Prediction of Ocular Drug Toxicity using Supervised Machine Learning Models

Manisha Jhunjhunwala
Masters of Engineering, Department of CSIS
Birla Institute of Technology, Pilani (Hyderabad Campus)
Student ID - 2020H1030125H

Surbhi Sharma
Masters of Engineering, Department of CSIS
Birla Institute of Technology, Pilani (Hyderabad Campus)
Student ID - 2020H1030148H

Under the supervision of

Dr. Chittaranjan Hota, Department of Computer Science

Dr. Nirmal Jayabalan, Department of Pharmacy

Birla Institute of Technology & Science, Hyderabad

Abstract

There are many systemic drugs which can cause ocular toxicity due to their accumulation at non-target site. Hydroxychloroquine is one of the example of a drug causing induced ocular toxicity. It is caused when the drugs gain access to the eyes through the transporters. So, structural fingerprints within the drug molecules will have to be recognized as substrate/non-substrate against a particular transporter. To protect people from potential eye injury caused due to these toxic drugs, several in-lab experiments have to be carried out including animal screening, which is both time-consuming and costly. To assist the pharmacological testing, several computational models can be made to tag potential chemical compounds as substrates or non-substrates and then testing only the high risk molecules (probably the substrates) for further research. This virtual screening using machine learning models is thus quite efficient and attractive to the toxicologists for tagging potential ocular toxicants. Various Supervised Machine Learning models have been used to identify whether the molecule is a substrate or non-substrate, so that the toxicity of the drugs can be predicted.

I INTRODUCTION

Humans are exposed to an abundance of chemical compounds via the environment, nutrition, cosmetics, and drugs. To protect humans from potentially harmful effects, these chemicals must pass reliable tests for adverse effects and, in particular, for toxicity.

Our main objective is to predict the drug induced ocular toxicity based on their structural and physiochemical properties using Artificial Intelligence tools which defines their interaction with the transporters. The aim would be to use several supervised machine learning algorithms to develop a model for the prediction of substrate specificity towards particular transporter. Many studies have confirmed the crucial role of drug transporters in accumulation of drugs at non-target site. Drugs may gain access to eye through transporters and accumulation overtime may lead to toxicity.

In silico study involves analysis of physicochemical parameter for substrate and non-substrate molecules in the dataset (systemic drugs inducing ocular toxicity). Classification Quantitative-Structural Activity Relationship (QSAR) study by Monte-Carlo optimization and Bayesian modelling methods, also, including machine learning models can be used for the development of models for substrate specificity towards particular transporter. In our work, we have used classification algorithms such as k-nearest neighbour (kNN), Random Forest, Naive Bayes, Decision Tress, C4.5, XG-BOOST and SVM to train our model across k folds. Logistic Regression was used as a stacking algorithm to get the ensembled prediction of the models trained for the molecules in testing and validation set. The final predictions along with the molecular properties are exported in a CSV file for further analysis (if required).

II BACKGROUND

First, Membrane transporter proteins or drug transporters are integral transmembrane proteins. They are known as gatekeepers of cells and organelles. This have roles in cell survival by controlling the influx and efflux of a variety of substances throughout these cells. These transmembrane proteins play vital roles in the maintenance of pharmacokinetics of the drugs along with its bio-distribution within the body. It is important to examine the role of these transporters in drug interaction in the presence of additional drugs or nutrients in order to avoid the danger of unwanted drug interactions.

In order to optimize drug concentrations at the site of action to reduce the toxicity as well as to predict pharmacokinetics in drug discovery and development program, it is necessary to figure out the transporter role in clearance mechanisms of these drugs.

We have used various Supervised Machine Learning Models to help predict, whether the molecule is a substrate or a non-substrate on the basis of various molecular properties like molecular volume, molecular weight, topological polar surface area, molecular ClogP value, molecular ROTB value, number of oxygen and nitrogen present in molecule (nON) and number of hydroxy and amine group present in molecule (nOHNH).

