Hindawi
Computational and Mathematical Methods in Medicine
Volume 2022, Article ID 9547317, 13 pages
https://doi.org/10.1155/2022/9547317

Research Article

Similarity-Based Method with Multiple-Feature Sampling for Predicting Drug Side Effects

Zixin Wu and Lei Chen

College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

Correspondence should be addressed to Lei Chen; chen_lei1@163.com

Received 20 July 2021; Revised 18 September 2021; Accepted 15 March 2022; Published 1 April 2022

Academic Editor: Rafik Karaman

Copyright © 2022 Zixin Wu and Lei Chen. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Drugs can treat different diseases but also bring side effects. Undetected and unaccepted side effects for approved drugs can greatly harm the human body and bring huge risks for pharmaceutical companies. Traditional experimental methods used to determine the side effects have several drawbacks, such as low efficiency and high cost. One alternative to achieve this purpose is to design computational methods. Previous studies modeled a binary classification problem by pairing drugs and side effects; however, their classifiers can only extract one feature from each type of drug association. The present work proposed a novel multiple-feature sampling scheme that can extract several features from one type of drug association. Thirteen classification algorithms were employed to construct classifiers with features yielded by such scheme. Their performance was greatly improved compared with that of the classifiers that use the features yielded by the original scheme. Best performance was observed for the classifier based on random forest with MCC of 0.8661, AUROC of 0.969, and AUPR of 0.977. Finally, one key parameter in the multiple-feature sampling scheme was analyzed.

1. Introduction

Drugs are important in treating various diseases; however, their therapeutic effects are accompanied by negative effects called side effects. In the pharmaceutical field, drug side effect is classified as an adverse drug reaction (ADR), the harmful or accidental reactions of qualified drugs that are irrelevant to the purpose of their use under normal usage and dosage. Some market-approved drugs may generate unaccepted side effects that can be harmful to the human body and bring high risks to pharmaceutical companies. For example, fluconazole and atorvastatin have potential hepatotoxicity and nephrotoxicity that can increase transaminase when used in specific patients such as those with liver disease. Side effects are one of the major obstacles in launching new drugs and delaying their development. Thus, determining all the side effects for a given drug is an important topic in drug development. Despite their efficiency in identifying side effects, solid clinical trials are time consuming and expensive and thus cannot meet the demand of large-scale tests. Thus, rapid and cheap methods for the identification of drug side effects must be developed.

Many advanced computational algorithms have been proposed [1–5] to provide strong technique support to deal with various medical problems. Several computational methods have been developed for the identification of drug side effects. Most of them are machine learning-based techniques that deeply investigate current information on drug side effects and develop proper patterns that can be used to predict side effects for a given new drug. Some early methods consisted of an individual binary classifier for each side effect [6–10]; hence, they always contain several binary classifiers that must be simultaneously executed to determine all side effects for a given drug. In view of this situation, some other techniques were directly built with multilabel classifiers [11–16] that identify side effects as labels and drugs as samples. Recommender systems were also proposed to predict drug side effects [17–19]. Recent works paired drugs and side effects as samples to convert the original problem as binary classification [20–22]. A key step in developing such binary classifiers is to extract essential properties from each drug–side effect pair. Some researchers used a similarity-based scheme to extract features [21, 22]; for convenience, they extracted only one feature from one type of drug association, a process called single-feature sampling scheme. However, some essential information may be omitted. For research continuation, a novel feature extraction scheme that can hold essential information for each drug–side effect pair must be developed.

In this study, an efficient binary classifier was proposed for the identification of drug side effects. Drugs and side effects were also paired as samples [20–22]. The single-feature sampling scheme [21, 22] was generalized to extract essential features from each pair. Named as multiple-feature sampling scheme, this newly proposed strategy can generate multiple features from each type of drug association. Classic machine learning algorithm, random forest (RF) [23], was adopted as the prediction engine. According to the 10-fold cross-validation results, the performance of such classifier was better than that of the previous classifier that uses original single sampling scheme for feature extraction. Further tests suggested that classifiers with other classification algorithms and features yielded by the multiple sampling scheme were all superior to those with the same classification algorithm and features generated by the original scheme. This finding indicated the power of the features generated by the proposed feature extraction scheme.

