple applications. Multiple regression allows a more complex relationship between the two features which is the input and output.

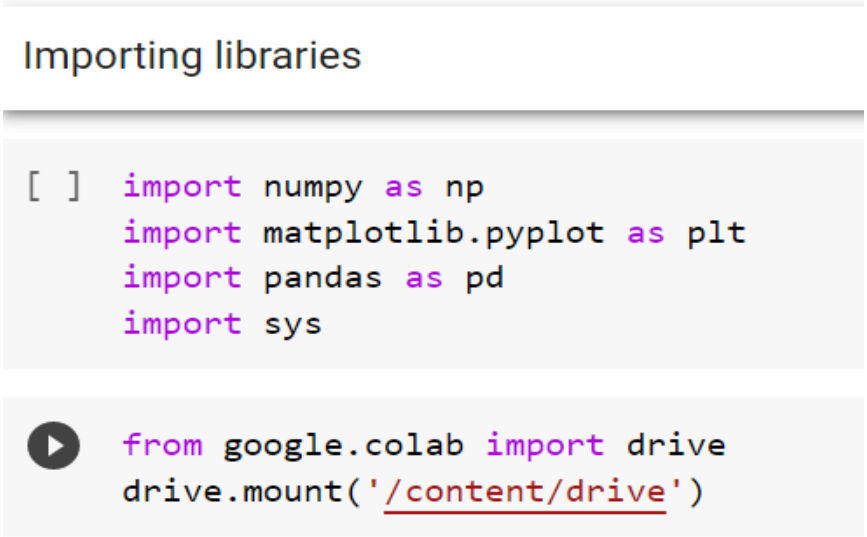
In this model we have split the data as 80:20. The training set is accustomed to train the model, while the test set is accustomed for the evaluation of the model. The trained set is composed of a larger portion of the dataset (80) and is used to learn the model's parameters, such as weights, Ph and the coefficients of the LR and MR model. It is used to minimize the error between the predicted and the true values

Test set on the other hand (20), is used to bench test the trained model on the unseen data. This is important because it allows the models performance to be assessed on data and it is not seen during training, which can help identify overfitting and ensures that the model generalizes well to new details. It is a principle to divide the data into test and training in a systematic way, such as stratified sampling to ensure that the two sets are representative of overall population. It is also required to carefully choose the size of the sets, in this case (80:20) as having too small a test set can lead to unreliable results, and too small of a training set can prevent the model from training to the required conditions.

Mean Squared Error (MSE) is used to assess the accuracy of a machine learning model. It is calculated as the average of the squared differences between predicted values and the true values.In detail, in order to calculate the MSE, the predicted values are taken first and they are subtracted from the true values to obtain error. These errors are then squared and summer, the result is divided by the total number of predictions to obtain MSE.It can be written as:

MSE = (1/N) * ∑ (y_pred - y_true) ^2.Where y_pred is the predicted value, y_true is the true value and N is the total number of predictions.MSE is used because of its differentiability and they are easy to optimize, which makes it apt for using in many ML algorithms.

Working: Machine Learning and Artificial Intelligence can be used to identify the apt agonist and antagonist. These were a few steps which were used to obtain the same:



An excel sheet which contains the properties of the agonist and antagonist are accumulated and it is mounted to the drive for easier access.

Libraries like numpy, matplotlib and pandas are imported. NumPy is used for operating with arrays and matrices. Pandas is a python library used for data analysis and data manipulation. Matplotlib is a library used for plotting graphs and 2D plots.

Importing dataset

```
[ ]  from pandas.core.internals.construction import dataclasses_to_dicts
     dataset = pd.read_csv('/content/drive/MyDrive/cmp list 2.csv')

     dataset =dataset[~dataset.isin([np.nan,np.inf, -np.inf]).any(1)]
     dataset.fillna(0,inplace=True)
     X = dataset.iloc[:,:-1].values
     y = dataset.iloc[:,-1].values
```

The iloc indexer is a method in the pandas library that is used in the selection of rows and columns from a DataFrame by their integer index position.