II.A k-Nearest Neighbour

It is one of the simplest and oldest machine learning algorithm known for pattern classification and is also known as instance-based learning. Usually, the accuracy is very high in addition to the time complexity which is the only drawback of the method. The k-nearest neighbors (KNN) is a simple, easy-to-implement supervised machine learning algorithm that can be used for both classification and regression. The algorithm assumes that similar things exist in close proximity, in other words, similar things are near to each other. The approximate distance between various points on the input vectors is computed and then assigning of the unlabelled points to the class of its k-nearest neighbours. k is an important parameter and different values for k cause different performances. We have trained the model with k=8.

II.B Random Forest

Random forest is a most used, flexible, easy to use, simple and diverse algorithm that can be used for both regression and classification problems and produces, even without hyper-parameter tuning, a great result most of the time. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging is that a combination of learning models increases the overall result. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Our model works on 100 estimators (also called decision trees) in forest and with random state as 0 which controls both, the randomness of the bootstrapping of the samples used when building trees and the sampling of the features to consider when looking for the best split at each node.

II.C XG-Boost

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. It is an efficient, effective and a very fast machine learning method that is used for handling very large datasets. It is a perfect combination of software and hardware optimization techniques to yield superior results using less computing resources in the shortest amount of time. We have set eval_metric as 'mlogloss' and use_label_encoder as 'False' to train our model on the dataset.

II.D Naive Bayes Classifier

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. Naïve Bayes is a probabilistic machine learning algorithm that is used in a wide variety of classification tasks.

II.E Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes. Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line). For our model, we have considered the kernel as 'linear' as the data is plotted on 2-D plane and can be seperated by a line.

II.F Decision Tree (C4.5 Single Decision Tree)

Decision Tree algorithm belongs to the family of supervised learning algorithms. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. The decisions or the test are performed on the basis of features of the given dataset. The decisions or the test are performed on the basis of features of the given dataset. To train our model on the given dataset, we have used the criterion as 'entropy', random_state as 100,max_depth as 3, min_samples_leaf as 5.

III APPROACH

Our work was divided into broad steps comprising of curating of data, data preprocessing, designing a consensus supervised machine learning model and finally predicting the various factors which leads to drug induced ocular toxicity and how they induce it.

First of all, the dataset of molecules or drugs causing ocular toxicity due to systemic administration was prepared by gaining information about the domain from various literatures. The dataset was then checked for structural diversity by different methods. Different features of the molecules were identified which plays an important role in distinguishing different compounds. Features such as, ClogP, Number of hydroxy (-OH) and amine (-NH) group present (nOHNH), Topological polar surface area (TPSA), Molecular weight, Molecular volume , Number of oxygen (-O) and nitrogen (-N) present (nON) and ROTB are used to train the model. Among these features, ClogP and nOHNH are of higher priority, then comes molecular weight, TPSA and others. Threshold values have also been researched upon from the available literature and is taken into consideration while constructing a decision tree.

Next comes the data pre-processing step. It is a very integral part before starting with any training. We have found that there were several compounds where some data in multiple features was missing. Also, the range of values was quite wide, which poses a difficulty in proper training. We have thus handled the missing values and have normalised the dataset before the training process. The NAN values in dataset were replaced by the mean of that particular feature.

Also, the dataset that we created was not evenly distributed and to train model for all possible combinations, dataset has to be shuffled. Initially, we applied Random shuffling on the dataset. But we noticed, on every run, because of random shuffling of data, different molecules were picked at random for training the testing the models, which led to different predictions for the molecules on each execution. To solve this issue, we added k-fold cross validation in our model. K-fold cross validation is a technique that divides the dataset into k-subsets (or folds) of equal size, test the data on 1 fold and use (k-1) folds to train the model. The process is repeated k-times, each time for a different fold for validation. This gave us more accurate estimate of a model's predictive performance. In our implementation, we have used k=5, so every data points get to be tested exactly once and is used in training 4 times.

Now, from a large pool of algorithm's, we need to choose which algorithms and techniques can help us get good results. In the present work, we have focussed on mainly supervised algorithms to build the model, which will be used based on the combination of different descriptors.

