Journal of Intelligent & Fuzzy Systems 35 (2018) 383–392
DOI:10.3233/JIFS-169596
IOS Press

Bioactive molecule prediction using majority voting-based ensemble method

Olutomilayo Olayemi Petinrin[a] and Faisal Saeed[b,c,*]
[a]*Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Johor, Malaysia*
[b]*College of Computer Science and Engineering, Taibah University, Medina, Saudi Arabia*
[c]*Department of Information Systems, Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Johor, Malaysia*

Abstract. The current rise in the amount of data generated has necessitated the use of machine learning in the drug discovery process to increase productivity. It is therefore important to predict molecular compounds which are biologically active and capable of drug-target interaction. Various machine learning methods have been used in predicting bioactive molecular compounds in order to deal with the large volume of data being generated. This study investigates the Majority Voting ensemble method using different combinations of 5 commonly-used machine learning algorithms, including Support Vector Machine, Decision Tree, Naïve Bayes, k-Nearest Neighbor, and Random Forest on three chemical datasets DS1, DS2, and DS3 which consist of structurally heterogeneous and homogeneous molecules and are commonly used in other studies. The results show that Majority Voting has a better performance, based on all the evaluation metrics used, compared to each of the machine learning algorithms as individual classifiers. It also shows the Majority Voting ensemble method as effective in the prediction of both heterogeneous and homogeneous bioactive molecular compounds, using statistical evaluation.

Keywords: Bioactivity prediction, chemoinformatics, drug discovery, ensemble classification, majority voting

1. Introduction

Drug discovery is an aspect of chemoinformatics whose importance cannot be over-emphasized. Chemoinformatics has become recognized as a distinct field over the years [8]. It is also known ascheminformatics, chembioinformatics, or chemical informatics [3]. It encompasses aspects like the Quantitative Structure Activity Relationship (QSAR), Quantitative Structure Property Relationship (QSPR), lead optimization, and drug target discovery, and in recent times it has been used with the aim of predicting the properties of biological molecules from their structural similarity [18]. In fact, it has been suggested that analysing and predicting bioactive molecules is the most popular task of chemoinformatics [7]. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization" [9].

Various known and unknown human diseases keep springing up frequently and to combat these new, more effort is put into the discovery of drugs efficient enough to tackle the disease. Pharmaceutical companies spend a great deal of time and resources in producing a new drug and after production, but 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. According to [37], it cost 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 end up not being endorsed by the US Food and Drug Administration (FDA),

*Corresponding author. Faisal Saeed. E-mail: fsaeed@taibahu.edu.sa.

and results in wasted effort, time, and resources. It is therefore essential to take the necessary steps in drug discovery. One of these necessary steps is the identification and prediction of biologically active molecular compounds from diverse chemical compounds to detect a drug-target relationship.

Bioactive compounds are compounds which have an effect on living organism, tissues, or cells. These compounds can be found in both plant and animal products or can be produced synthetically. They are very important in reducing deteriorative processes and degenerative disease. It is therefore essential to preserve the beneficial characteristics of these compounds. Bioactive molecules are also important in drug discovery. It is therefore important to identify and predict the molecular compounds which are highly bioactive and can thus assist in drug-target interaction. The bioactivity of a molecular compound and the property of the activity are known as endogenous and exogenous features respectively [28]. Although various forms of research have been carried out and computational approaches have been developed for the prediction of biologically active compounds, the performance of each approach used in the prediction varies due to the variety of methods and datasets used; thus, there is aquest for new approaches to replace the existing methods in use, which can predict bioactive molecular compounds with high performance. Hence, this work aims to obtain a better performance using voting based methods in the bid to avoid the frustration that comes with wasted effort and resources in drug discovery.

2. Methods

2.1. Machine learning methods

The process of drug discovery takes a long time and effort, from the mapping of the target disease to the development of a drug which has the capability to handle it [22, 31]. Compounds are being screened in a search for molecules with a high level of hits. High Throughput Screening (HTS) is a method which has been used over the years by trying different process to reach a solution. This process is known as the trial and error method [38]. A remarkable development from this trial and error method is the process known as Virtual Screening (VS). Virtual screening is an improvement from traditional approaches such as High Throughput Screening to the use of machine learning methods due to the increasing size of data being generated by the chemical field and need proper computation [6].

