Activity Prediction of Pre-Clinical Trial Drugs Of Estrogen Receptors Using Machine Learning

Hari Shankar Tiwari[1], Kumari Swati Gupta[2], Mitali Srivastava[3], Nitin Sharma[4], Vishan gupta[5]

[1,2,3,4] *(Bachelor of Technology, Final Year Students, 5 Assistant Professor,*

Department of Computer Science & Engineering, IMS Engineering College, Ghaziabad, U.P., India

***Abstract**: In previous days, the drugs to be produced were tested on the animals in in-vivo technique for the prediction of drug activity. This technique is time-consuming process and has a large threat to life of animals. To replace the technique , a computational model implementing the machine learning on the QSAR(Quantitative Structure-Activity Relationship) which deals with the physicochemical properties of molecules. Based on the previously recorded training dataset of values of different physiochemical properties. The machine learning models are applied on this data and predict the testing dataset and classify as the Active or Inactive. Based on the time consumed and accuracy of the different models conclusion is defines as which machine learning model is best suited for estrogen receptor dataset which has been given.*

***Keywords**: Active molecules, Inactive molecules, Estrogen Receptor, accuracy*

I- INTRODUCTION

According to the report compiled by National Centre for Biotechnology Information (NCBI), there has been more number of animals going to death when the drugs are tested on the animals in the in-vivo testing technique. When any drug is verified by the government to sell in the market before it some process is happened and this process goes in two steps. The first step is test these drugs on animals which is known as in-vivo technique after that the in-vitro technique is performed. So in the in-vivo technique these drugs tested on animals due to that many animals dies. To stop or to reduce the animal death we worked on the in-silico technique which is the computational method to know that the drug molecules are active or not using many machine learning algorithms to reduce the animal death.

The main purpose of the activity prediction using computational methods is to reduce the pressure of animal testing, cost and time reduction in early stages of drug discovery. The concept of Russell and Burch(1959) about the growing popularity of the 3Rs (replacement, reduction, refinement) focuses on the limited use of the animal for activity testing(in-vivo). In-silico methods also can predict ADMET (absorption, distribution, metabolism, excretion and toxicity) related properties in chemical space that reduce the dependency of chemical laboratory synthesis (in-vitro).

We develop the computational method (in-silico) for checking the activity of drug molecules rather then inside the living animals(in-vivo) or in glass(in-vitro) to save the life of animals, and money. In basics, the Active molecules are those Molecules that have the ability to bind with the ERs and various estrogenic effected by modulating the effect of ER while on the other hand inactive cannot bind with ER.

For the development of computational model, The machine learning model is employed on the physicochemical properties of the molecule.

For the study of physicochemical properties understandings, Quantitative Structure–Activity Relationship (QSAR) analysis is widely used. QSAR modelling provides an effective way for establishing and exploiting the relationship between chemical structures and their biological actions toward the development of novel drug candidates. Theoretically, QSAR analysis is the application of mathematical and statistical methods for the development of models for the prediction of biological activities or properties of compounds. Formally, a QSAR model can be expressed in the following generic format:

Predicted Biological Activity=Function (Chemical Str ucture)

A QSAR procedure tries to minimize the error of prediction, for example, in the form of the sum of squares between predicted and observed activities. The process of QSAR model development can be divided into three parts: data preparation, data analysis, and model validation. Model validation should include establishment of model applicability domain.

It has been reported that most QSAR models do not work well for evaluating in-vivo toxicity, especially for external compounds (Zvinavashe et al. 2008, 2009). Several reviews were published recently that challenge the feasibility and reliability of QSAR models of chemical toxicity (Johnson 2008; Stouch et al. 2003).

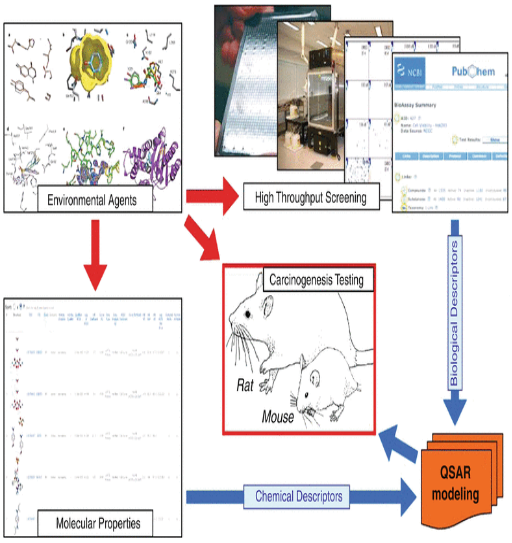


FIG. QSAR MODELING

Recently, the European Organization for Economic Co-operation and Development (OECD) developed a set of principles for the development and validation of QSAR models, which, in particular, requires “appropriate measures of goodness-of-fit, robustness, and predictivity” (Organisation 2008). The OECD guidance document especially emphasizes that QSAR models should be rigorously validated using external sets of compounds that were not used in the model development. So, adapting the machine learning model by QSAR helps to find its place in drug design and to move to the plateau of productivity.