The predictive models was obtained by running the parameters over various ML algorithms such as k-Nearest neighbors, Random Forest, a special class of decision tree C4.5, XG Boost, Support Vector machines and Naive Bayes probabilistic techniques. We have implemented a total of 6 supervised learning models initially. Then the performance of these models was compared with the existing methods of predicting drug toxicity. We have found that some models gave quite good results (around 90 – 92%) whereas some have a lower accuracy (around 75 %). We also noticed that for the similar performing models like RF and KNN, there were a set of molecules having different predictions.

To overcome this problem, we have used the concept of consensus modelling. It comprises of several aggregation techniques to get better results and make use of the strengths and weaknesses of the individual algorithms. We have used Logistic regression model as the consensus model to combine the predictions of the 6 base models and give a final prediction.

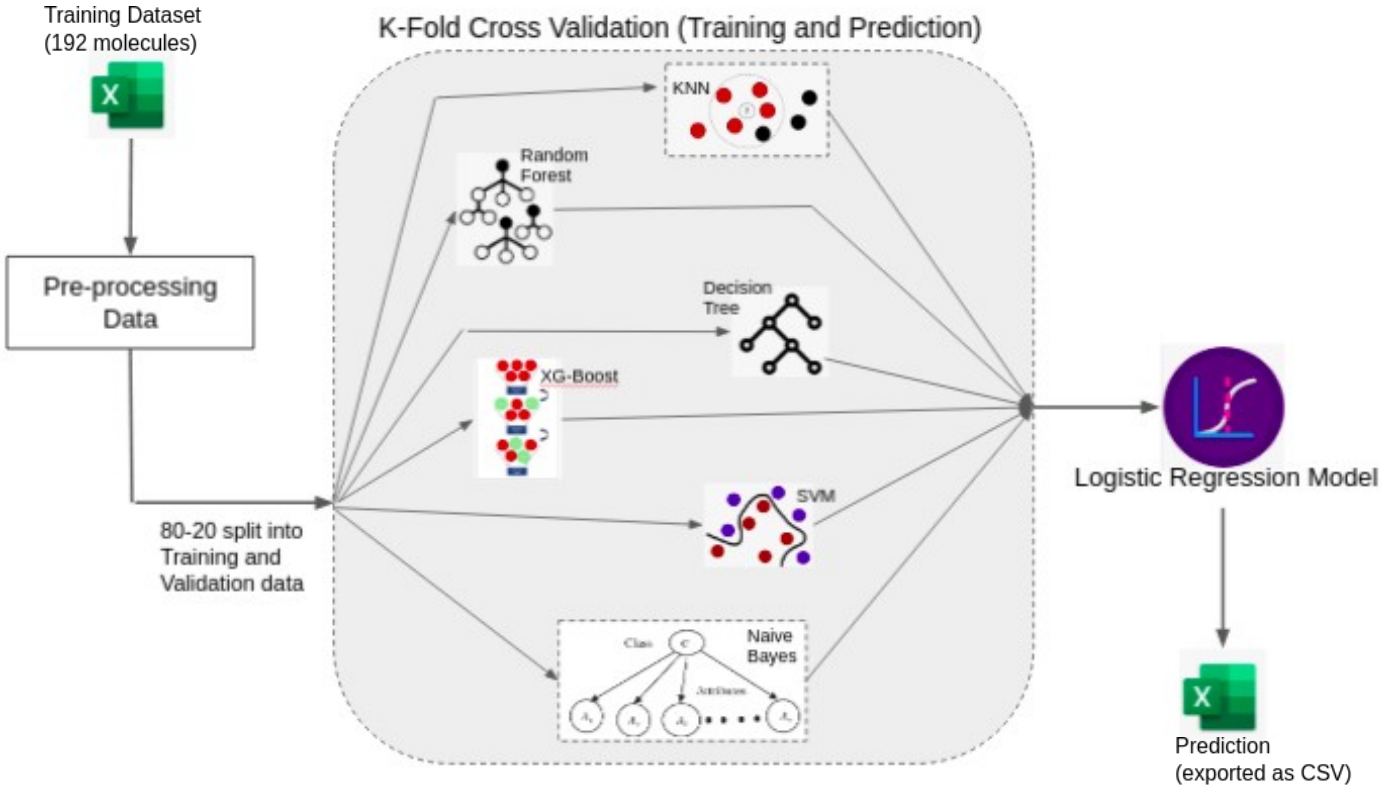

Figure 1: Architecture Diagram of Model created

IV IMPLEMENTATION

For the implementation of these supervised learning models, we have used the sklearn library of machine learning. In the curated dataset, we have denoted substrates as 1 and non-substrates as 0. Various normalization techniques were used to bring the values of all the features in a common range. Missing values were handles along with zero and infinite values. Since the training set was small and numeric features were used, there could be possibilities of overfitting. Hence, feature selection was done before model building, to reduce the risk of overfitting and enhance the performance of model. A special class of Decision trees (C4.5) was also used as one of the base models which avoids overfitting.

Our dataset had a total of 170 compounds in the training set and 23 compounds kept aside for the test set whose labels were unknown. We have first split the training dataset in the ratio of 80:20, to make one training set and validation set. All the 6 models were trained on the 80% split training set and different To assess the performance of these algorithms, we have used several metrics like Accuracy, Precision, Recall and RMSE (Root mean Square Error). Finally we have demonstrated the comparison of the results with the help of appropriate data visualizations and graphs like box plots and pie charts.

