Voting-Based Ensemble Method for Prediction of Bioactive Molecules

Olutomilayo Olayemi Petinrin[1], Faisal Saeed[1,2]*, Tawfik Al-Hadhrami[3]
[1]Information Systems Department, Faculty of Computing, Universiti Teknologi Malaysia
81310 Johor Bahru, Johor, Malaysia
[2]College of Computer Science and Engineering, University of Taibah, Medina, Saudi Arabia
[3]School of Science and Technology, Nottingham Trent University, Nottingham, United Kingdom
e-mail: olutomilayo.petinrin@gmail.com, faisalsaeed@utm.my, tawfik.al-hadhrami@ntu.ac.uk

***Abstract*—Machine learning has its tentacles spread over all major areas of science. The current rise in the amount of data being generated as necessitated its adoption in virtually all aspects including chemoinformatics. Several machine learning methods have been applied to the drug discovery process due to the importance of prediction of bioactivity before the release of drug into the market. The need for the most accurate method is hence evident. Majority voting ensemble is a method whose application is rare in predicting bioactive molecules. This study applies the method using different combination of commonly used classifiers as the base classifier on a chemical dataset of 8294 instances and 1024 attributes retrieved from the MDL Drug Data Report (MDDR). The accuracy of majority voting with the best combination of classifiers is found to be higher than the accuracy of the commonly used classifiers in the field, and makes it suitable for large chemical datasets.**

Keywords-bioactivity prediction; chemoinformatics; drug discovery; majority voting; voting ensemble

I. INTRODUCTION

According to [1], chemoinformatics is the mixing of information resources to transform data into information and information into knowledge for the intended purpose of making better and faster decisions in the area of drug lead identification and organization. Areas such as lead optimization, drug target discovery, Quantitative Structure Activity Relationship (QSAR), Quantitative Structure Property Relationship (QSPR) are part of chemoinformatics, in addition to predicting the properties of biological molecules from their structural similarity as discussed in [2]. According to [3], analysing and predicting bioactive molecules is the most popular task of chemoinformatics. It is essential that more effort is put into discovery of drugs to make it efficient to tackle several diseases. Pharmaceutical companies spend a lot of time and resources in producing a new drug and after production, if it does not meet the target requirement or tackle the disease for which it was produced, the drug will be called off the market hence resulting in waste of time, and resources. According to [4], it costs about $1.8 billion to bring a New Molecular Entity (NME) to the market after spending over $50 billion on its discovery. Some of these NMEs ends up not being endorsed by the US Food and Drug Administration (FDA), and results in wasted efforts and resources. It is therefore essential to take necessary steps in drug discovery by detecting drug-target relationship through the identification and prediction of bioactive molecules.

These bioactive molecular compounds are very important in reducing deteriorative processes and degenerative disease. It is therefore essential to preserve the beneficial characteristics of these compounds since they are important in drug discovery and, also, identify and predict compounds which are highly bioactive to assist in drug-target interaction. The bioactivity of a molecular compound and the property of the activity is known as endogenous and exogenous features respectively [5]. Computational approaches have been developed for the prediction of biologically active compounds over the years but, the performance of each approach used in the prediction is diverse due to variety of methods and datasets. Therefore, there is a need for approaches which denotes the previous existing methods that has been in use and can predict bioactive molecular compounds with high performance. This study therefore aims to get a better predicting accuracy using majority voting based method to assist the drug discovery process.

II. RELATED WORK

A. Machine Learning Methods

It takes a long period of time and effort to map the target disease to the drug which has the capability to handle it [6], [7]. Molecules with high hits are screened from the compounds. A method previously used is High Throughput Screening (HTS), and it follows a process known as trial and error method [8]. Machine learning because of its high computational ability has inherent capability to make predictive analysis [9] and there has been a resurging interest in it which has made data mining for drug discovery popular [10]. The existing data are big data since they possess volume, variety, velocity and veracity characteristics, and, this has made machine learning important to process these data [11], as a platform to create cheaper and powerful computational processing with ease and also provide sufficient storage for the large data [12]. Models are thus built automatically and faster for a more accurate and faster result delivery [13].