It has been proven that machine learning, because of its high computational ability, has the inherent capability to make predictive analyses [48]. There has been a resurgent interest in machine learning which has increased the popularity of data mining for drug discovery [19]. The existing data are mainly considered as big data, since they possess the characteristics of volume, variety, velocity, and veracity, and this has made machine learning an important way to process these data [49]. The machine learning platform has been able to create cheaper and more powerful computational processing with ease, while giving room for affordable storage capacity for the voluminous data [5]. This makes it is possible and easy to produce models automatically and within a target time, which will be able to analyse huge, and, more complex data while delivering the results faster and more accurately, even on a very large scale [23].

Various machine learning algorithms have been used in predictive analysis both in chemoinformatics and outside the field. One of these algorithms is Support Vector Machine which is widely used in classification. It was introduced by [14] and is a supervised learning method used for classification and regression [50]. 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 [40] showed that Support Vector Machine performs better than most of the investigated classifiers, and the performance can be improved by adjusting its parameters, even though more work still needs to be done on it. Radial Basis Function Network (RBFN) is commonly used as kernel in SVM [33, 40]. SVM can be used as a filter between drug and non-drug compounds in the early stage of drug discovery [33], even though real life data for such scenarios might be unbalanced. Support Vector Machine can be optimized to handle parameter optimization and feature selection [50]. In an area similar to chemoinformatics, SVM was also used in screening of drugs for hepatocellular carcinoma [45], where the chemicals to be predicted were accurately identified with SVM.

Another classifier frequently used is k-Nearest Neighbour (k-NN). k-NN is a non-parametric lazy algorithm, which does not make generalizations based on the training data points. It has little or no training phase, which influences the decision to consider it as a lazy algorithm, but this makes the training phase fast. k-NN makes predictions based on the

nearest training example in the feature space and it can also be used for regression [35]. It is simple but smart. k-NN can be used to model regression between bioactivity and molecular descriptors using manifold ranking [41]. Exploitation of similarity in structure helps in achieving a good predictive performance. It can be used in combination with docking-based intermolecular analysis to discover new inhibitors [29]. 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 [16]. Other areas apart from chemoinformatics where k-NN has been utilized include classification of heart disease [15] and detection of Parkinson's disease using fuzzy k-NN [11].

Naïve Bayes (NB), based on Bayes' theorem, is a classifier which was specifically introduced for text retrieval [47] and is also another machine learning algorithm which has gained widespread use in chemoinformatics. It is a highly scalable classifier and it requires a number of measures which are linear to the number of attributes (features/variables) in a learning problem. Naïve Bayesian classifiers are frequently used in chemoinformatics either in combination with another classifier or compared with other classifiers. It has been generally used in predicting biological properties rather than physicochemical properties. In predicting the toxicity of a compound [46], Naïve Bayes classifier performed better than other classifiers which it was examined against. Naïve Bayes has also been used for predicting phospholipidosis mechanism [36]. The size and diversity of the class of a dataset can have an effect on the predictive ability of the model as shown by [34] where the developed method had better performance compared to Naïve Bayes using the same dataset due to the size and diversity of the class. Naïve Bayesian classifiers can also be used for regression, even though this case is rarely seen nor implemented in chemoinformatics [17].

The Decision Tree is commonly used for classification. Output values of targets are predicted using the various input attributes of each instance. As the name implies, it is a tree depicted as an inverted tree with the roots at the top while the leaves are below. The root is the most essential attribute and it divides further into branches which are also further divided into branches until it reaches the leaf. The leaf is a node and cannot be further divided, and the nodes from the branches are known as internal nodes. Each leaf node is assigned with a target property, and the internal nodes are assigned with a molecular descriptor which checks if an instance is satisfying a condition before branching it out, based on its characteristics. The decision tree is constructed from class-labelled training tuples [35]. It is basically of two types: classification, which predicts the class an item of data belongs to, and regression, where a real number can be predicted. This is mainly referred to as a Classification and Regression Tree (CART). A decision tree can be pruned, either pre-pruned or post-pruned, to avoid overfitting. It is used in chemoinformatics for the identification of substructures that distinguish activity from nonactivity in a given chemical compound library [32], and also for classification of chemical compounds into drugs and non-drugs [39]. Apart from chemoinformatics, it has also been used in cancer classification [12] and also fault diagnosis [30]. The Decision Tree has proven to be an effective classifier in all areas of application although it works best with categorical data.