For the study of prediction of activity of pre-clinical trial drugs, we are using Estrogen Receptor (ER) molecules with their properties being applied on the classification model in machine learning to improve the results of QSAR. **Estrogen Receptor** are a protein found inside the cells of the female reproductive tissue as represented in Fig 3.1, some other types of tissue, and some cancer cells[1]. The hormone estrogen will bind to the receptors inside the cells and may cause the cells to grow. Also called ER.

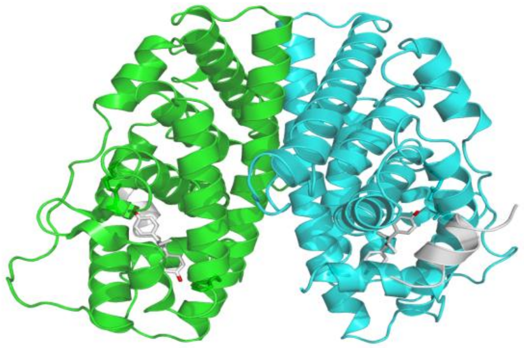


Fig 3.1 Physical structure of Estrogen Receptor

The classification machine learning models are applied on the ER dataset containing the ER molecules with their status of activity or inactive. Classification is also called categorization, is a supervised machine learning technique that uses known data to determine how the new data should be classified into a set of existing classes. There are many classification models like Decision Tree, Random Forest, Decision Tree, SVM, Neural Network, but we are only using Decision Tree, Random Forest, SVM (Support Vector Machine) for the prediction of the results. The results are being compared based on the classification model evaluation parameter.

2. METHODOLOGY

For the Drug Activity Prediction, we use Machine Learning technology, For which we developed a prediction pipeline that enables the use of machine learning for the activity prediction. This prediction pipeline is developed for the Estrogen Receptor datasets and use machine learning algorithms to the datasets to build a conclusion on the predicted results. We first introduce the dataset [Section 2.1] and the dataset Pre-processing [Section 2.2].Afterwards we present the machine learning models [Section 2.3] to be used in the development. We have used R studio as platform for development having packages as rpart , random Forest, Boruta and svm.

2.1 Dataset Description

In this Prediction process, We have used Estrogen Receptor dataset. The dataset consists of 748 drug molecules with 1444 features. In the dataset, out of 748 ER’s drug molecule, there are 374 molecules are active molecules and 373 are inactive molecules. Each molecule has 1444 features which includes their physiochemical properties. These physiochemical properties are also known as Molecular Descriptor. So, Molecular Descriptor defines a drug molecule based on their physiochemical properties. The activity of dataset is divided into binary classes. One class is Active class and other class in Inactive. Activity of drug molecules is classified as '0' and '1'. Out of these '0' is used for inactive molecules and '1' for inactive drug molecules.

Some of the common features are ALogP, AMR(molecular refractivity), Apolna Arom Atom, nAromBond, nAtom, nBase, nRing, surface area, Volume etc.

Out of these drug molecules, 70% data has been set for training dataset and remaining 30% as a testing dataset.

Training dataset= (total dataset*training/100);

Testing dataset= (total dataset*testing/100);

Therefore , out of 748 drug molecules , 524 is used as a training data and remaining 224 drug molecules is used as testing dataset.

2.2 Dataset Pre-processing

Each dataset in the raw format consists some of the null entries and redundant data. When we apply algorithm to these kind of dataset, the algorithm could misbehave or incorrect results is produced and the feasibility of the algorithm is reduced. In our dataset also there were so many null entries which are replaced by the mean value of that feature throughout the dataset. If X is a set of 'n' data items then mean is calculated as $\bar{x}$.

$$\bar{x} = (\Sigma x_i) / n$$

This makes the data consistent and does not allow the result to deviate from the actual behavior.