Predictive analysis has been made with several machine learning algorithms for chemoinformatics and other areas. Support Vector Machine is an algorithm which is widely

978-1-5386-2150-9/17/$31.00 ©2017 IEEE

used for classification. It was introduced by [14] and it is a supervised learning method used for classification and regression [15]. SVM represents examples as points in space, then separates them by a maximum margin so that each example falls into one of the categories. Its main concept is maximized marginalization and kernel function for non-linearly separable classes. It is also robust to accommodate high dimensional data. A review by [16] shows that Support Vector Machine performs better than most investigated classifiers, and the performance can be improved by adjusting its parameters, even though more work still needs to be done on it. Decision Tree is also a method commonly used for classification. It is constructed from class-labelled training tuples [17] and recursively divided into branches. It is used in chemoinformatics for the identification of substructures that distinguish activity from nonactivity in a given chemical compound library [18], and also for classification of chemical compounds into drug and non-drug [19]. It is an effective classifier which performs better with categorical data.

K-Nearest Neighbour (k-NN) is a non-parametric lazy algorithm, which does not make generalization based on the training data point. Prediction are made based on the nearest training example in the feature space and it can also be used for regression [17]. It is simple but smart. K-NN can be used to model regression between bioactivity and molecular descriptors using manifold ranking [20]. A good predictive performance can be achieved by exploiting the similarity structure of molecules. K-NN calculates the distance between each training set and test set in the dataset and gives the k closest sets. The time complexity is linear and it is guaranteed to find the needed and exact k nearest neighbours [21].

Random Forest (RF) on the other hand is an ensemble of Classification and Regression Trees (CART). It is referred to as either a classifier or an ensemble. A bootstrap sample of the original data is used in growing each tree. Trees in random forest are unpruned unlike decision tree where there is a possibility of applying pre-pruning or post-pruning [22]. Random forest has been effective for classification in chemoinformatics and it has shown some superiority when compared to other classifiers [23, 24] and also great ability to deal with class imbalance problems. Naïve Bayes (NB), on the other hand, is a classifier specifically introduced for retrieval of text [25]. It is based on Bayes theorem and frequently used in chemoinformatics either in combination with another classifier or comparison with other classifiers. It has been generally used in predicting biological properties rather than physicochemical properties. In predicting the toxicity of a compound [26], Naïve Bayes classifier showed better performance compared with other classifiers. The size and diversity of the class of a dataset can have effect on the predictive ability of the model as shown by [27] where the developed method had better performance compared to Naïve Bayes using the same dataset due to the size and diversity of the class.

These classifiers give good performance depending on the situation or dataset used, although no single classifier is said to be better and superior to the other when compared based on the performance, time and, computational cost [17]. Ensemble therefore helps in achieving high accuracy for prediction of bioactive compounds.

B. Majority Voting

The quest for better prediction accuracy transcends the capability of single classifiers. The hazard caused by neglecting prediction of bioactive molecules during drug discovery places greater importance on it. Studies have shown that combining two or more classifiers, also known as wisdom of crowds, or ensemble, can improve overall predicting accuracy of a model. Reference [28] in the classification of a large dataset which contains more than 24,700 compounds whose Cytochrome P450 (CYP) were known with five unique CYP isoforms: 1A2, 2C9, 2C19, 2D6, and 3A4 used the combination of different classifiers which were fused by Back Propagation Artificial Neural Network (BP-ANN) and validated using 5-fold cross-validation reported that the performance derived from the combined classifiers supersedes that of the single classifiers. In the same vein, Extreme Gradient Boosting (Xgboost) [29] which is a variant of the boosting ensemble and an ensemble of Classification and Regression Trees (CART), using seven different datasets and compared against Random Forest, Lib Support Vector Machine (LibSVM), Radial Basis Function Network (RBFN), and Naïve Bayes, had the best accuracy for bioactive molecule prediction.