features and patterns were learned. We then used the trained model on the 20% part of the dataset (comprising of almost 34 compounds) to predict the labels for the chemical compounds. The labels for these 34 compounds were known already and thus these actual labels were compared with the predicted labels to compute the accuracy of the model. We have used random shuffling every time the model was trained. Hence the accuracy of these different models ranged between 75-89%. We have observed random forest performed very well on our dataset but the results of other models were also promising.

To further enhance the accuracy and to take account of the strengths and weaknesses of the individual models, we have decided to use the combination of these base algorithms. We developed a combined method using the Logistic regression model to generate the final combination decision probability, which showed that the combined methods performed better and was superior to “single” methods.

To overcome the limitations of the single models, we proposed two aggregation schemes to construct consensus models: one simply averages the predictions of single models, taking no account of their contribution difference (averaged consensus), and the other involves performing logistic regression analysis, with the prediction of each individual model as the input variable and the consensus prediction as the output variable (weighted consensus).

Our consensus model was based on the concept of stacking and blending where in the predictions on the test dataset of 34 compounds was used as the training data for the logistic regression model, and results were predicted on the test dataset of 23 compounds. Only 4 out of these 23 compounds were predicted incorrectly and thus the accuracy of the consensus model stood at 81%. By changing the initial split of 80:20 to 70:30, accuracy got enhanced as the training set for the logistic regression got increased.

When we compared the results of the consensus model with the base models, the weighted consensus approach achieved consistently more favorable values across all evaluation metrics, indicating overall improvements in accuracy and stability.

Thus, our consensus model successfully captured the strengths and overcame the deficiencies of the highest performing model.

V RESULTS & CONCLUSION

The consensus model developed classifies the molecule as Substrate (indicated by 1) and Non-Substrate (indicated by 0). We are getting consistent accuracy of about 85-90% when tested over various molecules. On the latest dataset, out of total 22 molecules, 20 molecules were predicted correctly. The average accuracies of each model are as follows :

1. KNN - 86%
2. Random Forest - 90%
3. Decision Tree - 82%
4. Naive Bayes - 82%
5. XG-Boost - 94%
6. SVM --> 80%

Average accuracy of all models ranges from 83% to 89% approximately.