2. Materials and Methods

2.1. Benchmark Dataset. Data on 841 drugs and their side effects (824) [20–22] were extracted from SIDER (http://sideeffects.embl.de/) [24], a public database collecting the information of marketed drugs and their ADRs. The original data contained 888 drugs and 1385 side effects. The side effects that were annotated to no more than five drugs were excluded. Furthermore, drugs without the properties mentioned in Section 2.2 were discarded. From the remaining 841 drugs and 824 side effects, 57,058 drug–side effect pairs were obtained. Each pair indicated that the specific drug in the pair has the side effect in the same pair. Given that these pairs indicate the relationship between one drug and one side effect, they were termed as positive samples and comprised the positive dataset (PDS).

In addition to PDS, a negative dataset (NDS) was necessary in building an efficient binary classifier. A total of 57,058 drug–side effect pairs were produced by randomly pairing one drug and one side effect [20, 21]. However, no pairs can be labeled as positive samples. Therefore, these pairs constituted one NDS. Different NDSs may influence the performance of the classifier. Therefore, four other NDSs were also generated. Finally, five datasets each containing the PDS and one NDS were produced and denoted by DS_1, DS_2, DS_3, DS_4, and DS_5.

2.2. Drug Association Obtained from Different Drug Properties. Two drugs with strong associations always share similar functions [25–29]. Side effects can be deemed as one type of drug function. Thus, classifiers can be constructed by adopting features derived from drug associations. From different aspects of drugs, several types of drug associations can be measured and quantified. For easy comparisons, the drug associations adopted in a previous study [21] were adopted, and their brief descriptions are as follows.

2.2.1. Drug Fingerprint Association. Simplified molecular input line entry specification (SMILES) string [30] is a widely used scheme for drug representation. Fingerprints can be extracted from this string using existing software, such as RDKit [31]. The associations of two drugs can be evaluated by comparing their fingerprints. Here, ECFP_4 fingerprints and Tanimoto coefficient were used to measure such association between any two drugs. For formulation, this association for drugs d_1 and d_2 was denoted by $G^f(d_1, d_2)$.

2.2.2. Drug Structural Association. In addition to SMILES string, another popular drug representation scheme is graph-based method. Here, each drug is represented by a graph with nodes depicting atoms and edges indicating bonds. The association of two drugs can be assessed by considering the similarity of two corresponding graphs. "SIMCOMP" (https://www.genome.jp/tools/simcomp/) reported in the KEGG [32, 33] was set up based on such idea. This tool can output the associations of a given drug with other drugs as measured by scores between 0 and 1. Such association for drugs d_1 and d_2 was denoted by $G^s(d_1, d_2)$.

2.2.3. Drug Anatomical Therapeutic Chemical (ATC) Code Association. The ATC system is a widely accepted and used in drug classification. Each drug in such system is assigned five-level ATC codes that indicate its essential properties. For two drugs, their association can be measured according to their ATC codes. This study used the same method in [21] to evaluate drug association based on their ATC codes. For convenience, the association of drugs d_1 and d_2 was denoted by $G^a(d_1, d_2)$.

2.2.4. Drug Literature Association. Given the extensive literature on drugs, the association of two drugs can be measured from their cooccurrence in some literature and natural language processing methods. The well-known public database, STITCH (version 4.0, http://stitch4.embl.de/) [34], provides such associations, which were directly employed in this study. "Textmining" score was extracted from the downloaded file "chemical_chemical.links.detailed.v4.0.tsv." For drugs d_1 and d_2, their literature association was denoted by $G^{tm}(d_1, d_2)$.

2.2.5. Drug Target Protein Association. Target protein is the basic property of drugs. Hence, the association of two drugs can be estimated by comparing their target proteins. In this study, the target proteins of drugs were retrieved from DrugBank (https://go.drugbank.com/) [35]. Each drug was encoded into a binary vector by applying one-hot scheme to its target proteins. The direction cosine of two vectors was defined as such association of two drugs. For formulation, this association between drugs d_1 and d_2 was denoted as $G^t(d_1, d_2)$.