Majority voting ensemble has been used in other areas but not been for predicting bioactive molecules in chemoinformatics. It is an effective method which handles incomplete data without making assumptions about missing values. It can either be weighted or not. Due to the inability of Extreme Learning Machines (ELM) in handling incomplete data which are collected from real-life applications, voting method was implemented by [30] on the ELM to determine the importance of each data using the training set. Using weighted majority voting in predicting, the recorded performance is better than that of single classifiers, and the computational efficiency of the neural network ensemble was improved. In bioinformatics, multiple voting which consisted of several single voting different from each other because of the random partitioning, fused by majority voting to provide solution to the negligence of data partitioning while classifying was introduced [31]. It was pointed out that partitioning of data during single voting is important to the detection of mislabelled data, and multiple voting was introduced to reduce the problem of dependency of mislabelled data on data partitioning. Since multiple voting is a conglomeration of single voting, multiple voting can be used on single voting to put to check, the unreliability of single voting. Feature construction was introduced to voting based method to make it more effective in addressing the analysis of financial distress in the real world and predicting the probability of a bank being involved in financial distress [32]. Majority voting uses indecisive rules to construct good rules and make decision and has shown great performance in other areas which warrants implementing it in the prediction of bioactive molecules.

III. METHODOLOGY

A. Dataset

The dataset used in the implementation of this research is available in the MDL Drug Data Report (MDDR) database and is already converted to Pipeline Pilot's ECFC_4 fingerprint and folded into 1024 elements fingerprints. It is a commonly used dataset for bioactivity prediction and has been used for Ligand based Virtual Screening (LBVS) [33]-[35], and bioactive molecule prediction [29], [36]. The prediction was made based on the activity of the biological molecules. The dataset consists of 8294 bioactive molecules and 11 classes which contains both structurally heterogeneous (diverse) and homogeneous molecules. It also consists of activity class, diversity of the class, number of molecules attributed to each class, and the average pairwise Tanimoto similarity index which is calculated for all the molecule pair in the class.

B. Majority Voting Ensemble Method

Majority voting is a simple and effective ensemble algorithm which can be used for classification and regression problems. It reduces misclassification for multi-classifiers and has shown considerable performance in improving prediction accuracy. It has been applied in various areas and its result has been impressive. Series of classifiers make predictions based on their algorithms, then majority voting makes prediction based on mode of the base classifiers. It is important to select classifiers with uncorrelated predictions. A good rule of thumb requires the selection of classifiers from tree, Bayesian, function, lazy classifiers, to have classifiers with varied predictions.

If the final computed class probability of an instance is given by $LC_i(X)$, the final prediction of the ensemble of classifiers is given as (1):

$$H(X) = arg_{i=1,\ldots,n} \max(LC_i(X)) \quad (1)$$

Similarly, the computation of majority voting is given as (2):

$$arg_{i=1,\ldots,m} \max\{S_i = \sum_{j=1}^{m} I(h_j(X) = Y)\} \quad (2)$$

Using WEKA, which is a data analytic software, five classifiers which are commonly used for bioactivity prediction were used as the base classifiers and combined differently to find the best combination. These classifiers are Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), k-Nearest Neighbour (k-NN), and Random Forest (RF) which is in fact an ensemble of trees itself. These classifiers make predictions differently on the datasets, using different algorithms, and the prediction is validated with 10-fold cross validation. Majority voting gives the overall prediction based on the mode of the prediction from the base classifiers for each instance. For each instance in the dataset, majority voting assigns the instance to the class where majority of the base classifiers classify it. It should be noted that if more half of the base classifiers classify incorrectly for an instance, the overall prediction for majority voting will also be incorrect.

