IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 33, No. 5, October, 2019, pp. 387-392
Journal homepage: http://iieta.org/journals/ria

Optimized Ensemble Machine Learning Framework for High Dimensional Imbalanced Bio Assays

Rohit Sharma*, Nishtha Hooda

Computer Science and Engineering Department, Chandigarh University, Mohali, Punjab 140413, India

Corresponding Author Email: nishthae7689@cumail.in

https://doi.org/10.18280/ria.330509

Received: 22 July 2019
Accepted: 30 September 2019

Keywords:
machine learning, ensemble, bioassays, SMOTE, drug prediction

ABSTRACT

In pharmaceutical research, a recent hotspot is the study of the activity of bioactive compounds and drugs with computational intelligence. The relevant studies often adopt machine learning techniques to speed up the modelling, and rely on bioassay to evaluate the effect and potency of a compound or drug. This paper aims to design an efficient and accurate method to assess the activity of bioactive compounds and drugs. First, the authors performed virtual screening on the data on bioactive compounds and drugs, eliminating the imbalanced classes and high dimensionality of drug descriptors. Next, eight machine learning algorithms, namely Bayes Net, Naive Bayes, SMO, J48, Random Forest, AdaBoost, AdaBag and logistic regression, were trained by the virtually screened data, and used to predict the activity or inactivity of a drug through bioassays. The synthetic minority oversampling technique (SMOTE) was employed to solve the numerous imbalanced datasets in bioassay. On this basis, the ensemble machine learning model of random forest was optimized. Experimental results show that the optimized random forest machine learning framework achieved better results than the other ensemble-based machine learning methods. The research provides an effective way to perform bioassays on high-dimensional imbalanced data.

1. INTRODUCTION

Pharmaceutical drugs substantially affecting the life of people. A human requires drugs for the prevention of disease or to help in diagnosis or illness. But the process of drug design and development is quite expensive to buy. Computational Intelligence can help to reduce the cost of the drug development process [1]. The report of Global view Research in 2018 implies that global drug discovery market estimation in 2016 was $713.4 million and it is anticipated to the progress at a CAGR (Compound Annual Growth Rate) of 12.6% by 2025. Artificial intelligence and machine learning noble techniques have helped many researchers in finding cost effective solution in diverse domains like drugs discovery, audits, etc. [2-4]. By using Artificial Intelligence in drug discovery, it increases the drugs market rapidly. By 2028, Bekyle indicates that Artificial Intelligence has the power to save $70 million in drug discovery [1].

Virtual screening is the computational technique of biological compounds and supplements the HTS procedure, used in drug discovery to search libraries of small molecules. It automatically evaluates very large libraries of compounds using computer programs. There are two types of compounds present in the bioassays. The compounds which are used for the prevention of disease are called active drugs and other compounds that are used for balancing the drug molecules and power of drugs, are called inactive drugs [5].

PubChem is a public source database of chemical molecules and their activities against the organic assays [6]. It is kept up by the National Center for Biotechnology Information (NCBI), a segment of the National Library of Medicine, which is a piece of the United States National Institutes of Health (NIH) [7]. The discovery of new drugs required High-Throughput Screening (HTS) and it is a very tedious task that takes 10-15 years to bring drugs into the market. In HTS, the number of compound batches are screened against bioassay to examine the compound capability to combine the objective. The strenuous efforts are required to process such high-dimensional virtual screening data with class imbalance issues. It means that there are numerous features in the data and the number of active compounds is very less than inactive compounds [5].

In this paper, the machine-learning framework is proposed to overcome the high dimensionality and class imbalance problems in virtual screening data and to predict the active and inactive compounds in BioAssays. To handle high dimensionality issues, feature selection is employed. Gini index ranker is used to extract the important features and remove the irrelevant ones. For class balancing SMOTE algorithm is used. This algorithm balances the imbalance classes by taking the nearest neighbors and generates the same features for the target class [8].

Rest of the paper organized as follows: Section 2 briefly describes various methods and the related work. Section 3 defines the dataset used and the tentative view of the frameworks. Section 4 describe dataset description and experimental settings. Section 5 shows the results and comparison of the framework. Finally, Section 6 conclusion of the paper.