2.3 Feature Selection

Feature Selection[2] is used to filter out correlated variables, molecular descriptors having many zero values, missing values, and noise from the dataset. As per discussion our dataset contains 1444 features, which are large in number and requires high time complexity while executing a model.

Feature selection is a process which extracts essential features improves the performance of the model and removing irrelevant and redundant features. For the Feature Selection process, we have used Boruta[3] package in R. Boruta package has many functions for selecting the important features from the dataset. Precisely, it works as a wrapper algorithm around Random Forest. This package derive its name from a demon in Slavic mythology who dwelled in pine forests. We know that feature selection is a crucial step in predictive modelling. This technique achieves supreme importance when a data set comprised of several variables is given for model building. Firstly, it adds randomness to the given data set by creating shuffled copies of all features (which are called shadow features). Then, it trains a random forest classifier on the extended data set and applies a feature importance measure (the default is Mean Decrease Accuracy) to evaluate the importance of each feature where higher means more important. At every iteration, it checks whether a real feature has a higher importance than the best of its shadow features (i.e. whether the feature has a higher Z score than the maximum Z score of its shadow features) and constantly removes features which are deemed highly unimportant. Finally, the algorithm stops either when all features gets confirmed or rejected or it reaches a specified limit of random forest runs. The Boruta technique extracted 48 features that are going to be applied on the model. These features have least number of the missing values or noisy data and these are used for the activity prediction using classification model.

Table 1-Some Descriptors extracted after feature selection

Feature Name	Feature Description
AMR (Atomic Molar Refractivity)	The total polarizability of a mole of a substance and is dependent on the temperature, the index of refraction, and the pressure.
Nn	Number of Nitrogen atoms
AATS (Autocorrelation of topological Structure)	Index that measures the Linear correlation between lagged of time series y.
ETA (Extended Topological Atom)	Index for modelling chemical and drug-induced toxicities and physicochemical properties relevant to such toxicities.
VP (Vapour Pressure)	The **pressure** exerted by a **vapor** that is in equilibrium with its solid or liquid form.
VPC4,5,6 (Valence	This descriptor signifies

Path)	valence molecular connectivity index of order 4th, 5th, 6th path-clusters.
NsssCH	This descriptor defines total number of –CH bond connected with three single bond

2.2 Target class for classification

The target class is Activity which contains two instances Active(1) and Inactive(0). Active molecules have the ability to bind with the ERs and various estrogenic effected by modulating the effect of ER while on the other hand inactive can not bind with ER.

2.2 Machine Learning models

Model Name	Required package in R
Random forest(RF)	Random Forest
Decision Tree	Rpart
Support Vector machine (SVM)	Kern lab

Models which are used for classification of activity are explained below with their source codes. Table2 displays the various models with their packages used in multilevel prediction model. All models are implemented in R. We are using random forest, decision tree and SVM for the prediction.

2.5.1 Decision Tree

Decision tree induction is the learning of decision trees from class-labelled training tuples. A **decision tree** is a flowchart-like tree structure, where each **internal node** (non-leaf node) denotes a test on an attribute, each **branch** represents outcome of the test, and each **leaf node** (or *terminal node*) holds a class label. The topmost node in a tree is the **root** node. [3]

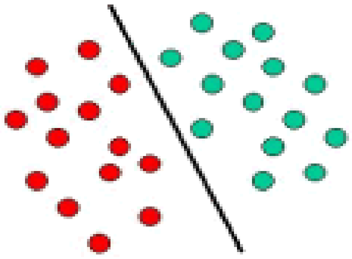
rpart package is used for the implementation of decision tree. Rpart library uses rpart(…) function to apply the model.

2.5.2 Random Forest

The model builds a number of decision tree models and predicts using the ensemble. An individual decision tree is built by choosing a random sample from the training data set as the input. At each node of the tree, only a random sample of predictors is chosen for computing the split point. This introduces variation in the data used by the different trees in the forest[3]. Library used for the model is random Forest in the R. This library has method random Forest(..) to implement the model on the dataset.

2.5.3 Support Vector Machine (SVM)

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A schematic example is shown in the illustration below. In this example, the objects belong either to class GREEN or RED. The separating line defines a boundary on the right side of which all objects are GREEN and to the left of which all objects are RED. Any new object (white circle) falling to the right is labelled, i.e., classified, as GREEN (or classified as RED should it fall to the left of the separating line).[4]



Package used in R for the Support Vector Machine model is kernlab. This package include kernlab library and function used is ksvm.