IV. RESULTS AND DISCUSSION

The classifiers were evaluated based on their performance in prediction accuracy. Accuracy is a widely used evaluation metric in classification. Accuracy of a classifier is also known as the overall recognition rate of the classifier. It refers to the predictive abilities of the classifier. It is the percentage ratio of correctly classified instances. The accuracy of a classifier or built model cannot be determined based on the training data but on the test data which have class-labelled instances that were not part of the training data. The 10-fold cross validation was used in this instance to validate the accuracy of the classifiers. The effectiveness of accuracy is shown when the distribution of the class is relatively balanced. The general formula for deriving the accuracy of a classifier from the confusion matrix is given as:

$$Accuracy = \frac{TP+TN}{P+N} \quad (3)$$

where, TP are the true positives, TN are the true negatives, P are the positives, N are the negatives, FP are the false positives, and FN are the false negatives. The five classifiers were combined differently using all the classifiers first and subsequently with the exemption of one classifier each to generate a total of six different combinations for the majority voting ensembles. Table I shows the accuracy of the majority voting with the different combinations.

TABLE I. ACCURACY OF MAJORITY VOTING WITH DIFFERENT BASE CLASSIFIER COMBINATION

Combination of Base Classifiers	Accuracy (%)
SVM, DT, NB, k-NN, RF	96.9134
DT, NB, k-NN, RF	95.8765
SVM, NB, k-NN, RF	97.0943
SVM, DT, k-NN, RF	**97.1546**
SVM, DT, NB, RF	95.5872
SVM, DT, NB, k-NN	95.6716

From Table I, it is shown that the combination of Support Vector Machine, Decision Tree, k-Nearest Neighbour, and Random Forest as the base classifiers gave the best accuracy for the majority voting as highlighted. It should be noted that the absence of Naïve Bayes which had the lowest accuracy as an individual classifier as shown in Table II made it have a better result. Since majority voting makes prediction based on the prediction of the base classifiers, the presence of a non-suitable individual classifier will affect its accuracy. This is why the combinations which had both Naïve Bayes and Decision Tree in it, which were the two lowest performing classifiers in Table II, had low performance in their accuracies. Therefore, when building a model with majority voting method, it is essential to choose classifiers with good predictive accuracy as this influences the overall accuracy of majority voting.

TABLE II. COMPARISON BETWEEN MAJORITY VOTING AND BASE CLASSIFIERS

Classifiers	Accuracy (%)
Majority Voting (SVM, k-NN, DT, RF)	**97.1546**
Naïve Bayes	77.6585
Support Vector Machine	96.0694
k-Nearest Neighbour	96.7929
Decision Tree	87.714
Random Forest	96.9375

As shown in Table II, it is seen that among the base classifiers, Random Forest had the highest predicting accuracy of 96.9375% and it is closely followed by k-Nearest Neighbour (96.7929%), then Support Vector Machine (96.0694%), Decision Tree (87.714%) and lastly Naïve Bayes with a low accuracy of 77.6585%. Random Forest having the best accuracy can be pinned on the fact that Random Forest itself is an ensemble of trees which will hence improve its performance, although, k-Nearest Neighbour closely follows it.

The final comparison between the individual classifiers and majority voting can then be highlighted in this decreasing order: Majority Voting (with the combination of Support Vector Machine, Decision Tree, k-Nearest Neighbour, and Random Forest) (**97.1546%**) > Random Forest (96.9375%) > k-Nearest Neighbour (96.7929%) > Support Vector Machine (96.0694%) > Decision Tree (87.714%), and finally Naïve Bayes (77.6585%). This shows majority voting ensemble method having the best accuracy with the right and appropriate combination of base classifiers.

V. CONCLUSION

Bioactive molecule prediction is an integral step in drug discovery. Machine learning algorithms are introduced to handle the complex chemical data and several algorithms have been used, although the need to have better predicting methods exist. This study implements majority voting which is frequently used in other areas but rarely for prediction in chemoinformatics. The right choice of base classifiers is essential and from the study, the best choice of base classifier combination had better accuracy than the single classifiers examined. The method is suitable for high dimensional dataset and effectively applicable in bioactivity prediction.