2.3. Feature Engineering. In Section 2.2, five types of drug associations that have been used to extract features to represent drug–side effect pairs [21, 22] were employed. These features indicated the linkage between one drug and one side effect in a drug–side effect pair. However, they extract only one feature from each type of drug association and thus cannot fully capture the essential linkage between the drug and the side effect. This study proposed a novel feature extraction scheme called multiple-feature sampling scheme, which can extract multiple features from one type of drug association. For a clear description, some denotations are necessary. For one drug–side effect pair $p = <d, s>$, where d and s indicate one drug and one side effect, respectively, let S be a set consisting of drugs having side effect s that have been extracted from the training dataset. If d also has side effect s, then, it would not be included in S. For one type of drug association, all values between d and drugs in S are selected. Denoted by $\Psi^k(p)$ (where $k \in \{f, s, a, tm, t\}$ represents the type of drug association used to construct such list), a candidate feature list for p is then constructed with the decreasing order of above values. The top value in this list has been previously chosen as exclusive feature [21, 22]. Selection of several values in this list can contain more information to represent the linkage of drug d and side effect s. On the basis of the different selection models, two strategies were proposed, namely, discrete and continuous strategies. Their procedures are shown in Figure 1.

2.3.1. Discrete Strategy. In this strategy, several values from the list $\Psi^k(p)$ are selected to indicate the distribution of values in the list. In this way, these selected values can fully indicate the linkage between drug d and side effect s. This process can be achieved by selecting some discrete values in the list. For example, the value at the first place or that at the top q% place can be selected. These values comprise a set of features from one type of drug association.

2.3.2. Continuous Strategy. This strategy differs from the first one. Given that the linkage of drug d and side effect s is highly indicated by some top values in the list, these values must be properly selected because they may fully contain the essential information. For an integer q between 1 and 100, the top q% values in the list $\Psi^k(p)$ were selected as features.

2.4. Classification Algorithm. A proper classification algorithm is important in building an efficient classifier. In this study, RF [23] was adopted to construct the classifier. RF is one of the most classic classification algorithms and has been used to set up many classifiers in bioinformatics [36–41].

RF is an integrated classification algorithm containing several decision trees, each of which is constructed by two random selection procedures. The first procedure is to select samples. Given a dataset with n samples, randomly select n samples with replacement from such dataset. The second procedure is to select features to split each node. The selected features should be much less than overall features. After the predefined number of decision trees has been constructed, RF integrates them by major voting. For a query sample, each decision tree gives its prediction. The majority prediction is the predicted result of RF. Although a decision tree is a relative weak classification algorithm, RF is extremely powerful and has always been an important candidate to build different classifiers.

In this study, "RandomForest" in Weka [42] was directly used to implement the abovementioned RF. Default parameters were adopted, and the number of decision trees was set to 100.

In addition to RF, the following classification algorithms were used to build corresponding classifiers: support vector machine (SVM) (polynomial kernel, RBF kernel) [43], Adaboost M1 [44], Bagging [45], Bayesian network [46], Naive Bayes [47], K-nearest neighbor (KNN) [48], decision tree (C4.5) [49], PART [50], logistic regression [51], multilayer perceptron (MLP) [52], and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [53]. The goal is to confirm that the features yielded by the multiple sampling scheme are more effective than those yielded by the single sampling scheme. For convenience, corresponding tools in Weka were used to implement the above classification algorithms under default parameters. These classification algorithms adopt different principles and procedures for classification. Therefore, their usage can fully test the utility of the proposed feature sampling scheme. If the classifier with features yielded by the multiple sampling scheme is superior to that with previous features for any of these classification algorithms, then, the robustness of the novel features obtained by the multiple sampling scheme is confirmed.

2.5. Accuracy Measurement. Ten-fold cross-validation [54–59] was adopted to evaluate the performance of all constructed classifiers. Such method randomly divides the original dataset into ten parts. Each part is singled out one by one as the test set, and the remaining parts constitute the training set. Samples in the test set are predicted by the classifier based on the training set. Thus, each sample is tested exactly once.

For a binary classification problem, four entries can be counted by comparing the predicted and true classes of each sample, that is, true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The following measurements were based on these four entries: sensitivity (SN) (also called recall), specificity (SP), prediction accuracy (ACC), Matthews correlation coefficient (MCC) [20, 21, 37, 60–63], precision, and $F1$-measure. Their definitions are as follows:

$$\mathrm{SN(recall)} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{1}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{2}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}}, \tag{3}$$