2.6 CLASSIFICATION MODEL EVALUATION PARAMETERS

2.6.1 Gini Coefficient: The Gini coefficient measures the inequality among values of a frequency distribution. A Gini coefficient of zero expresses perfect equality where all values are the same. Gini coefficient ranges from 0 to 1. The 0 values denotes the perfect equality of data.[4]

2.6.2 Sensitivity: *Sensitivity or Recall is also known as True positive rate. It is the ratio of actual positives which are correctly identified as positive[5]. It is calculates as,*

Recall = TP/(TP+FN)

2.6.3 Specificity: *It is also known as true negative rate. It is the ratio of actual negatives which are correctly identified as negative[5]. It is computed as,*

Specificity=TN/(TP+FN)

2.6.4 Precision: *Precision can be thought of as a measure of exactness; it means what percentage of tuples labelled as positive are actually positive[6].*

Precision= TP/(TP+FP)

2.6.5 Accuracy: Accuracy is most important criteria to measure exactness of any classifier[6]. The accuracy can be calculated as

Accuracy= ((TP+FP)/(TP+FP+FN+TN))*100

3. RESULT ANALYSIS AND CONCLUSION

3.1 RESULT

Model	Ginn index	Sensitivity	Specificity	precision
Decision tree	0.699	0.79	0.783	0.803
Random forest	0.859	0.86	0.891	0.94
SVM	0.796	0.705	0.939	0.919

The Accuracy and Time consumed are given in table below.

Model	Accuracy	Total time(in sec)
Decision tree	78.67	7.02
Random forest	88.55	8.86
SVM	82.4	7.36

This section presents the analysis of predicted results of all three machine learning classification models that are decision tree, random forest, support vector machine which have already been discussed.

The above table represents the values of performance parameters of different models. The accuracy of decision tree, random forest and svm are **78.67, 88.55** and **82.4** respectively. Accuracy of random forest is highest and the accuracy of decision tree is minimum. But the time complexity of Decision tree is least.

3.2 ANALYSIS

From the above discussion, Random forest model predict the 88.55 out of 100 drug molecules as active when it was actually active. While the decision tree predicted the correct result 78.67 out of 100 and SVM predicted 82.4 out of 100 drug molecules as active when it was actually active.

3.3 CONCLUSION

In this project, we have proposed a machine learning based multilevel prediction model to the assessment of quality of ER drug molecules based on QSAR/QSPR. The target class for the prediction was Activity. The dataset used in the prediction process was improper an has large number of features. Then we applied Feature reduction model i.e. Boruta in R. After the pre-processing we developed a multilevel prediction model which include three different algorithms in classification. These models were Decision tree, Random Forest and Support Vector Machine. Based on the confusion matrix, it was possible to calculate the values of evaluation parameters for the different models.

These parameters were Gini coefficient, sensitivity, specificity, precision, accuracy and time taken. We believe that by changing the subsets of features by using some different feature selection tool the performance of the models can be enhanced and get optimized parameter values. The limitation of this project is that we can use those kind of drug molecules for which our model has been trained for predicting its activity. Since different kind of drug have different physiochemical properties. Therefore, different kind of drug molecules cannot be identified by this model.

REFERENCES

[1] Estrogen Receptors and human disease by Bonnie J. Deroo and Kenneth S. sKorach https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2373424/

[2] http://r-statistics.co/Variable-Selection-and-Importance-With-R.html

[3] Data Mining: Concepts and Techniques, 3rd Edition by Micheline Kamber, Jian Pei, Jiawei Han

[4] Hare singh nayak, https://www.crazyengineers.com/threads/what-is-gini-index-why-it-is-used-in-data-mining.59082

[5] https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38

[6] Vishan Kumar gupta and Prashant Singh Rana , Activity assessment of small drug molecules in estrogen receptor, www.ietdl.org

[7] https://datascience.stackexchange.com/questions/5345/ide-alternatives-for-r-programming-rstudio-intellij-idea-eclipse-visual-stud

[8] https://www.tutorialspoint.com/r/r_overview.htm

[9] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[10] https://www.rdocumentation.org/packages/Boruta/versions/6.0.0/topics/Boruta

[11] https://www.researchgate.net/publication/232715019_Implementation_of_Multiple-Instance_Learning_in_Drug_Activity_Prediction
https://github.com/gaurangmhatre/Drug-Activity-Prediction-with-Classification