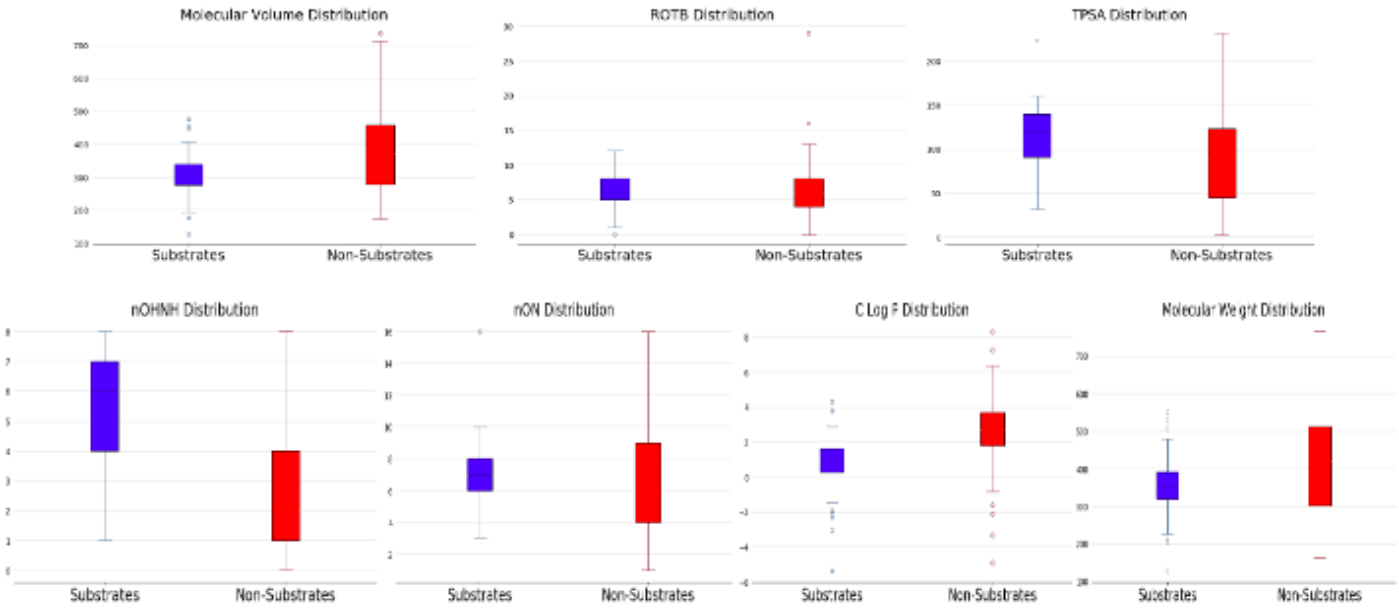

Figure 2: Feature distribution Box plot on Training Set

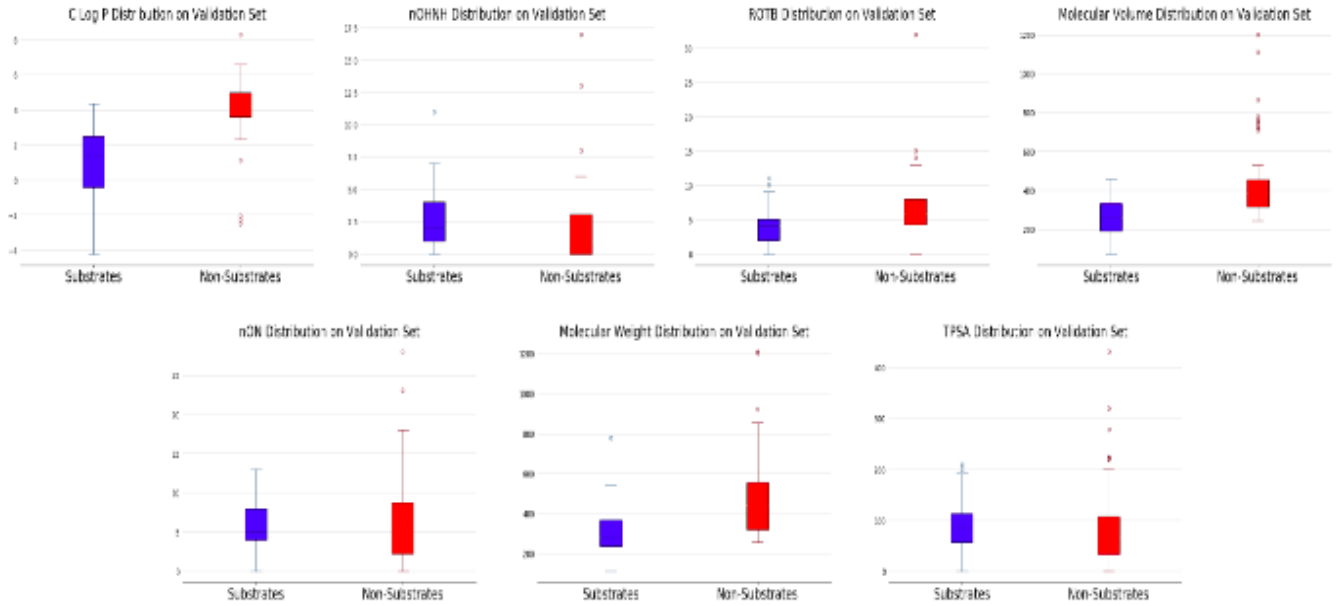

Figure 3: Feature Distribution Box plot on Validation Set

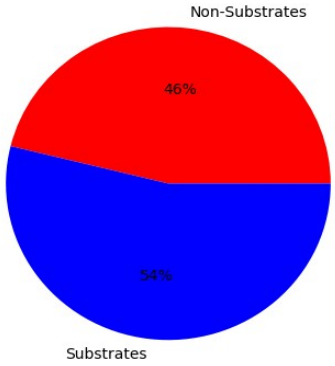

Figure 4: Prediction Statistics of Molecules

VI FUTURE WORK

We have worked upon diverse types of supervised learning models to classify the chemical drug molecules as substrates and non-substrates. We have used an ensemble learning technique using all the supervised models as their base models to give more accurate predictions. In our future work, we intend to use even powerful techniques of computation, Deep Learning to train and test the model. We intend to use Deep neural networks to assess the performance of Deep Learning in computational toxicity prediction. Deep Learning is founded on novel algorithms and architectures for artificial neural networks together with the recent availability of very fast computers and massive datasets. It discovers multiple levels of distributed representations of the input, with higher levels representing more abstract concepts. Currently our dataset is limited, but to train deep learning models we need comparatively large dataset than the supervised learning algorithms. In our next stage of work, we will also try to employ various methods to extend the dataset, both using computation techniques and further study of available literature.

VII ACKNOWLEDGMENT

We are very thankful to Prof. Dr. Chittaranjan Hota and Prof. Dr. Nirmal Jayabalan for their constant guidance and help throughout the project. It would have been difficult to accomplish this without your support. We would also like to thank Ms.Manisha Malani to guide and assist us with the datasets and helping us understand the pharmalogical terminologies, concepts and techniques.

