Predicting the Toxicity of Chemicals and Drugs using Machine Learning Models

Shraddha Surana[1], Pratik Mahankal[2], Prateek Bihani[3]

[1]Lead Data Scientist at Thought Works Pune Maharashtra, India

[2]Student, Symbiosis skills and professional University, Pune, Maharashtra, India

[3]HBA, Ekklavya Infosys, Pune, Maharashtra, India

Abstract: ***Toxicity is the way to find out if the drug/medicine is harmful to human body. Currently, the toxicity of the medicine is calculated using in-vivo method, where the medicine is tested on the animals and their results are generated. However, this method of toxicity testing for all existing compounds biologically may not be viable financially and logistically. We try to solve this problem by using machine learning and deep learning techniques. We have used the ensemble learning algorithm voting based classifier [logistic regression, decision tree, support vector machines] to predict the toxicity of theTox21 dataset. Where we get the AUC (Area under Curve) of NR-AR-LBD: 0.87 SR-mmp: 0.84 NR-Ahr: 0.81 on these assays.***

Keywords: Machine Learning, Ensemble Learning, Toxicity Prediction, Chemicals, Tox21 dataset

1. Introduction

Drugs have go through certain studies to see if the drugs are not toxic to the human body. They have to go through the necessary clinical trials for their approvals. There is a certain degree of risk involved in the clinical trials because the drugs are tested on the animals. The results offer little guidance to the human body reactions, due to inter-species differences and differential disease models [1].

Therefore, animal experiments cannot simulate the human body's response to new drugs and offer no risk exemption. Also, animals are caused harm due to the testing og drugs on them. One of the statics shows the around 100 million animals are used every year for the clinical trials. The research shows that almost around half of the new drugs were found unsafe for animals [9]. The research also shows that clinical trials provide very less information about the human toxicity reactions. For example, Sitaxentan caused no explicit liver injury in animal experiments, whereas the hepatotoxicity was prominent in humans. This research shows that the drug reacted to the animals differently in comparison to humans. [5]

Current methods for testing the toxicity of a high number of chemicals rely on High-Throughput Screening (HTS). HTS experiments can investigate whether a chemical compound at a given concentration exhibits a certain type of toxicity, for a number of different compounds in parallel. These experiments are repeated with varying concentrations of the chemical compound, which allows us to determine those response curves [2]. Conducting these HTS experiments are time and cost intensive processes. Typically, a compound has to be tested for several types of toxicity at different concentration levels. Thus, the whole procedure has to be re-run multiple times for each compound. Usually, a cell line has to be cultivated to obtain a single data point.

Quantitative structure activity relationship (QSAR) relationships are used to predict the toxicity, behavioral parameters and physiological properties. QSAR are the chemical models that show the patterns between the chemical structures. QSAR relations can be used to predict the toxicity by finding out the similar relationship between the chemicals of the same toxicity [15]. Advanced machine learning algorithms have helped to predict the toxicity of the chemicals when combined with the QSAR relations and HTS. Commonly adopted machine learning algorithms are Support vector machines, decision trees, and k-nearest neighbors [15]. The Tox21[17] program aimed to identify new methods for assessing chemical toxicity in the form of QSAR models in order to improve the identification of chemicals that may affect the functions of seven nuclear receptors (AR, AR-LBD, ER, ER-LBD, Ahr, Aromatase, PPAR-gamma) and five stress response pathways (ARE, ATAD5, HSE, MMP, p53) in the human body.

A chemical structure can be characterized by a series of numerical values known as molecular fingerprints or descriptors. They can show molecule properties such as log P, molecular weight, hydrogen bond donors, acceptors, rotatable bonds, and so on, which be relate to experimental proof. They may also be 2D Fragmentbased fingerprints, such as MACCS, which are represented by bit arrays of 0s and 1s, with each bit location indicating the presence or absence of structural fragments (166 bits). ECFP's are another kind of the fingerprints, which is defined as the Extended Connected fingerprints. The properties of the ECFP include:

- They reflect molecular structures by means of circular atom neighborhoods
- They can be measured very easily
- They are built to reflect both the presence and lack of functionality, both of which are important for the study of molecular behavior

2. Literature Review

Toxicity of the medicine/drug is one of the most important criteria for the examination of the consumables drug. To assess the toxicity of chemicals and medicines, the regulatory agencies require in-vivo testing for several toxic endpoints, leading to many animal experiments conducted annually [19]. This process can be simplified using machine learning and deep learning methods.