Random Forest is a classifier which consist of ensembles of Classification and Regression Trees (CART). It is an ensemble of multiple decision trees where a bootstrap sample of the original data is utilized for growing each tree. However, the trees in random forest are not pruned, unlike decision tree, where there is a possibility of applying pre-pruning or post-pruning [27]. Two techniques, bagging and random feature selection are used by Random Forest, where majority vote is also implemented to make predictions. It uses weak learners to make predictions [42]. Weak learners are predictors with low bias and high variance, which are essential for good accuracy. This can be achieved by growing a tree to its maximum depth without pruning. According to [26], Random Forest is constructed using the following process:

i. Draw n_{tree} bootstrap samples from the original data. n_{tree} here represents the number of ensemble trees;
ii. Grow an unpruned CART for all the bootstrap samples, and randomly select m_{try} variables at each node of the tree for splitting. m_{try} here can be a positive integer;
iii. Make prediction using aggregation of information from the n_{tree} using majority vote;
iv. Determine the error rate using data outside the bootstrap sample.

Random Forest has been effectively used for classification and prediction in chemoinformatics and it has shown some superiority when compared to other classifiers [25, 43] and also great ability to deal with

class imbalance problems. It is a fast classifier and rarely predicts wrongly.

Each machine learning algorithm has shown good performance in various situations where it has been used, depending on the type of dataset or its dimensionality. Despite this, there is no approved standard for predicting bioactive molecular compounds, since no method can be said to be superior to the other. If each method is compared based on the performance, time and, computational cost, it can be easily concluded that no method can claim to be superior to the other [35]. It is however of great importance that high performance be obtained from these methods when predicting.

2.2. Voting based methods

Although single classifiers have been able to perform well at predicting, there is still the quest to improve the performance of models which are used for prediction, especially in the case of predicting bioactive molecules, due to their importance in drug discovery and the result of their neglect. Analyses by various researchers have shown that the combination of two or more classifiers might result in better performance than that of single classifiers. In the classification of a large dataset which contains more than 24,700 compounds whose Cytochrome P450 (CYP450) are known and five unique CYP isoforms: 1A2, 2C9, 2C19, 2D6, and 3A4, [13] used the combination of different classifiers which were combined by Back Propagation Artificial Neural Network (BP-ANN) and validation using 5-fold cross-validation and reported that the performance derived from the combined classifiers superseded that of the single classifiers. Bagging, Boosting, Stacking, and Voting are popular ensemble methods which can be used to improve prediction performance of a model. Extreme Gradient Boosting (Xgboost) [6], which is a variant of the boosting ensemble and an ensemble of Classification and Regression Trees (CART), was used in an experimental prediction of bioactive molecules, using seven different datasets; when compared against Random Forest, Lib Support Vector Machine (LibSVM), Radial Basis Function Network (RBFN), and Naïve Bayes, it had the best prediction accuracy overall.

Various explorations of ensemble methods have been made in chemoinformatics, but the voting-based ensemble which has been used in other areas has not yet been tapped into in this field. The voting-based method is an ensemble which consists of base classifiers and works to minimize misclassification. It improves the overall prediction based on the prediction of the base classifiers. It uses a combination rule on the prediction of the base classifiers. These rules are the product of probability, average of probability, majority voting, maximum probability and minimum probability, and they can either be weighted or not [10]. The voting-based method is effective in handling incomplete data without making assumptions about missing values. The voting method was used by [44] in Extreme Learning Machine (ELM) to handle incomplete data. ELMs are efficient learning algorithms for single-hidden layer feedforward neural network (SLFN). It was discovered that ELM cannot handle incomplete data, which are mostly common in data gathered from real life applications. The voting-based extreme learning machine uses the training set data to determine how important each item of data is, and trains using the ELM. Using weighted majority voting in predicting, the recorded performance is better than that of single classifiers. It also improves the computational efficiency of the neural network ensemble.

Moreover, in implementing voting ensemble methods in bioinformatics, [20] pointed out that partitioning of data during single voting is important to the detection of mislabelled data, and therefore introduced multiple voting, which consists of several single votes different from each other because of the random partitioning combinedwith majority voting to provide a solution to the neglect of data partitioning while classifying. Multiple voting is able to reduce the problem of dependency of mislabelled data on data partitioning. Since multiple voting is a conglomeration of single votes, multiple voting can be used on single voting to check the unreliability of single voting. The introduction of feature construction to the voting-based method made 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 [21]. The method uses indecisive rules to construct good rules and make decisions. The performance of the algorithm was better than that of the other algorithms compared with it.