ACKNOWLEDGMENT

This work is supported by the Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under the Research University Grant Category (VOT Q. J130000.2528.16H74).

REFERENCES

[1] F. K. Brown, "Chemoinformatics: what is it and how does it impact drug discovery," *Annual reports in medicinal chemistry,* vol. 33, pp. 375-384, 1998.

[2] B. Gaüzére, L. Brun, D. Villemin, and M. Brun, "Graph kernels based on relevant patterns and cycle information for chemoinformatics," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, 2012, pp. 1775-1778.

[3] J. Bajorath, "Chemoinformatics: Recent advances at the interfaces between computer and chemical information sciences, chemistry, and drug discovery," *Bioorganic & Medicinal Chemistry,* vol. 20, p. 5316, 2012.

[4] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg*, et al.*, "How to improve R&D productivity: the pharmaceutical industry's grand challenge," *Nature reviews Drug discovery,* vol. 9, pp. 203-214, 2010.

[5] A. Iwaniak, P. Minkiewicz, M. Darewicz, M. Protasiewicz, and D. Mogut, "Chemometrics and cheminformatics in the analysis of biologically active peptides from food sources," *Journal of Functional Foods,* vol. 16, pp. 334-351, 2015.

[6] D. Harnie, M. Saey, A. E. Vapirev, J. K. Wegner, A. Gedich, M. Steijaert*, et al.*, "Scaling machine learning for target prediction in drug discovery using apache spark," *Future Generation Computer Systems,* vol. 67, pp. 409-417, 2017.

[7] M. A. Khamis, W. Gomaa, and W. F. Ahmed, "Machine learning in computational docking," *Artificial intelligence in medicine,* vol. 63, pp. 135-152, 2015.

[8] A. S. Reddy, S. P. Pati, P. P. Kumar, H. Pradeep, and G. N. Sastry, "Virtual screening in drug discovery-a computational perspective," *Current Protein and Peptide Science,* vol. 8, pp. 329-351, 2007.

[9] S. Zhang, "Application of machine learning in drug discovery and development," *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques: Complex Computational Methods and Collaborative Techniques,* p. 235, 2010.

[10] M. Glick and E. Jacoby, "The role of computational methods in the identification of bioactive compounds," *Current opinion in chemical biology,* vol. 15, pp. 540-546, 2011.

[11] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine Learning on Big Data: Opportunities and Challenges," *Neurocomputing,* 2017.

[12] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, "Efficient machine learning for big data: A review," *Big Data Research,* vol. 2, pp. 87-93, 2015.

[13] D. Hecht, "Applications of machine learning and computational intelligence to drug discovery and development," *Drug Development Research,* vol. 72, pp. 53-65, 2011.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, pp. 273-297, 1995.

[15] S. Zhou, G.-B. Li, L.-Y. Huang, H.-Z. Xie, Y.-L. Zhao, Y.-Z. Chen*, et al.*, "A prediction model of drug-induced ototoxicity developed by an optimal support vector machine (SVM) method," *Computers in biology and medicine,* vol. 51, pp. 122-127, 2014.

[16] S. Sengupta and S. Bandyopadhyay, "Application of Support Vector Machines in Virtual Screening," *International Journal for Computational Biology (IJCB),* vol. 1, pp. 56-62, 2014.

[17] A. Lavecchia, "Machine-learning approaches in drug discovery: methods and applications," *Drug discovery today,* vol. 20, pp. 318-331, 2015.

[18] J. Klekota and F. P. Roth, "Chemical substructures that enrich for biological activity," *Bioinformatics,* vol. 24, pp. 2518-2525, 2008.

[19] N. Schneider, C. Jäckels, C. Andres, and M. C. Hutter, "Gradual in silico filtering for druglike substances," *Journal of chemical* information and modeling, vol. 48, pp. 613-628, 2008.