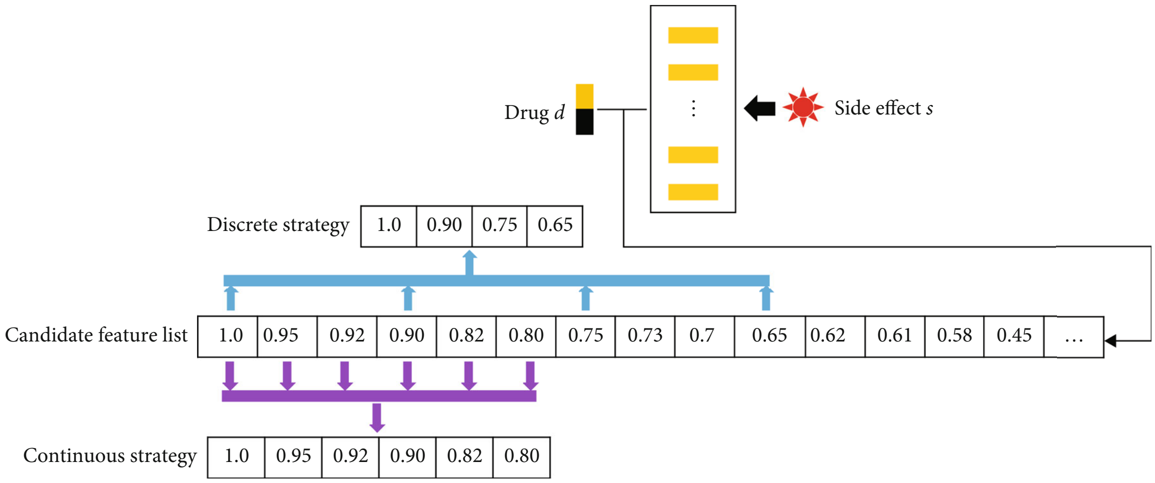


Figure 1: Procedures of multiple-feature sampling scheme to extract essential features from a drug–side effect pair. For a pair of drug d and side effect s, drugs having the side effect s are extracted from the training dataset. The association scores between d and these drugs constitute a candidate feature list. The discrete strategy selects discrete values in such list as features, and the continuous strategy picks up some top values in this list as features.

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}\text{-}\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TN}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})}}, \tag{4}$$

$$\mathrm{precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{5}$$

$$F1-\mathrm{measure}=\frac{2\times\mathrm{precision}\times\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}}. \tag{6}$$

ACC, MCC, and $F1$-measure use all four entries and thus are more important than the other three measurements. Receiver operating characteristic (ROC) curve [64] and precision-recall (PR) curve were further employed to fully assess the performance of constructed classifiers. These curves indicate the performance of classifiers under different thresholds. ROC curve takes 1-SP as x-axis and SN as the y-axis, and PR curve takes recall as x-axis and precision as y-axis. Areas under these two curves (AUROC and AUPR) are important measurements to evaluate the performance of classifiers. Among the abovementioned parameters, MCC was selected as the main measurement.

3. Results and Discussion

A novel feature extraction method was proposed to extract essential features from drug–side effect pairs. On the basis of these features, efficient classifiers to predict drug side effects were established. All procedures are illustrated in Figure 2.

3.1. Performance of the RF Classifiers with Discrete Strategy. The discrete strategy picks some discrete values in the candidate feature list. Given that the top value in such list is the most important and has been previously selected as the exclusive feature [21, 65], this top value is always picked up as one feature. As mentioned in Section 2.3, the value located at top q% place in the list was also selected. In this study, q was set as 5, 10, 15, and 20. Values with high ranks in the candidate feature list are more important than those with low ranks, that is, the top value is the most important, followed by values at 5%, 10%, 15%, and 20%. Incremental feature selection was adopted to generate four feature subsets as listed in column 1 of Table 1. With each feature subsets derived from five types of drug associations, a RF classifier was built on each of five datasets and evaluated by 10-fold cross-validation. The average performance is listed in Table 1. MCC followed an increasing trend when the values at top 5%, 10%, 15%, and 20% were added. Other five measurements also generally followed such trend. The RF classifiers with all selected features (top values and those at 5%, 10%, 15%, and 20%) generated the highest MCC of 0.7172. This finding indicated that the features yielded by such multiple-feature sampling scheme were quite efficient for the identification of drug side effects.

The ROC and PR curves of these four RF classifiers were investigated, and the results are shown in Figure 3. All AUROCs and AUPRs were higher than 0.900 and 0.910, respectively, thus, further suggesting the good performance of RF classifiers with discrete strategy.