Selecting columns from the csv file to create a training set and test set. The columns from second to the thousand four forty fifth columns are selected. The first column is not chosen as it contains the name of the compounds in string format. Dataset.axes is used to check the data type of each column and print the length of all the columns in the file.

To split a dataset into specific test and training sets, we can use the train_test_split () function from the scikit-learn library.X_train and Y_train are variables that contain the input data for the training set. They are one of the variables that is returned by the train_test_split () function from the library. It contains the input data that will be used to train a machine learning model. This data is typically a matrix or array with one row per data point and one column per feature. It is often important to scale the features of a dataset before training a model, as features with different scales can have a disproportionate impact on the model's performance. In this model we have given the standard range such as 0 to 1 or -1 to 1. In machine learning, feature scaling is a crucial pre-processing step.

In order to train the model, a relationship between the coefficients(input) and the target variables has to be learnt by the model. When you fit a model to data, you provide the model with a set of input data and the corresponding target values, and the model uses this information to learn and to form a new data. The model uses the training data to determine the coefficients of the line of best fit, which it can then use to make predictions on new data.

Making predictions on the test set allows you to evaluate the performance of the model and see how well it can generalize to new data.

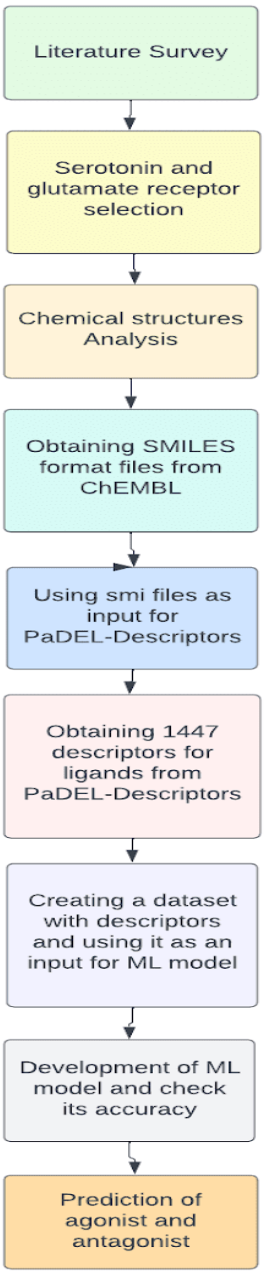
To understand the efficiency of the model, we can visualize the performance of the model on the test set to plot the model's predictions against the true values for the test set. A scatter plot may be used to do this, with the real values shown on the x-axis and the anticipated values on the y-axis. If the model is performing well, the points on the scatter plot should fall close to the line y=x, which indicates perfect prediction.

The accuracy of a machine learning model is a measure of how well it can correctly classify or predict the target values for a given dataset. A model with high accuracy can make correct

predictions on a large proportion of the data, the accuracy of an ML model is an important performance metric, and is often used to contrast the performance of different models on a given dataset. In our model we got an accuracy of 79.7%.

III. Flowchart

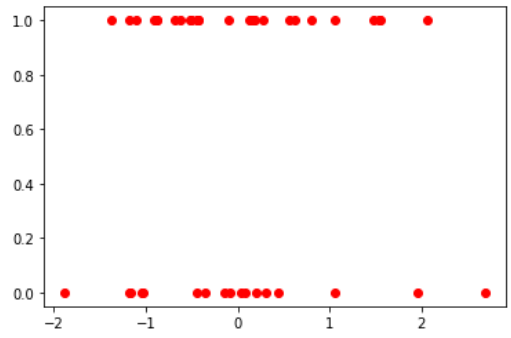
The process of determination of agonist and antagonist of a given receptor involves the study and analysis of various research papers. The receptor chosen is of the GPCR family and initially tested with serotonin which has the highest number of agonists and antagonists 43 and 31 respectively. Glutamate receptor was selected later with 6 agonists and 16 antagonists in an attempt to increase the accuracy.