3. Experimental design

3.1. Datasets

The three datasets DS1, DS2 and DS3 used in the implementation of this research are found in the

Table 1
Activity Class for Dataset DS1

Activity Index	Activity Class	Active Molecules	Pairwise Similarity (Mean)
31420	Renin inhibitors	1130	0.573
71523	HIV protease inhibitors	750	0.446
37110	Thrombin inhibitors	803	0.419
31432	Angiotensin II AT1 antagonists	943	0.403
42731	Substance P antagonists	1246	0.339
06233	5HT3 antagonists	752	0.351
06245	5HT reuptake inhibitors	359	0.345
07701	D2 antagonists	395	0.345
06235	5HT1A agonists	827	0.343
78374	Protein kinase C inhibitors	453	0.323
78331	Cyclooxygenase inhibitors	636	0.268

Table 2
Activity Class for Dataset DS2

Activity Index	Activity Class	Active Molecules	Pairwise Similarity (Mean)
07707	Adenosine (A1) agonists	207	0.424
07708	Adenosine (A2) agonists	156	0.484
31420	Renin inhibitors	1130	0.584
42710	Monocyclic β-lactams	111	0.596
64100	Cephalosporins	1301	0.512
64200	Carbacephems	158	0.503
64220	Carbapenems	1051	0.414
64300	Penicillin	126	0.444
65000	Antibiotic, macrolide	388	0.673
75755	Vitamin D analogous	455	0.569

Table 3
Activity Class for Dataset DS3

Activity Index	Activity Class	Active Molecules	Pairwise Similarity (Mean)
09249	Muscarinic (M1) agonists	900	0.257
12455	NMDA receptor antagonists	1400	0.311
12464	Nitric oxide synthase inhibitors	505	0.237
31281	Dopamine β-hydroxylase inhibitors	106	0.324
43210	Aldose reductase inhibitors	957	0.37
71522	Reverse transcriptase inhibitors	700	0.311
75721	Aromatase inhibitors	636	0.318
78331	Cyclooxygenase inhibitors	636	0.382
78348	Phospholipase A2 inhibitors	617	0.291
78351	Lipoxygenase inhibitors	2111	0.365

MDL Drug Data Report (MDDR) database and are already converted to Pipeline Pilot's ECFC_4 fingerprint and folded into 1024 elements fingerprints. They are commonly used datasets and have been used for Ligand-based Virtual Screening (LBVS) [2, 4, 24], and bioactive molecule prediction [1, 6]. The prediction was made based on the activity of the biological molecules. DS1 contains 8294 bioactive molecules and 11 classes, which comprise both structurally heterogeneous and homogeneous molecules. DS2 and DS3 contain 5083 and 8569 bioactive molecules respectively, with 10 classes of homogeneous molecules for DS2 and 10 classes of heterogeneous molecules for DS3. Tables 1, 2, and 3 show the activity class of the molecules, diversity of the class, number of molecules attributed to each class, and the average pairwise Tanimoto similarity index of the molecule pair of the class for datasets DS1, DS2, and DS3, respectively.

3.2. Voting-based mechanism

Five separate classifiers which are commonly used for bioactivity prediction were used as the base

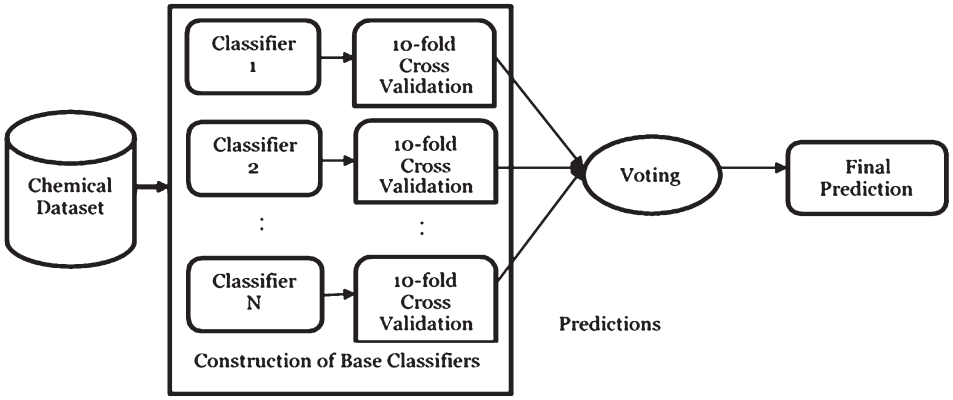

Fig. 1. Voting Mechanism.

classifiers and combined differently to find the best combination. These classifiers first make predictions differently on the datasets and the predictions are validated with 10-fold cross validation. Majority voting gives the overall prediction based on the prediction of the base classifiers. For each instance in the dataset, majority voting assigns the instance to the class where the majority of the base classifiers classify it. This implementation was carried out using WEKA software. 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. It is important to select classifiers which have uncorrelated predictions. A good rule of thumb requires the selection of classifiers from tree, Bayesian, function, and lazy classifiers, in order to have classifiers with varied predictions. The process of majority voting is shown in Fig. 1.