2. RELATED WORK

Most part of the research has concentrated regarding the issue of drug discovery over virtual screening by different strategies with the point of discover active chemicals. Daniel

P. Russo et al. proposed an automated extraction using a computational method for bioassay data from a public source library and predict the toxicity in animals using a novel bio profile-based read-across approach. In this work, the relevant toxicity mechanism and acute oral toxicity were identified by using a novel subspace-clustering algorithm [9]. Conrad Stork et.al proposed a machine-learning model in which they use some silico methods for the prediction of the frequent hitters or problematic compounds. The models were taken from the PubChem BioAssay database consisting of 311k compounds [10]. Manole-Stefan Niculescu proposed an automated optical method experimented on a Siemens Dimension EXL200 analyzer. This method is used for improving the accuracy of biochemical assays. It has a cuvette window, which is used to examine the quantity of analyte from the cuvette. The cuvette window analyzes the samples for every time to maintain some assay conditions [11]. Ming Hao el.al proposed an algorithm, GLM Boost that combines with SMOTE to get the better results of the problem of several mismatch datasets from PubChem BioAssay. By applying the proposed model, those samples have poor results generated can be detected as high balance accuracy [12]. Bin Chen et.al proposed predictive models that enabled "virtual screens" to identify compounds in a large dataset. In this paper, they examine the quality of Naive Bayesian predictive models constructed using BioAssay data [6].

3. PROPOSED WORK

Figure 1 presented the abstract view of the prediction model. The BioAssay Data, which contains Virtual Screening Primary and Confirmatory BioAssays is given to the machine learning framework. In the machine learning framework first, it splits that data into two parts training and testing. The given data is highly imbalanced and has many features. Therefore, for balance the unbalanced data SMOTE algorithm is applied. In the next step, the Gini index is applied to the training data. It aims to choose the best subset features which are useful and good relation with the target feature. After selecting important features, examine different classifiers on the training data by validating with the test data and in last check the overall performance of the model, which classifier gives the best results.

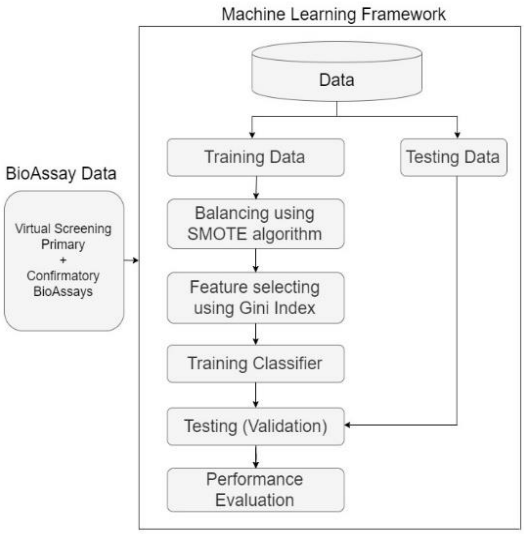


Figure 1. Abstract view

3.1 Class balancing by SMOTE algorithm

Preprocessing means preparation of data before training the classifier. In our dataset, the biochemical composites used in virtual screening are imbalanced classes of active and inactive compounds. Therefore, to balance the dataset Synthetic minority Oversampling Technique (SMOTE) is applied, which increases the number of minority classes until the data was balanced.

SMOTE algorithm is used to solve the problem of an imbalance of data. It is the oversampling method, which generates the virtual training records for the minority class. These records are generated by randomly for selecting the nearest neighbors by using k-nearest neighbor's algorithm for the minority class. After this, the information is remade and few classification models can be applied for the processed information [6].

Algorithm 1. SMOTE algorithm

Comment: {rand (0, 1) denotes the random numbers between zero and one.}

1. Set the minority class N for each x ∈ K. The k-nearest neighbor is calculating the Euclidean distance between x and in set N.
2. The imbalanced proportion sets the testing rate k. For each x ∈ K, K samples are casually chosen from its k-nearest Neighbor and they build the set N1.
3. Every sample xr ∈ N1(r = 1,2,3,....N), this is used to generate the new sample x = x + rand(0,1) ∗ |x–xr| [6].