3.2. Performance of RF Classifiers with Continuous Strategy. Different from discrete strategy, continuous strategy selected values from the candidate feature list in a continuous way. As mentioned in Section 2.3, top q% values in the candidate feature list can be chosen as features. Here, some q values including 10, 20, 30, and 40 and four feature subsets were tested. A RF classifier was also built on each of the five datasets by using the feature subsets derived from the five types of drug associations. Each classifier was assessed by 10-fold cross-validation, and the average performance is listed in Table 2. When $q = 20$ (top 20%), the RF classifier yielded the highest MCC of 0.8661 and generated the ACC of 0.9312, $F1$-measure of 0.9278, SN of 0.8852, SP of 0.9771,

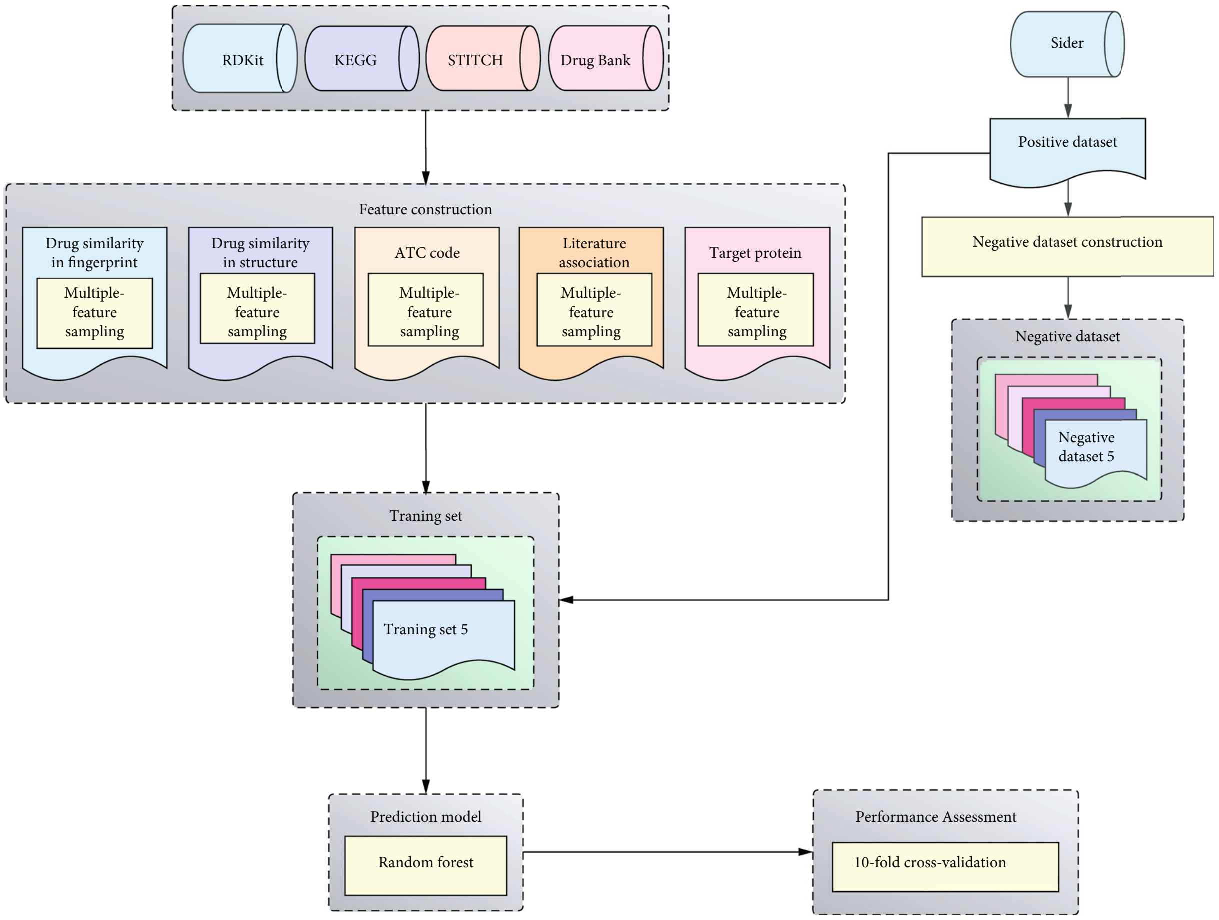


Figure 2: Entire procedures of the method for identification of drug side effects. Positive dataset (reported drug–side effect pairs) is retrieved from SIDER, and five negative datasets are randomly generated. From the four public databases or tools, five drug properties are employed and used to extract features with multiple-feature sampling scheme. Random forest is adopted to build the model and is further evaluated by 10-fold cross-validation.