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:

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

4. Results and discussion

The classifiers were evaluated based on six different evaluation metrics. These metrics are generally used in evaluating the performance of a classifier. They are accuracy, sensitivity, specificity, precision, recall, and f-measure. 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. Sensitivity can be referred to as true positive recognition rate. It is the percentage ratio of positive instances which are correctly classified as positive or the measure of positives correctly identified as such. It depicts how much the classifier avoids false negatives. Specificity is known as true negative recognition rate. It is the percentage ratio of negative instances which are correctly identified as negative or the total measure of negatives correctly classified as such. It depicts how much the classifier avoids false positives. Precision is a measure of exactness. It shows the percentage ratio of instances which are classified as positive and are actually positive. It is based on relevance. That is, how relevant are the instances classified as being positive? Recall is a measure of exactness. It shows how many of the actual positives are predicted to be such. Recall and Precision are both combined into a single metric known as the F-measure. The F-measure is the harmonic mean of both recall and precision and the approximate average of both precision and recall. The F-measure is also known as F_1, since recall and precision are evenly weighted. It can be a bias evaluation metric.

The performance 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 performance of the classifiers. The effectiveness of an evaluation metric is shown when the distribution of the class is relatively balanced. The general formula of the evaluation metrics of a classifier from the confusion matrix is given as:

$$Accuracy = \frac{TP + TN}{P + N} \tag{2}$$

$$Sensitivity = Recall = \frac{TP}{TP + FN} = \frac{TP}{P} \quad (3)$$

$$Specificity = \frac{TN}{N} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Fmeasure = \frac{2 \times recall \times precision}{recall + precision} \quad (6)$$

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 base classifiers were further combined in six different ways. The accuracy evaluation metric was chosen to determine the best combination. The combination with the best accuracy was hence chosen to compare with the base classifiers. The accuracy of the combinations with majority voting for datasets DS1, DS2 and DS3 is shown in Tables 4, 5, and 6 respectively.

Table 4
Accuracy of Majority Voting with Different Base Classifier Combination for Dataset DS1

Majority Voting Combination	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

Table 5
Accuracy of Majority Voting with Different Base Classifier Combination for Dataset DS2

Majority Voting Combination	Accuracy (%)
SVM, DT, NB, k-NN, RF	**98.3081**
DT, NB, k-NN, RF	98.072
SVM, NB, k-NN, RF	98.1114
SVM, DT, k-NN, RF	98.19
SVM, DT, NB, RF	98.131
SVM, DT, NB, k-NN	98.0917

Table 6
Accuracy of Majority Voting with Different Base Classifier Combination for Dataset DS3

Majority Voting Combination	Accuracy (%)
SVM, DT, NB, k-NN, RF	95.5065
DT, NB, k-NN, RF	94.4795
SVM, NB, k-NN, RF	95.6699
SVM, DT, k-NN, RF	**95.6816**
SVM, DT, NB, RF	93.8725
SVM, DT, NB, k-NN	94.1527

Tables 4 and 6 show that the combination of Support Vector Machine, Decision Tree, k-Nearest Neighbour, and Random Forest as the base classifiers gave the best accuracy of 97.1546% and 95.6816% for datasets DS1 and DS3 respectively, for the majority voting. In Table 5, which shows the results of dataset DS2, the combination of all the base classifiers, Support Vector Machine, Decision Tree, Naïve Bayes, k-Nearest Neighbour, and Random Forest gives the best accuracy of 98.3081% compared to all other combinations. Since majority voting makes the overall prediction based on the prediction of the base classifiers, the presence of a non-suitable individual result will affect its accuracy. Therefore, when building a model with the majority voting method, it is essential to choose base classifiers whose accuracy as individual classifiers is good, so that majority voting makes the overall accuracy better. The comparison between the selected individual/base classifiers and the majority voting combination with best accuracy is shown in Tables 7, 8, and 9, for datasets DS1, DS2, and DS3 respectively.

Majority Voting had the best performance in accuracy, sensitivity, precision, recall, and f-measure in datasets DS1 and DS2, as shown in Tables 7 and 8, but in dataset DS3, as shown in Table 9, k-Nearest Neighbor had better specificity compared to Majority Voting. It is also noted that Naïve Bayes had low performance generally in all the datasets and combinations of base classifiers which included Naïve Bayes also had low performance. This shows the importance of selecting good base classifiers to aid better prediction with Majority Voting.