3.2 Feature selection

It is the most significant preprocessing stage that applied before the training classifier. It aims to pick the subset features, which are extra important and have a good relationship with the target. In this paper, the Gini index is applied to the training data to select the best features from the highly imbalanced data. Gini (D) is defined as:

$$\text{Gini (D)} = 1 - \sum_{j=1}^{n} p_j^2 \quad (1)$$

$$\text{pj} = \frac{\text{count of specific class label}}{\text{total count of D}}$$

where, pj describes the relative frequency of class j in D [13].

3.3 Ensemble classification

The machine learning framework is proposed for the extrapolation of active and inactive compounds in BioAssay data. There are eight distinct models namely, Bayes Net (BN), Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM/SMO), Random Forest (RF), Adaboost, Adabag, and J48 are used. The Bayes Net classifier is related to the Direct Acyclic Graph whose knobs represent variables in the Bayes Net perception [14]. Naive Bayes classifier are a collection of algorithms where each algorithm has the same principle. It is based on the Bayes Theorem [15]. The purpose of using Logistic Regression is to model probability of events like true/false, binary 0/1, pass/fail, win/loose alive/dead or healthy/sick [16]. The Support Vector Machine (SVM/SMO) is used to solve the quadratic optimization problems during train the support vector machine [17].

Random Forest is ensemble classifier and is made by a combination of multiple decision trees. It requires more computer memory as compared to other classifiers [18]. Adaboost is the combination of multiple classifiers into a single classifier and it is used to solve both classification and regression problems [19]. Adabag is the combination of individual classifiers that built in the training sets of the bootstrap copies [20]. J48 labeled the input data and based on that data it creates a decision tree that is used for classification [21]. After analyzing these algorithms, optimize the best machine learning algorithm to improve the results.

4. EXPERIMENTAL INVESTIGATION

This segment briefly explains the dataset description and experimental view of the model.

4.1 Dataset description

21 data set were created using High Throughput Screening (HTS) technology. Each dataset has 21 training and 21 testing subsets in the database. The further whole dataset is divided into two parts, primary and confirmatory in which 7 datasets are of primary screening, 10 datasets are of confirmatory screening and four datasets are both primary and secondary (confirmatory) screening [5]. The primary screening of bioassay allows direct high throughput binding measurement of small compounds without classifying it. On the other hand, mistakes occur during primary HTS for that secondary screening performs on the primary screening data. It is designed to confirm successes competence by a series of useful assays. The major role of secondary screening is to identify the useful response of composites somewhat creating it be a high throughput format [22].

In this research, the primary dataset is taken for the prediction of bioassays. The dataset contains 4279 compounds. The biological compounds are selected on the basis of preliminary virtual screening of approximately 480,000 drug-like small molecules from the Chemical Diversity Laboratories. It contains 144 features of 3423 of different drug instances. The parameters include binary features that help in the prediction of active and inactive drugs. This dataset is taken from the UCI machine learning repository [23].

4.2 Experimental setting

Various dissimilar model building methods are executed using R language. The main goal of this experiment is to calculate the accuracy and make predictions of the classifier. The dataset is divided into two forms training and testing data. To train the model, training data is fed to the classifier after preprocessing the data. Then it is validated by using test data. To evaluate results of the anticipated framework numerous evaluation metrics such as accuracy, TP rate, FP rate, F-measure, MCC, SAW score, and ROC are used.

5. RESULT AND DISCUSSION

This segment describes the results, performance estimation, and comparison of the framework on benchmark dataset.