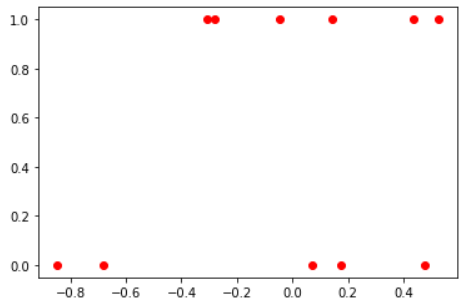


The chemical structures were analyzed using ChEMBL to determine the essential factors on the basis of which the agonists and antagonists bind to the respective receptor. The SMILES format of the chemical structures was obtained. SMILES stands for "Simplified Molecular Input Line Entry System" which is used to convert the three-dimensional structure of a chemical into a series of symbols that computer programs can easily comprehend.
An open source software known as PaDEL-Descriptors was used to calculate the molecular descriptors. The SMILES format was used as input and a folder was created to store the output from PaDEL. The output folder obtained after a runtime of 0.43 seconds contained 1447 descriptors such as hybridization ratio, Hbond acceptor count, largest chain and so on.
The 1447 descriptors were used as input to the machine learning model. The accuracy was determined and was found to be 79%.

The agonist and antagonist of the respective receptor was predicted and represented in the form of 0 and 1.

IV. Analysis and Results






```
[ ]  y_pred = regressor.predict(X_test)
     for i in range(len(y_pred)):
       if y_pred[i]>1:
         y_pred[i]=1
       else:
         y_pred[i]=0
     #np.set_printoptions(precision=2)
     #print(np.concatenate((y_pred.reshape(len(y_pred),1), y_test.reshape(len(y_test),1)),1))
     from sklearn.metrics import mean_squared_error
     mean_squared_error(y_pred,y_test,squared=False)
     #print(y_pred)           #y_pred and y_test
```

```
0.7977240352174656
```

```
[ ]  pred=regressor.predict(X)
     res = []
     for i in range(len(pred)):
       if pred[i]>1:
         pred[i]=1
         res.append('antagonist')
         #print('It is an Antagonist')
       else:
         pred[i]=0
         res.append('agonist')
         #print('It is an Agonist')
     result=[]
     for i,j in zip(pred,res):
       temp = (i,j)
       result.append(temp)

     print(result[50])
```

```
(0.0, 'agonist')
```

The accuracy of a machine learning model is a measure of how well it can correctly classify or predict the target values for a given dataset. In classification tasks, accuracy is typically calculated as the number of true predictions made by the model, divided by the total number of predictions made. In regression tasks, accuracy is often measured using metrics such as mean absolute error or root mean squared error. A model with high accuracy can make correct predictions on a large proportion of the data, while a model with low accuracy may struggle to make accurate predictions. The accuracy of an ML model is an important performance metric, and is often used to contrast the performance of different models on a given dataset. In our model we got an accuracy of 79.7%.

V.Abbreviations:

1. ML – Machine Learning
2. AI – Artificial Intelligence
3. ChEMBL – European Molecular Biology Laboratory
4. LR- Linear Regression
5. MR- Multiple Regression
6. MSE – Mean Squared Error
7. CSV – Comma Separated Values
8. AUC – Area Under the Curve
9. GPCR- G -Protein Coupled Receptor
10. ER – Estrogen Receptor
11. ROC – Receiver Operating Characteristic Curve
12. PaDEL descriptors

VI.ACKNOWLEDGMENT:

We would like to express our deepest appreciation and gratitude to all those who have contributed to the development of this Machine Learning computational method for classification of agonist or antagonist.

Firstly, we extend our sincere thanks to our supervisor for providing us with guidance, support, and valuable insights throughout the research process. Your continuous encouragement and constructive feedback were invaluable in shaping the direction and focus of this work.

We would also like to acknowledge the contributions of our colleagues and collaborators, who have provided us with data, technical expertise, and valuable discussions. Their willingness to share their knowledge and experience has been crucial in the successful completion of this project.