The quantitative structure activity relationship (QSAR) approach is one of the most popular and commonly used approaches [20]. QSAR models were widely employed to predict physicochemical properties, environmental behavioral parameters and toxicity of diverse chemicals. The basic QSAR assumption is that similar molecules have similar activities. Thus, by studying the relationship between chemical structures and biological activities, it is possible to predict the activities of new molecules without actually conducting lab experiments. Researchers have implemented the algorithms like ensemble learning combined with the Random forest. There are chemical descriptors which can be used for predicting the toxicity of the chemical compounds. Chemical descriptors are those that are calculated mainly based on the molecular structure-derived information's, atomic types, atomic charges and atomic distances. Among which the most commonly uses is the molecular fingerprints. In which there are different types of fingerprints which can be used to predict the toxicity of the chemical compounds. The MACCS is the most common fingerprint Each of the 166 bits encodes a specific structural characteristic. [20]. In practical combinations of the molecular fingerprints and machine learning algorithms are Pubchem-SVM and MACCSRF. The merits of SVM and RF are apparent. SVM performs the best among many machine learning models, among RF, k-nearest neighbor (k-NN), and naive Bayes [14][6][4][22]. MAP4, ECFP4, MHFP6 are the best performing fingerprints that give good results for the toxicity prediction.[8][12].

3. Methods and Materials

Data

We have used Tox21 data to build the machine learning model. The data contains almost 10,000 samples of the molecules. The toxicity is divided into seven nuclear receptors (AR, AR-LBD, ER, ER-LBD, AhR, Aromatase, PPAR-gamma) and 5 stress response pathways (ARE, ATAD5, HSE, MMP, p53) in the human body. There are four possible assay outcomes for each compound: active, inactive, inconclusive or not tested. Only those chemicals labeled as either active (1) or inactive (0) were retained for this study. For each molecular smiles (molecular formulae) MACCS and ECFP4 fingerprints were calculated using the RdKit and combined them [13]. As the data is highly imbalanced, for each of the toxicity we have used the overbalancing technique to balance the imbalanced the data.[10]

Algorithm

Logistic regression is a statistical model that, in its basic form, models a binary dependent variable using a logistic function. The hypothesis of logistic regression tends to limit the cost function between 0 and 1.

$$Recall = Sensitivity = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$Balanced\ Accuracy\ (BA) = \frac{Sensitivity + Specificity}{2}$$

We have used naive bayes algorithm as it offers fast computational sifier systems, or simply assembly systems. In this study we have used two ensemble models [18].

- Decision Tree, Random Forest and Support Vector Machine
- Logistic Regression, Decision tree, Support Vector Machines

In our study, the Tox21 dataset was highly dimensional and highly imbalanced. For datasets with such large features and a small number of minority class samples, classification often suffers from over fitting. Thus, we have implemented overbalancing technique.

4. Performance Evaluation Matrix

We have used these as performance evaluation matrices to evaluate the models. Binary classification models performances are also being expressed primarily by four ways i.e. 1. True positive (TP) The number of true active chemicals correctly predicted as active by the model 2. False positive (FP) As number of true inactive chemicals incorrectly predicted as active 3. True negative (TN) The number of true inactive chemicals correctly predicted as inactive 4. False negative (FN) The number of true active chemicals incorrectly predicted as inactive. The True Positive Rating (TPR) can also be referred as Sensitivity or Recall which is a fraction of the correctly expected active chemicals. True negative rate (TNR) or precision gives a comparable metric (accuracy) to the inactive (majority) class. Precision derives the probability of a model for making correct active class predictions. Most evaluation metrics can be derived from the above four terms. Precision measures of the above model are a chance of making a successful active class. The F1 score is the harmonic mean of the accuracy and memory. Equally, balanced accuracy (BA) is the mean of sensitivity and specificity all groups.

Time during training and prediction as well as it allow parameter complexity and is not affected by irrelevant features [14].

Support Vector Machine, is a linear model that can be used to solve classification and regression problems. It can solve linear and nonlinear problems and is useful for a wide range of practical applications. SVM is based on a simple concept: The algorithm draws a line or a hyperplane to divide the data into classes.

A Decision tree is a tree structure that looks like a flowchart, with each internal node representing a test on an attribute, each branch represents the test outcome and each leaf node (terminal node) holding a class label [15].