VIII REFERENCES

1 *Baidya, A.T., et al., In silico modelling, identification of crucial molecular fingerprints, and prediction of new possible substrates of human organic cationic transporters 1 and 2. New Journal of Chemistry, 2020.* ***44****(10): p. 4129-4143.*

2 *Predicting Chemical Ocular Toxicity Using a Combinatorial QSAR Approach*

3 *DeepTox: Toxicity Prediction using Deep Learning [https://www.frontiersin.org/articles/10.3389/fenvs.2015.00080/full]*

4 *Helma, C. and J. Kazius, Artificial intelligence and data mining for toxicity prediction. Current Computer-Aided Drug Design, 2006.* ***2****(2): p. 123-133*

5 Jensen, O., J.r. Brockmöller, and C. Dücker, *Identification of Novel High-Affinity Substrates of OCT1 Using Machine Learning-Guided Virtual Screening and Experimental Validation.* Journal of Medicinal Chemistry, 2021

6 Baidya, A.T., et al., *In silico modelling, identification of crucial molecular fingerprints, and prediction of new possible substrates of human organic cationic transporters 1 and 2.* New Journal of Chemistry, 2020. **44**(10): p. 4129-4143.

7 Khuri, N. and S. Deshmukh. *Machine Learning for Classification of Inhibitors of Hepatic Drug Transporters*. in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2018. IEEE.

8 Wu Y, Wang G. Machine learning based toxicity prediction: from chemical structural description to transcriptome analysis. International journal of molecular sciences. 2018 Aug;19(8):2358.

9 Solimeo R, Zhang J, Kim M, Sedykh A, Zhu H. Predicting chemical ocular toxicity using a combinatorial QSAR approach. Chemical research in toxicology. 2012 Dec 17;25(12):2763-9.