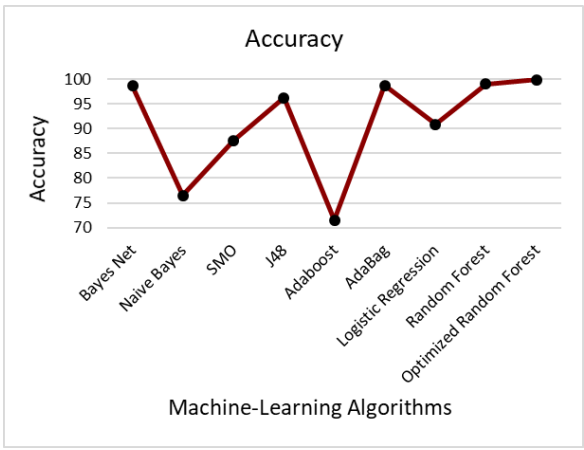
5.1 Performance evaluation

The performance of the proposed framework is tested using various metrics [24] like Accuracy, TP Rate, FP Rate, F-Measure, MCC, and AUC presented in Table 1. It shows the performance comparison of optimized Random Forest with other classifiers. On the basis of Table 1, various graphs are plotted to show the comparative analysis of different classifiers. In Figure 2, a performance comparison of the machine learning framework is presented graphically. Figure 2(a) shows the accuracy of different classifiers. In this Figure, Optimized Random Forest has the highest accuracy among all because it is an ensemble tree-based structure model that is made up of many trees. It creates different small models by taking two or more features randomly and make a final model based on their results. Figure 2(b) shows the ROC curve, which means how classifiers are performing generally. It plots the true positive rate against the false positive rate. The model which has a higher positive rate gives better performance. In this, Optimized Random Forest has the highest ROC among all classifiers. Figure 2(c) shows the TP Rate, which means the correctly classified instances. Optimized Random Forest, Adabag and Bayes Net perform good to classify the instances correctly because Random Forest and Adabag bothe are ensembled algorithms and Bayes Net is the type of probabilistic graphical model that uses bayesian inferences. Figure 2(d) shows the FP Rate, which means falsely classified instances. Bayes Net perform well to falsely classify the instances among all because it is a graph-based model that uses the concept of probability distribution and probability theory for predictions [20].

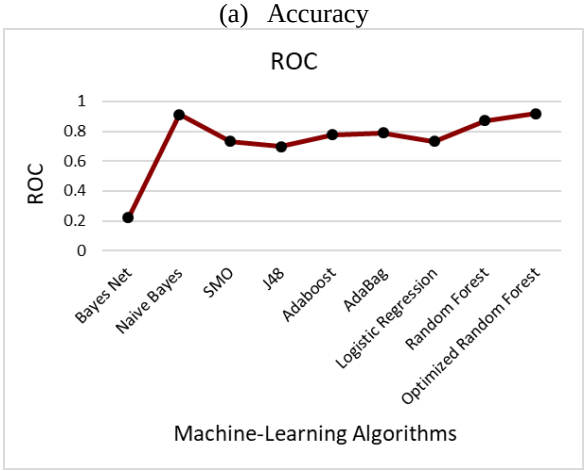
These metrics performed on different classifiers like Bayes Net (BN), Naive Bayes (NB), Logistic Regression (LR), SVM/SMO, Random Forest (RF), Adaboost, Adabag, and J48 [2]. In all these classifiers, it can be observed that Random Forest gives the highest accuracy and Adaboost has the lowest, which is 71%. The ROC curve value of Naive Bayes and Random Forest is highest among others.

Table 1. Performance comparison of optimized random forest with other classifiers