Table 1: Performance of the RF classifiers with discrete strategy.

Feature sampling	SN	SP	ACC	MCC	Precision	$F1$ measure
Top + 5%	0.8072	0.8694	0.8383	0.6780	0.8608	0.8331
Top + 5% + 10%	0.8209	0.8829	0.8519	0.7058	0.8751	0.8467
Top + 5% + 10% + 15%	0.8214	0.8907	0.8561	0.7145	0.8825	0.8505
Top + 5% + 10% + 15% + 20%	0.8201	0.8944	0.8573	0.7172	0.8860	0.8514

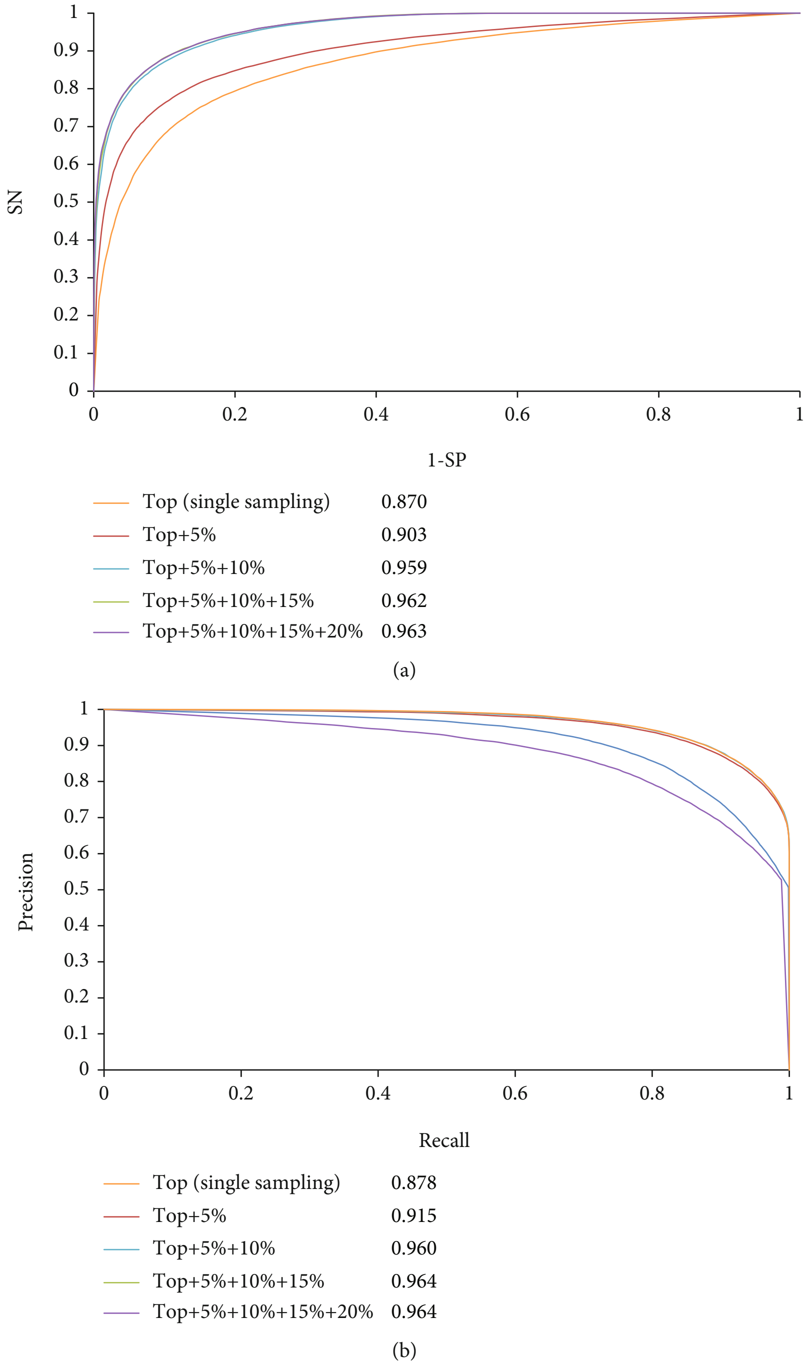
and precision of 0.9747. Compared with the RF classifiers with discrete strategy, the best RF with continuous strategy had higher measurements, particularly for MCC (by 15%), ACC (by 7%), and $F1$-measure (by 7%). These results indicated that the features obtained by continuous strategy were more powerful in identifying drug side effects than those yielded by discrete strategy.