A statistical method, Kendall's Coefficient of Concordance, also known as Kendall's W test, was used to rank the classifiers using the evaluation metrics as the raters and the results for the three datasets are shown in Table 10. Kendall's W gives the measurement of agreement, which shows how well the raters agreed on the ranking of the objects of consideration. Kendall's W ranges from 0 to 1. When W is 0, it means there was no agreement between the raters, while 1 means there was perfect agreement between the raters.

The test ranked Majority Voting highest in all the datasets examined. Moreover, the W, which signifies the level of agreement between the raters, is between 0.953 and 0.984 for the three datasets. Thus, apart from the evaluation metrics used, statistically, Majority Voting also performed better than the other classifiers with a high level of W, which signifies an almost perfect agreement between the raters.

Table 7
Performance Comparison Between Individual Classifiers and Majority Voting for DS1

Classifiers	Accuracy (%)	Sensitivity	Specificity	Precision	Recall	F-Measure
Majority Voting (SVM, DT, k-NN, RF)	**97.1546**	**0.972**	**0.997**	**0.971**	**0.972**	**0.971**
Random Forest	96.9375	0.969	**0.997**	0.969	0.969	0.969
K-Nearest Neighbor	96.7929	0.968	**0.997**	0.968	0.968	0.968
Support Vector Machine	96.0694	0.961	0.996	0.961	0.961	0.961
Decision Tree	87.714	0.877	0.987	0.877	0.877	0.877
Naïve Bayes	77.6585	0.777	0.978	0.782	0.777	0.777

Table 8
Performance Comparison Between Individual Classifiers and Majority Voting for DS2

Classifiers	Accuracy (%)	Sensitivity	Specificity	Precision	Recall	F-Measure
Majority Voting (SVM, DT, k-NN, RF)	**98.3081**	**0.983**	**0.996**	**0.983**	**0.983**	**0.983**
Random Forest	98.1704	0.982	**0.996**	0.981	0.982	0.981
K-Nearest Neighbor	97.8163	0.978	**0.996**	0.978	0.978	0.978
Support Vector Machine	97.8359	0.978	**0.996**	0.978	0.978	0.978
Decision Tree	97.1867	0.972	0.994	0.971	0.972	0.971
Naïve Bayes	94.7669	0.948	0.994	0.954	0.948	0.949

Table 9
Performance Comparison Between Individual Classifiers and Majority Voting for DS3

Classifiers	Accuracy (%)	Sensitivity	Specificity	Precision	Recall	F-Measure
Majority Voting (SVM, DT, k-NN, RF)	**95.6816**	**0.957**	0.992	**0.957**	**0.957**	**0.957**
Random Forest	95.2264	0.952	0.99	0.953	0.952	0.952
K-Nearest Neighbor	95.5299	0.955	**0.993**	0.955	0.955	0.955
Support Vector Machine	93.4524	0.935	0.989	0.934	0.935	0.934
Decision Tree	87.2666	0.873	0.98	0.872	0.873	0.872
Naïve Bayes	65.9197	0.659	0.958	0.697	0.659	0.668

Table 10
Ranking of Methods for datasets DS1, DS2, and DS3 with Kendall's W Test

Dataset	Method	Mean Average	Kendall's W
DS1	Majority Voting	5.83	0.984
	Random Forest	5.00	
	k-Nearest Neighbor	4.17	
	Support Vector Machine	3.00	
	Decision Tree	2.00	
	Naïve Bayes	1.00	
DS2	Majority Voting	5.75	0.953
	Random Forest	4.92	
	Support Vector Machine	3.75	
	k-Nearest Neighbor	3.58	
	Decision Tree	1.92	
	Naïve Bayes	1.08	
DS3	Majority Voting	5.83	0.984
	k-Nearest Neighbor	5.17	
	Random Forest	4.00	
	Support Vector Machine	3.00	
	Decision Tree	2.00	
	Naïve Bayes	1.00	

5. Conclusion

Drug discovery is an important aspect of chemoinformatics and bioactive molecule prediction is an integral step that needs to be undertaken to avoid resource wastage. Machine learning algorithms are introduced to handle the complex chemical data and several algorithms have been used, even though none can claim superiority over the other. Furthermore, the combinations of more than one classifier or an ensemble, have shown better accuracy than single classifiers. This research has utilized majority voting, which is frequently used in other areas but hardly ever for prediction in chemoinformatics. The results show that majority voting has better accuracy compared to single classifiers, provided the right base classifiers are chosen. The method is also suitable in handling high-dimensional datasets which are both homogeneous and heterogeneous. It is therefore recommended as a suitable method in drug discovery for bioactivity prediction, to avoid wastage of resources and accommodate both homogeneous and heterogeneous chemical datasets, however diverse these might be.