Additionally, we are grateful to the reviewers and editors whose constructive feedback has helped to improve the quality and clarity of this work.

Finally, we would like to thank our families and friends for their unwavering support and encouragement throughout this endeavour. Your belief in us has been a constant source of inspiration, and we could not have accomplished this without your love and support.

Once again, I extend my heartfelt gratitude to everyone who has played a part in the development of this Machine Learning computational method for classification of agonist or antagonist.

VII.References”

1.Kurosaki, K., Wu, R., & Uesawa, Y. (2020). A toxicity prediction tool for potential agonist/antagonist activities in molecular initiating events based on chemical structures. *International journal of molecular sciences*, *21*(21), 7853.

2.Zorn, K. M., Foil, D. H., Lane, T. R., Hillwalker, W., Feifarek, D. J., Jones, F., ... & Ekins, S. (2020). Comparison of Machine Learning Models for the Androgen Receptor. *Environmental science & technology*, *54*(21), 13690-13700.

3.Matsuzaka, Y., & Uesawa, Y. (2020). DeepSnap-deep learning approach predicts progesterone receptor antagonist activity with high performance. *Frontiers in bioengineering and biotechnology*, *7*, 485.

4.Jabeen, A., de March, C. A., Matsunami, H., & Ranganathan, S. (2021). Machine learning assisted approach for finding novel high activity agonists of human ectopic olfactory receptors. *International journal of molecular sciences*, *22*(21), 11546.

5.Sakamuru, S., Zhao, J., Xia, M., Hong, H., Simeonov, A., Vaisman, I., & Huang, R. (2021). Predictive models to identify small molecule activators and inhibitors of opioid receptors. *Journal of Chemical Information and Modeling*, *61*(6), 2675-2685.

6.Piir, G., Sild, S., & Maran, U. (2021). Binary and multi-class classification for androgen receptor agonists, antagonists and binders. *Chemosphere*, *262*, 128313.

7.Matsuzaka, Y., Totoki, S., Handa, K., Shiota, T., Kurosaki, K., & Uesawa, Y. (2021). Prediction Models for Agonists and Antagonists of Molecular Initiation Events for Toxicity Pathways Using an Improved Deep-Learning-Based Quantitative Structure–Activity Relationship System. *International journal of molecular sciences*, *22*(19), 10821.

8.Sugaya, N. (2013). Training based on ligand efficiency improves prediction of bioactivities of ligands and drug target proteins in a machine learning approach. *Journal of chemical information and modeling*, *53*(10), 2525-2537.

9.Ciallella, H. L., Russo, D. P., Aleksunes, L. M., Grimm, F. A., & Zhu, H. (2021). Predictive modeling of estrogen receptor agonism, antagonism, and binding activities using machine-and deep-learning approaches. *Laboratory investigation*, *101*(4), 490-502.

10.Matsuzaka, Y., & Uesawa, Y. (2020). molecular image-based prediction models of nuclear receptor agonists and antagonists using the deepsnap-deep learning approach with the Tox21 10K library. Molecules, 25(12), 2764.

11.Zhu, X. L., Cai, H. Y., Xu, Z. J., Wang, Y., Wang, H. Y., Zhang, A., & Zhu, W. L. (2011). Classification of 5-HT1A receptor agonists and antagonists using GA-SVM method. *Acta Pharmacologica Sinica*, *32*(11), 1424-1430.

12.Wang, F., & Xing, J. (2019). Classification of thyroid hormone receptor agonists and antagonists using statistical learning approaches. *Molecular Diversity*, *23*(1), 85-92.

13.Caballero-Vidal, G., Bouysset, C., Grunig, H., Fiorucci, S., Montagné, N., Golebiowski, J., & Jacquin-Joly, E. (2020). Machine learning decodes chemical features to identify novel agonists of a moth odorant receptor. *Scientific reports*, *10*(1), 1-9.

14.Singh, P. (2021). *Deploy Machine Learning Models to Production*. Apress.