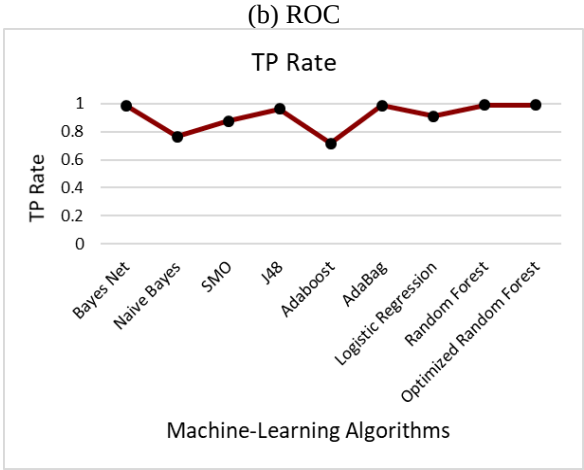
Classifier	Bayes Net	Naive Bayes	SVM	J48	Adaboost	Adabag	Logistic Regression	Random Forest	Optimized Random Forest
Accuracy(%)	98.60	76.52	87.62	96.26	71.49	98.71	90.89	98.94	**99.81**
TP Rate	**0.99**	0.77	0.88	0.96	0.72	**0.99**	0.91	0.99	**0.99**
FP Rate	**0.99**	0.00	0.41	0.58	0.42	0.49	0.58	0.58	0.66
F-Measure	**0.99**	0.85	0.92	0.97	0.82	**0.99**	0.94	**0.99**	**0.99**
MCC	0.32	0.21	0.16	0.25	0.08	0.52	0.14	**0.54**	**0.54**
AUC	0.22	0.91	0.73	0.70	0.78	0.79	0.73	0.87	**0.92**
SAW Score	0.78	0.74	0.78	0.85	0.72	0.92	0.82	0.97	**1.00**



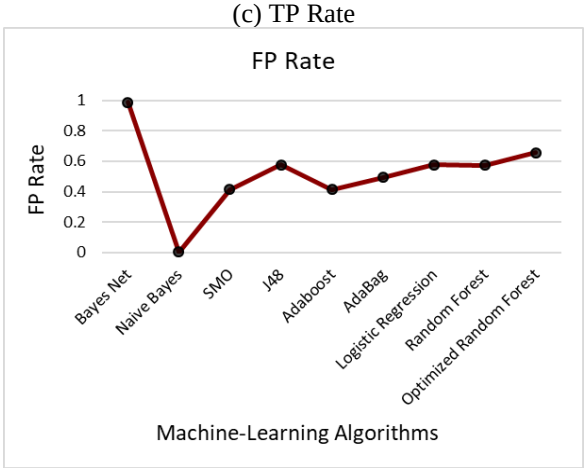
(a) Accuracy



(b) ROC



(c) TP Rate



(d) FP Rate

Figure 2. Comparison chart of machine learning framework with classifiers using accuracy, ROC, TP rate, and FP rate

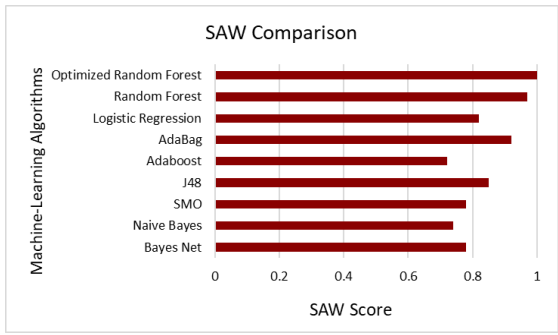
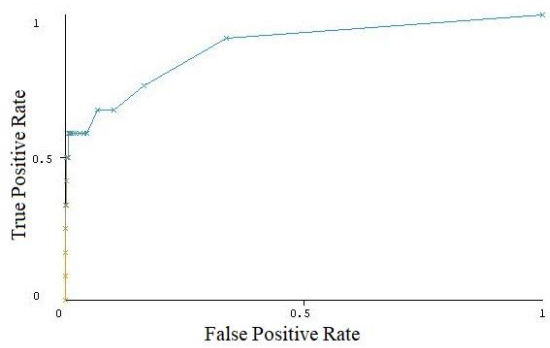
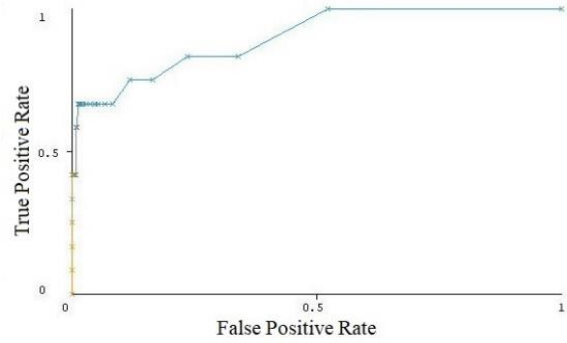