The ROC and PR curves of RF classifiers with continuous strategy were plotted as shown in Figure 4. All ROC curves were close to the point (0, 1), and all PR curves were close to the point (1, 1). The AUROCs and AUPRs were all quite high. Compared with AUROCs and AUPRs for discrete strategy, those for continuous strategy were generally higher. This finding further confirmed that the features yielded by continuous strategy were more powerful than those yielded by discrete strategy.

3.3. Comparison of RF Classifiers with Single- and Multiple-Feature Sampling.

3.3. Comparison of RF Classifiers with Single- and Multiple-Feature Sampling. A multiple-feature sampling scheme was proposed to extract essential features from each drug–side effect pair. Previous studies [21, 22] only picked up the top value as the feature, and this technique was called single sampling scheme. This section compares the RF classifiers with these two feature sampling schemes.

The average performances of RF classifiers with single-feature sampling scheme are listed in Table 3. The MCC



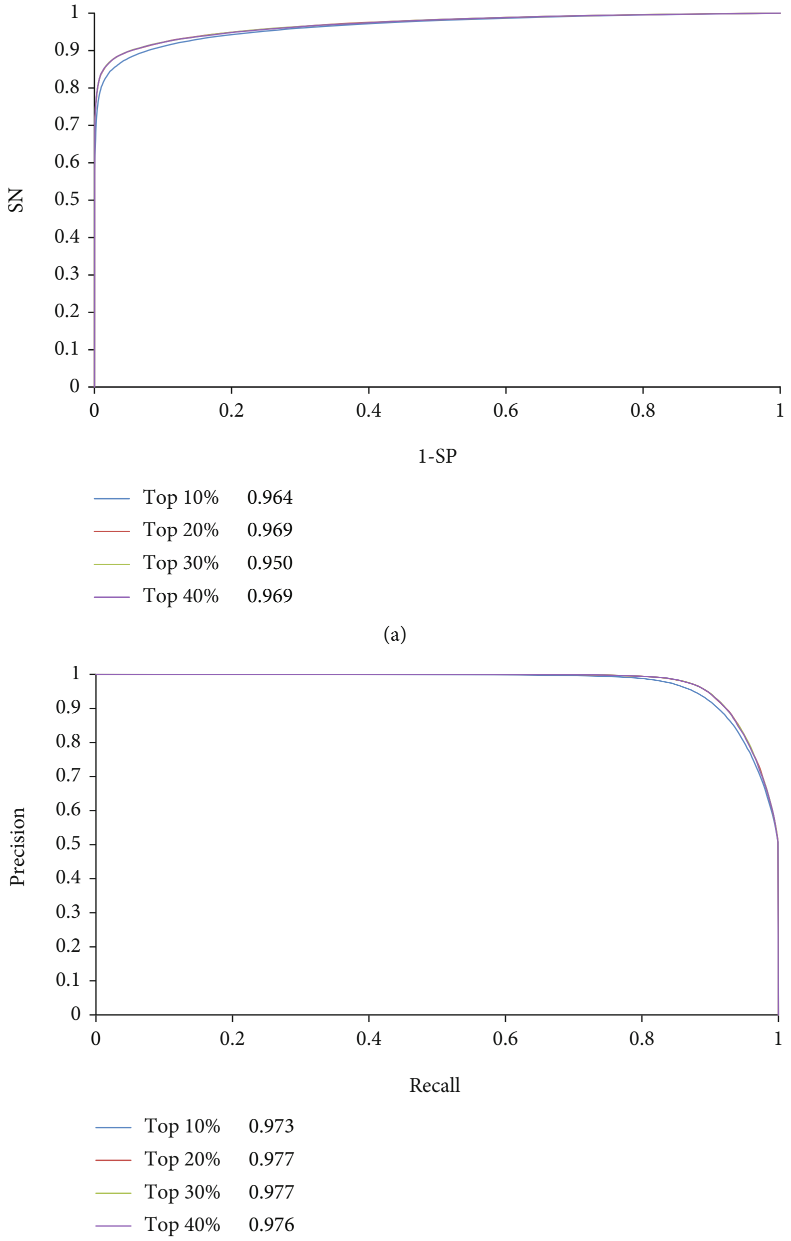
(a)

(b)

Figure 3: Receiver operating characteristic (ROC) curve and precision-recall (PR) curve of RF classifiers with single-feature sampling scheme and multiple-feature sampling scheme (discrete strategy). (a) ROC curves and (b) PR curves.