Acknowledgments

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] 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* **54**(1) (2014), 30–36.

[2] 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* **26**(3) (2012), 279–287.

[3] M.W. Aktar and S. Murmu Chemoinformatics: Principles and Applications, 1-28, *Agricultural Chemistry*, 2008.

[4] M.M. Al-Dabbagh, N. Salim, M. Himmat, A. Ahmed and F. Saeed, A quantum-based similarity method in virtual screening, *Molecules* **20**(10) (2015), 18107–18127.

[5] 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* **2**(3) (2015), 87–93.

[6] I. Babajide Mustapha and F. Saeed, Bioactive molecule prediction using extreme gradient boosting, *Molecules* **21**(8) (2016), 983.

[7] J. Bajorath, Chemoinformatics: Recent advances at the interfaces between computer and chemical information sciences, chemistry, and drug discovery, *Bioorganic & Medicinal Chemistry* **20**(18) (2012), 5316.

[8] B.F. Begam and J.S. Kumar, A study on cheminformatics and its applications on modern drug discovery, *Procedia Engineering* **38** (2012), 1264–1275.

[9] F.K. Brown, Chemoinformatics: What is it and how does it impact drug discovery, *Annual Reports in Medicinal Chemistry* **33** (1998), 375–384.

[10] J. Cao, Z. Lin, G.-B. Huang and N. Liu, Voting based extreme learning machine, *Information Sciences* **185**(1) (2012), 66–77.

[11] H.-L. Chen, C.-C. Huang, X.-G. Yu, X. Xu, X. Sun, G. Wang and S.-J. Wang, An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach, *Expert Systems with Applications* **40**(1) (2013), 263–271.

[12] K.-H. Chen, K.-J. Wang, K.-M. Wang and M.-A. Angelia, Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data, *Applied Soft Computing* **24** (2014), 773–780.

[13] F. Cheng, Y. Yu, J. Shen, L. Yang, W. Li, G. Liu and Y. Tang, Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers, *Journal of Chemical Information and Modeling* **51**(5) (2011), 996–1011.

[14] C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning* **20**(3) (1995), 273–297.

[15] B. Deekshatulu and P. Chandra, Classification of heart disease using k-nearest neighbor and genetic algorithm, *Procedia Technology* **10** (2013), 85–94.

[16] Z. Deng, X. Zhu, D. Cheng, M. Zong and S. Zhang, Efficient kNN classification algorithm for big data, *Neurocomputing* **195** (2016), 143–148.

[17] E. Frank, L. Trigg, G. Holmes and I.H. Witten, Technical note: Naive Bayes for regression, *Machine Learning* **41**(1) (2000), 5–25.

[18] B. Gaüzére, L. Brun, D. Villemin and M. Brun, *Graph kernels based on relevant patterns and cycle information for chemoinformatics*, Paper Presented at the Pattern Recognition (ICPR), 2012 21st International Conference on, 2012.

[19] M. Glick and E. Jacoby, The role of computational methods in the identification of bioactive compounds, *Current Opinion in Chemical Biology* **15**(4) (2011), 540–546.

[20] D. Guan, W. Yuan, T. Ma and S. Lee, Detecting potential labeling errors for bioinformatics by multiple voting, *Knowledge-Based Systems* **66** (2014), 28–35.

[21] H.A. Güvenir and M. Çakir, Voting features based classifier with feature construction and its application to predicting financial distress, *Expert Systems with Applications* **37**(2) (2010), 1713–1718.

[22] D. Harnie, M. Saey, A.E. Vapirev, J.K. Wegner, A. Gedich, M. Steijaert and W. De Meuter, Scaling machine learning for target prediction in drug discovery using apache spark, *Future Generation Computer Systems* **67** (2017), 409–417.

[23] D. Hecht, Applications of machine learning and computational intelligence to drug discovery and development, *Drug Development Research* **72**(1) (2011), 53–65.

[24] J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, 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* **46**(2) (2006), 462–470.

[25] 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* **60** (2016), 59–71.