[20] L. Shen, D. Cao, Q. Xu, X. Huang, N. Xiao, and Y. Liang, "A novel local manifold-ranking based K-NN for modeling the regression between bioactivity and molecular descriptors," *Chemometrics and Intelligent Laboratory Systems,* vol. 151, pp. 71-77, 2016.

[21] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, "Efficient kNN classification algorithm for big data," *Neurocomputing,* vol. 195, pp. 143-148, 2016.

[22] J.-H. Huang, H.-L. Xie, J. Yan, H.-M. Lu, Q.-S. Xu, and Y.-Z. Liang, "Using random forest to classify T-cell epitopes based on amino acid properties and molecular features," *Analytica chimica acta,* vol. 804, pp. 70-75, 2013.

[23] J. Hu, Y. Li, J.-Y. Yang, H.-B. Shen, and D.-J. Yu, "GPCR–drug interactions prediction using random forest with drug-association-matrix-based post-processing procedure," *Computational biology and chemistry,* vol. 60, pp. 59-71, 2016.

[24] Z.-S. Wei, K. Han, J.-Y. Yang, H.-B. Shen, and D.-J. Yu, "Protein–protein interaction sites prediction by ensembling SVM and sample-weighted random forests," *Neurocomputing,* vol. 193, pp. 201-212, 2016.

[25] L. Zhang, L. Jiang, C. Li, and G. Kong, "Two feature weighting approaches for naive Bayes text classifiers," *Knowledge-Based Systems,* vol. 100, pp. 137-144, 2016.

[26] H. Zhang, Y.-L. Kang, Y.-Y. Zhu, K.-X. Zhao, J.-Y. Liang, L. Ding, *et al.*, "Novel naïve Bayes classification models for predicting the chemical Ames mutagenicity," *Toxicology in Vitro,* vol. 41, pp. 56-63, 2017.

[27] A. Koutsoukas, R. Lowe, Y. KalantarMotamedi, H. Y. Mussa, W. Klaffke, J. B. Mitchell, *et al.*, "In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt window," *Journal of chemical information and modeling,* vol. 53, pp. 1957-1966, 2013.

[28] F. Cheng, Y. Yu, J. Shen, L. Yang, W. Li, G. Liu, *et al.*, "Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers," *Journal of chemical information and modeling,* vol. 51, pp. 996-1011, 2011.

[29] I. Babajide Mustapha and F. Saeed, "Bioactive molecule prediction using extreme gradient boosting," *Molecules,* vol. 21, p. 983, 2016.

[30] Y.-T. Yan, Y.-P. Zhang, J. Chen, and Y.-W. Zhang, "Incomplete data classification with voting based extreme learning machine," *Neurocomputing,* vol. 193, pp. 167-175, 2016.

[31] D. Guan, W. Yuan, T. Ma, and S. Lee, "Detecting potential labeling errors for bioinformatics by multiple voting," *Knowledge-Based Systems,* vol. 66, pp. 28-35, 2014.

[32] H. A. Güvenir and M. Çakır, "Voting features based classifier with feature construction and its application to predicting financial distress," *Expert Systems with Applications,* vol. 37, pp. 1713-1718, 2010.

[33] A. Abdo, F. Saeed, H. Hamza, A. Ahmed, and N. Salim, "Ligand expansion in ligand-based virtual screening using relevance feedback," *Journal of computer-aided molecular design,* vol. 26, pp. 279-287, 2012.

[34] M. M. Al-Dabbagh, N. Salim, M. Himmat, A. Ahmed, and F. Saeed, "A quantum-based similarity method in virtual screening," *Molecules,* vol. 20, pp. 18107-18127, 2015.

[35] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, *et al.*, "New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching," *Journal of chemical information and modeling,* vol. 46, pp. 462-470, 2006.

[36] A. Abdo, V. Leclère, P. Jacques, N. Salim, and M. Pupin, "Prediction of new bioactive molecules using a bayesian belief network," *Journal of chemical information and modeling,* vol. 54, pp. 30-36, 2014.