RF is a robust supervised learning algorithm that has been commonly used for classifications in many data science applications. The RF model consists of a number of individual decision trees that function as a set. The individual decision trees are created by randomly selecting the features of each node to decide the break. During classification, each tree vote's and the class with the most votes is the model's prediction. [4]

Ensemble learning is a method in which several models, such as classifiers are strategically created and combined to solve a specific computational intelligence problem. Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.) performance of the model. The Ensemble based method is achieved by integrating a variety of models (classifiers). As a result, such systems are also known as multiple class.

In addition, the two widely used matrices AUC [7] were also calculated using Scikit learn [10] to evaluate and compare the overall performance of the classifiers.

5. Results

Data Curation and Preprocessing

Table 1 shows an overview of the preprocessed training and test datasets of chemicals and their behaviors as determined by 12 in vitro assays. The raw Tox21 datasets contains over 12 thousand chemicals, some of them were retained for each assay after preprocessing. Due to the lack of labels of some of the chemicals were not used. As the data was highly imbalanced we have used the oversampling technique to train the data. The test data was not sampled.

Table 1: Comparison between the training (before sampling and after sampling) data and test data

Assays	Training Data Before Sampling	Training Data After Sampling	Testing Data
NR-AR	7485	14366	1872
NR-Ahr	6531	11542	1633
NR-AR-lbd	6875	13282	1719
NR-armotase	5776	10954	1445
NR-ER	6152	10824	1539
NR-ER-lbd	6998	13304	1750
NR-ppar-gamma	6543	12718	1636
SR-ARE	5731	9696	1433
SR-atad5	7268	13986	1818
SR-HSE	6516	12348	1630
SR-MMP	5852	19784	1464
SR-p53	6903	12918	1726

Selecting the ensemble learning as the base classifier

A comparison of six popular machine learning algorithms, i.e., random forest (RF), logistic regression(LR), decision trees (DT), Naive Bayes (NB), support vector machine(SVM) and Ensemble learning(EL), was done using the training datasets of all the twelve assays. These algorithms were implemented in Scikit-learn with default parameter settings. The purpose of this preliminary study was to select a base classifier from these algorithms. AUC score was calculated and used as the metric to evaluate classification performances. As shown in Table 2, Ensemble learning was the gave the best results for 8 of the 12 assay datasets, including NRAhr, NR-armotase, NR-PPAR-GAMMA, SR-ARE, SR-ATAD5, SR-HSE, SR-MMP, SR-P53. Ensemble learning was the second best performer for others 4 assays. The average score of the Ensemble learning algorithm was 0.78 AUC. The second best algorithm was support vector machine (SVM) with an average score of 0.766 AUC. Thus, ensemble learning algorithm outperformed the other 6 algorithms on the Tox21 dataset. Thus, we chose the ensemble learning algorithm as the best classifier. [Table-2]

Comparing the results of two Ensemble Learning Models

As mentioned in the earlier section [Algorithms Used], we have used two types of ensemble learning algorithm's to explore which works better to predict the toxicity. In the first algorithm we have used the Decision Tree, Random Forest and Support Vector Machine algorithms and in the second we have used Logistic Regression, Decision tree, Support Vector Machine algorithms. In the previous section [Selecting the ensemble learning as the base classifier], we have mentioned the results of the second algorithm.

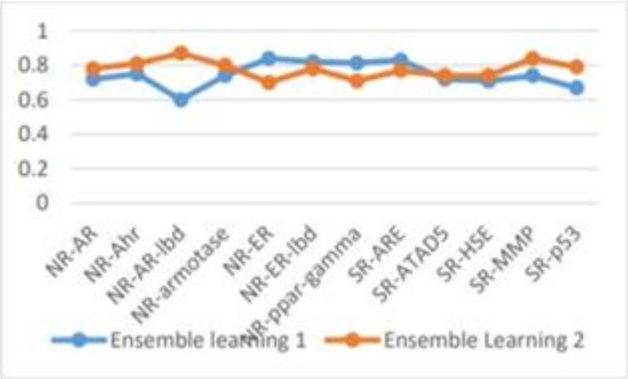

Figure 1: Comparison between two Ensemble learning model

Table 2: Comparison between the results of the different models according to the AUC