Figure 3. SAW Comparison graph



(a) Threshold curve of random forest



(b) Threshold curve of random forest with SMOTE

Figure 4. Threshold curve of random forest (class value active)

In Figure 3, represents the comparison of machine learning algorithms and compute the SAW score for all of them, which is displayed in the Table 1. In SAW comparison, the graph shows that the optimized Random Forest has the highest SAW score. Adabag also shows a good score as it is an inbuilt ensemble classifier. The purpose of computing SAW score is to explore the comprehensive performance of the machine learning models and to give the best prediction comparison with the other models. In Figure 4, a threshold curve is represented in which the line describes the probability of positive instances which are higher than 0.50. The graph is drawn between the False Positive Rate and True positive Rate. Figure 4(a) displayed the Threshold curve of Random Forest for active instances. Figure 4(b) presented the Threshold curve of Random Forest with SMOTE for the active instances.

5.2 Performance comparison of optimized ensemble on benchmark dataset

This section shows the comparison of the proposed optimized ensemble on a benchmark dataset available at UCI

machine learning repository [21]. The results are presented in the Table 2. The evaluation metrics namely [20] TP Rate, FP Rate, Precision, and Accuracy of the proposed model is compared with the results of the past research work on the same dataset. In the previous work, the researcher has implemented Naive Bayes, Random Forest, SMO, and J48 for training the model. Highest accuracy of 85.16% has been achieved by J48 algorithm. Random forest gave an accuracy of 81.19% only. The proposed optimized version of Random forest in the present research work worked more efficiently than the past research work after the employment of pre-processing techniques like feature selection and class balancing. It has been observed that the proposed Optimized Random Forest gives the highest accuracy and better predictive performance as compared to the other models.

Table 2. Comparison of proposed ensemble with methods implemented on benchmark dataset

Classifier	TP Rate	FP Rate	Precision	Accuracy(%)
Naive Bayes	0.75	0.19	0.99	80.84
Random Forest	0.83	0.18	0.99	81.19
SMO	0.75	0.14	0.99	84.93
J48	0.75	0.14	0.99	85.16
Optimized Random Forest(Proposed)	**0.99**	0.66	0.98	**99.00**

6. CONCLUSION

In this paper, an efficient ensemble machine learning framework is proposed to train a high dimensional and highly imbalance virtual screening Bioassay data. The most popular feature selection ranker namely, the Gini index is employed to find the best and relevant features for training the computational model. Additionally, the SMOTE algorithm is implemented to balance the active and inactive classes in the training data. On testing the proposed model using the benchmark dataset, the accuracy, TP rate, FP rate, F-Measure, MCC, AUC are found to be 98.94%, 0.99, 0.66, 0.99, 0.51, and 0.92 respectively. This efficient framework based on ensemble machine learning algorithm can also be used as a decision system for the extrapolation of active and inactive compounds. In the future, the machine-learning framework will be enhanced by applying it to various big data techniques like Hadoop, Spark, etc.

REFERENCES

[1] Budek, K. (2019). Machine learning in drug discovery. Retrieved from https://deepsense.ai/machine-learning-in-drug-discovery, accessed on 14 November 2019.

[2] Hooda, N., Bawa, S., Rana, P.S. (2018). B2FSE framework for high dimensional imbalanced data: A case study for drug toxicity prediction. Neurocomputing, 276: 31-41. https://doi.org/10.1016/j.neucom.2017.04.081

[3] Hooda, N., Bawa, S., Rana, P.S. (2018). Fraudulent firm classification: A case study of an external audit. Applied Artificial Intelligence, 32(1): 48-64. https://doi.org/10.1080/08839514.2018.1451032

[4] Hooda, N., Bawa, S., Rana, P.S. (2019). Optimizing fraudulent firm prediction using ensemble machine learning: A case study of an external audit. Applied Artificial Intelligence, 34(1): 20-30. https://doi.org/10.1080/08839514.2019.1680182

[5] Schierz, A.C. (2009). Virtual screening of bioassay data. Journal of Cheminformatics, 1(1): 1-21. https://doi.org/10.1186/1758-2946-1-21