[26] J.-H. Huang, M. Wen, L.-J. Tang, H.-L. Xie, L. Fu, Y.-Z. Liang and H.-M. Lu, Using random forest to classify linear B-cell epitopes based on amino acid properties and molecular features, *Biochimie* **103** (2014), 1–6.

[27] 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, *Analyticachimicaacta* **804** (2013), 70–75.

[28] 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* **16** (2015), 334–351.

[29] N.J. Jaradat, M.A. Khanfar, M. Habash and M.O. Taha, Combining docking-based comparative intermolecular contacts analysis and k-nearest neighbor correlation for the discovery of new check point kinase 1 inhibitors, *Journal of Computer-Aided Molecular Design* **29**(6) (2015), 561–581.

[30] R. Jegadeeshwaran and V. Sugumaran, Comparative study of decision tree classifier and best first tree classifier for fault diagnosis of automobile hydraulic brake system using statistical features, *Measurement* **46**(9) (2013), 3247–3260.

[31] M.A. Khamis, W. Gomaa and W.F. Ahmed, Machine learning in computational docking, *Artificial Intelligence in Medicine* **63**(3) (2015), 135–152.

[32] J. Klekota and F.P. Roth, Chemical substructures that enrich for biological activity, *Bioinformatics* **24**(21) (2008), 2518–2525.

[33] S. Korkmaz, G. Zararsiz and D. Goksuluk, Drug/nondrug classification using support vector machines with various feature selection strategies, *Computer Methods and Programs in Biomedicine* **117**(2) (2014), 51–60.

[34] A. Koutsoukas, R. Lowe, Y. KalantarMotamedi, H.Y. Mussa, W. Klaffke, J.B. Mitchell and A. Bender, 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* **53**(8) (2013), 1957–1966.

[35] A. Lavecchia, Machine-learning approaches in drug discovery: Methods and applications, *Drug Discovery Today* **20**(3) (2015), 318–331.

[36] R. Lowe, H.Y. Mussa, F. Nigsch, R.C. Glen and J.B. Mitchell, Predicting the mechanism of phospholipidosis, *Journal of Cheminformatics* **4**(1) (2012), 2.

[37] S.M. Paul, D.S. Mytelka, C.T. Dunwiddie, C.C. Persinger, B.H. Munos, S.R. Lindborg and A.L. Schacht, How to improve R&D productivity: The pharmaceutical industry's grand challenge, *Nature Reviews Drug Discovery* **9**(3) (2010), 203–214.

[38] 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* **8**(4) (2007), 329–351.

[39] N. Schneider, C. Jäckels, C. Andres and M.C. Hutter, Gradual in silico filtering for druglike substances, *Journal of Chemical Information and Modeling* **48**(3) (2008), 613–628.

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

[41] 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* **151** (2016), 71–77.

[42] A. Verikas, A. Gelzinis and M. Bacauskiene, Mining data with random forests: A survey and results of new tests, *Pattern Recognition* **44**(2) (2011), 330–349.

[43] 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* **193** (2016), 201–212.

[44] Y.-T. Yan, Y.-P. Zhang, J. Chen and Y.-W. Zhang, Incomplete data classification with voting based extreme learning machine, *Neurocomputing* **193** (2016), 167–175.

[45] W.-L.R. Yang, Y.-E. Lee, M.-H. Chen, K.-M. Chao and C.-Y.F. Huang, In-silico drug screening and potential target identification for hepatocellular carcinoma using Support Vector Machines based on drug screening result, *Gene* **518**(1) (2013), 201–208.

[46] H. Zhang, Y.-L. Kang, Y.-Y. Zhu, K.-X. Zhao, J.-Y. Liang, L. Ding and J. Zhang, Novel naïve Bayes classification models for predicting the chemical Ames mutagenicity, *Toxicology in Vitro* **41** (2017), 56–63.

[47] L. Zhang, L. Jiang, C. Li and G. Kong, Two feature weighting approaches for naive Bayes text classifiers, *Knowledge-Based Systems* **100** (2016), 137–144.

[48] 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*, **235** 2010.

[49] L. Zhou, S. Pan, J. Wang and A.V. Vasilakos, Machine Learning on Big Data: Opportunities and Challenges, *Neurocomputing* **237** (2017), 350–361.

[50] S. Zhou, G.-B. Li, L.-Y. Huang, H.-Z. Xie, Y.-L. Zhao, Y.-Z. Chen and S.-Y. Yang, A prediction model of drug-induced ototoxicity developed by an optimal support vector machine (SVM) method, *Computers in Biology and Medicine* **51** (2014), 122–127.