Assays	Ensemble Learning	Logistic Regression	Naive Bayes	Support Vector Machine	Random Forest	Decision Tree
NR-AR	0.786	0.80	0.704	0.786	0.789	0.77
NR-Ahr	0.81	0.786	0.61	0.78	0.74	0.74
NR-AR-LBD	0.87	0.86	0.88	0.87	0.86	0.84
NR-armotase	0.80	0.78	0.60	0.78	0.72	0.74
NR-ER	0.70	0.69	0.55	0.71	0.68	0.67
NR-ER-lbd	0.785	0.781	0.70	0.786	0.770	0.772
NR-ppargamma	0.66	0.67	0.65	0.69	0.61	0.663
SR-ARE	0.770	0.75	0.557	0.76	0.71	0.70
SR-atad5	0.74	0.73	0.66	0.72	0.65	0.68

SR-HSE	0.74	0.73	0.66	0.72	0.65	0.68
SR-mmp	0.84	0.82	0.57	0.83	0.79	0.798
SR-p53	0.797	0.794	0.593	0.78	0.71	0.75

Performance evaluation for the Ensemble Learning Model

The goal behind using ensemble learning algorithm is to combine the predictions of several base estimators to enhance the robustness over one estimator. The ensemble learning algorithm [Logistic Regression, Decision tree and Support Vector Machines] is trained to classify the toxicity. In this, we have used the voting based ensemble learning method, where we have used the hard voting method to classify the toxicity for every assay. We have used the hard voting method to get the prediction from all the classifiers. The class voted the most by the classifiers is the output of ensemble learning model. The decision tree and logistic regression were trained with default parameters, whereas the support vector machine was trained with the kernel as a polynomial with a degree of two. For each of the 12 assays, we have trained all 6 machine learning algorithms on the training dataset and then the trained models were used to determine the score of the performance evaluation matrices. We have omitted the accuracy (the ratio of accurate predictions to the total number of chemicals) from the matrices panel shown in Table 3 because the accuracy can be deceptive while assessing the model's success for a strongly imbalanced classification. In particular, a high accuracy does not mean that the prediction model is capable of accurately predicting the uncommon classes.

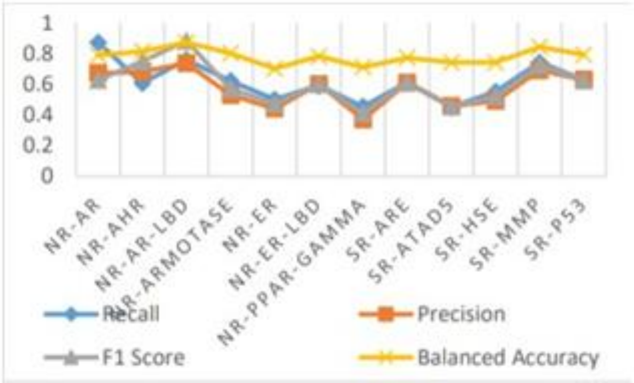

Figure 2: Performance Evaluation For Ensemble Learning model

Table 3: Performance Evaluation For Ensemble Learning

Assays	Recall	Precision	F1 Score	Balanced Accuracy
NR-AR	0.87	0.67	0.626	0.787
NR-Ahr	0.606	0.68	0.746	0.812
NR-ARLBD	0.757	0.735	0.88	0.87
NR-armotase	0.62	0.53	0.57	0.800
NR-ER	0.50	0.44	0.47	0.70
NR-ER-lbd	0.59	0.60	0.60	0.78
NR-ppargamma	0.453	0.37	0.411	0.71
SR-ARE	0.612	0.608	0.612	0.77
SR-atad5	0.45	0.457	0.453	0.74
SR-HSE	0.55	0.49	0.52	0.74
SR-mmp	0.74	0.69	0.72	0.84
SR-p53	0.626	0.626	0.626	0.79

6. Conclusion

Due to the high imbalance ratio, the algorithms tend to easily overfit on the data and do not give correct results. Thus, we have applied the sampling techniques to avoid overfitting. However, using the undersampling technique might result in the loss of information. Thus, to avoid the overfitting of the data we have used the oversampling technique. In this study, we have used a combination of MACCS and ECFP4 as fingerprints. Evaluation matrices were used to predict the reliability of the machine learning algorithms. The highest average AUC achieved was 0.78 by the ensemble learning model. The outputs will benefit and enable greater use of in-silico toxicity models as a decision-making tool to assess the potential health risks of chemicals and drugs.