[6] Chen, B., Wild, D.J. (2010). PubChem BioAssays as a data source for predictive models. Journal of Molecular Graphics and Modelling, 28(5): 420-426. https://doi.org/10.1016/j.jmgm.2009.10.001

[7] Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., Zhou, Z.G., Han, L.Y., Karapetyan, K., Dracheva, S., Shoemaker, B.A., Bolton, E., Gindulyte, A., Bryant, S.H. (2011). PubChem's BioAssay database. Nucleic acids Research, 40(D1): D400-D412. https://doi.org/10.1093/nar/gkr1132

[8] Luengo, J., Fernández, A., García, S., Herrera, F. (2011). Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling. Soft Computing, 15(10): 1909-1936. https://doi.org/10.1007/s00500-010-0625-8

[9] Russo, D.P., Strickland, J., Karmaus, A.L., Wang, W., Shende, S., Hartung, T., Aleksunes, L.M., Zhu, H. (2019). Nonanimal models for acute toxicity evaluations: applying data-driven profiling and read-across. Environmental Health Perspectives, 127(4): 047001. https://doi.org/10.1289/EHP3614

[10] Stork, C., Wagner, J., Friedrich, N.O., de Bruyn Kops, C., Šícho, M., Kirchmair, J. (2018). Hit dexter: A machine-learning model for the prediction of frequent hitters. ChemMedChem, 13(6): 564-571. https://doi.org/10.1002/cmdc.201700673

[11] Niculescu, M.S. (2017). Optical method for improving the accuracy of biochemical assays. In 2017 E-Health and Bioengineering Conference (EHB), pp. 381-385. https://doi.org/10.1109/EHB.2017.7995441

[12] Hao, M., Wang, Y., Bryant, S.H. (2014). An efficient algorithm coupled with synthetic minority over-sampling technique to classify imbalanced PubChem BioAssay data. Analytica Chimica Acta, 806: 117-127. https://doi.org/10.1016/j.aca.2013.10.050

[13] Manek, A.S., Shenoy, P.D., Mohan, M.C., Venugopal, K.R. (2017). Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World Wide Web, 20(2): 135-154. https://doi.org/10.1007/s11280-015-0381-x

[14] Nielsen, T.D., Jensen, F.V. Bayesian networks and decision graphs. Springer Science & Business Media., accessed on 4 June 2009.

[15] Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E. (2006). Machine learning: A review of classification and combining techniques. Artificial Intelligence Review, 26(3): 159-190. https://doi.org/10.1007/s10462-007-9052-3

[16] King, G., Zeng, L. (2001). Logistic regression in rare events data. Political Analysis, 9(2): 137-163. https://doi.org/10.1093/oxfordjournals.pan.a004868

[17] Platt, J.C. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. ISBN: 0-262-19416-3, 185-208.

[18] Belgiu, M., Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing,

114: 24-31. https://doi.org/10.1016/j.isprsjprs.2016.01.011
[19] Dietterich, T.G. (2000, June). Ensemble methods in machine learning. In International Workshop on Multiple Classifier Systems, pp. 1-15. https://doi.org/10.1007/3-540-45014-9_1
[20] Alfaro, E., Gamez, M., Garcia, N. (2013). Adabag: An R package for classification with boosting and bagging. Journal of Statistical Software, 54(2): 1-35. https://doi.org/10.18637/jss.v054.i02
[21] Bhargava, N., Sharma, G., Bhargava, R., Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, 3(6): 1114-1119.
[22] Baell, J.B., Holloway, G.A. (2010). New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. Journal of Medicinal Chemistry, 53(7): 2719-2740. https://doi.org/10.1021/jm901137j
[23] Uci machine learning repository, Pubchem bioassay data URLhttps://archive.ics.uci.edu/ml/datasets/PubChemBioas sayData, accessed on 17 November 2019.
[24] Jeni, L.A., Cohn, J.F., De La Torre, F. (2013). Facing imbalanced data--recommendations for the use of performance metrics. In 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 245-251. https://doi.org/10.1109/ACII.2013.47