References

[1] Aysha Akhtar. “The Flaws and Human Harms of Animal Experimentation”. In: Cambridge quarterly of healthcare ethics: CQ: the international journal of healthcare ethics committees24 (Sept. 2015) pp. 407–19.DOI:10. 1017 /S0963180115000079.

[2] Amancio Carnero. “High throughput screening in drug discovery”. In: Clinical translational oncology: official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico8 (Aug. 2006) pp. 482–90.DOI:10.1007/s12094-006-0048-2.

[3] ChemAxon. ChemAxon Dataset. Feb. 2014.URL: https://docs.chemaxon.com / display / docs / extended connectivity fingerprint -ecfp.md.

[4] Adele Cutler, David Cutler, and John Stevens. “Random Forests”. In: vol. 45. Jan.2011, pp. 157–176.ISBN: 978-14419-9325-0.DOI:10.1007/978-1-4419-9326-7_5.

[5] John Erve et al. “Bioactivation of Sitaxentan in Liver Microsomes, Hepatocytes, and Expressed Human P450S with Characterization of the Glutathione Conjugateby Liquid Chromatography Tandem Mass Spectrometry.” In: Chemical research intoxicology26 (May 2013) DOI:10.1021/tx4001144

[6] Theodoros Evgeniou and Massimiliano Pontil. “Support Vector Machines: Theory and Applications”. In: vol. 2049. Jan. 2001, pp. 249–257.DOI:10.1007/3-540-44673-7_12.

[7] Peter Flach, Jose Hernandez-Orallo, and C‘esar Ferry. “A Coherent Interpretation of AUC as a Measure of Aggregated Classification Performance.” In: Jan. 2011, pp. 657–664.

[8] Martin G utlein and Stefan Kramer. “Filtered circular fingerprints improve either prediction or runtime performance while retaining interpret ability”. In: Journal of Cheminformatics8 (Oct. 2016) DOI: 10.1186/s13321-016-0173-z.

[9] Thomas Hwang et al. “Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results”. In: JAMA internal medicine 176(Oct. 2016)

DOI:10.1001/jamainternmed.2016.6008.

[10] Guillaume Lemaitre, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". In: Journal of Machine Learning Research18.17 (2017) pp. 1–5.URL:http://jmlr.org/papers/v18/16-365.html.

[11] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: Journal of Machine Learning Research12 (2011) pp. 2825–2830.

[12] Daniel Probst and Jean-Louis Reymond. "A Probabilistic Molecular Fingerprint for Big Data Settings". In: (Oct.2018) DOI:10.26434/chemrxiv.7176350.
rdKit. .URL:http://rdkit.org/.

[13] Irina Rish. "An Empirical Study of the Na ıve Bayes Classifier". In:IJCAI 2001 Work Empir Methods Artif Intell3 (Jan. 2001)

[14] Lior Rokach and Oded Maimon. "Decision Trees". In: vol. 6. Jan. 2005, pp. 165–192.DOI:10.1007/0-387-25465-X_9.

[15] Weihao Tang et al. "Deep learning for predicting toxicity of chemicals: a mini review". In: Journal of En vironmental Science and Health, Part C36 (Mar. 2019)pp. 1–20.DOI:10.1080/10590501.2018.1537563.

[16] Tox21.Tox21 Dataset. Feb. 2014.URL:https://ncats.nih.gov/tox21

[17] Giorgio Valentini and Francesco Masulli. "Ensembles of Learning Machines". In:vol. 2486. May 2002, pp. 3–22.ISBN: 978-3-540-44265-3.DOI:10.1007/3-540-45808-5_1

[18] Wikipedia. URL:https://en.wikipedia.org/wiki/In_v*ivo*.

[19] Yunyi Wu and Guanyu Wang. "Machine Learning Based Toxicity Prediction: From Chemical Structural Description to Transcriptome Analysis". In: International Journal of Molecular Sciences19 (Aug. 2018) p. 2358.DOI:10. 3390 /ijms19082358.

[20] Hongbin Yang et al. "In Silico Prediction of Chemi cal Toxicity for Drug Design Using Machine Learning Meth ods and Structural Alerts". In: Frontiers in Chemistry6 (Feb. 2018)DOI:10.3389/fchem.2018.00030.

[21] Zhongheng Zhang "Introduction to machine learning: Knearest neighbors". In: Annals of Translational Medicine4 (June 2016) pp. 218–218.DOI:10.21037/atm.2016.03.37.