Table 2: Performance of the RF classifiers with continuous strategy.

Feature sampling	SN	SP	ACC	MCC	Precision	*F*1 measure
Top 10%	0.8737	0.9644	0.9190	0.8416	0.9609	0.9152
Top 20%	0.8852	0.9771	0.9312	0.8661	0.9747	0.9278
Top 30%	0.8844	0.9770	0.9307	0.8652	0.9747	0.9273
Top 40%	0.8834	0.9775	0.9305	0.8648	0.9751	0.9270



(a)

(b)

Figure 4: Receiver operating characteristic (ROC) curve and precision-recall (PR) curve of RF classifiers with multiple-feature sampling scheme (continuous strategy). (a) ROC curves and (b) PR curves.

Table 3: Comparison of RF classifiers with single- and multiple-feature sampling schemes.

Scheme		SN	SP	ACC	MCC	Precision	*F*1 measure
Single sampling		0.7948	0.8049	0.7999	0.5997	0.8030	0.7988
Multiple sampling	Discrete strategy	0.8201	0.8944	0.8573	0.7172	0.8860	0.8514
	Continuous strategy	0.8852	0.9771	0.9312	0.8661	0.9747	0.9278

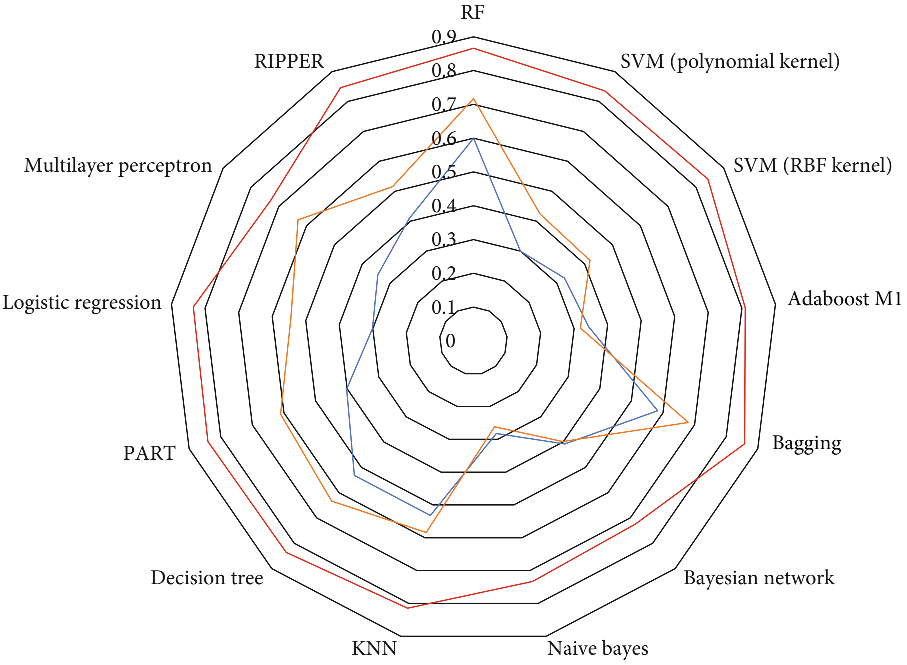
was 0.5997, ACC was 0.7999, and *F*1-measure was 0.7988. Other three measurements (SN, SP, and precision) were 0.7948, 0.8049, and 0.8030, respectively. The best performing (highest MCC) RF classifiers with discrete and continuous strategies were selected for comparison and are also listed in Table 3. The MCCs for two strategies were 0.7172 and 0.8661, which were higher than that for the RF classifier with single-feature sampling scheme. Same conclusions can

Table 4: Performance of classifiers with different classification algorithms and feature extraction schemes.

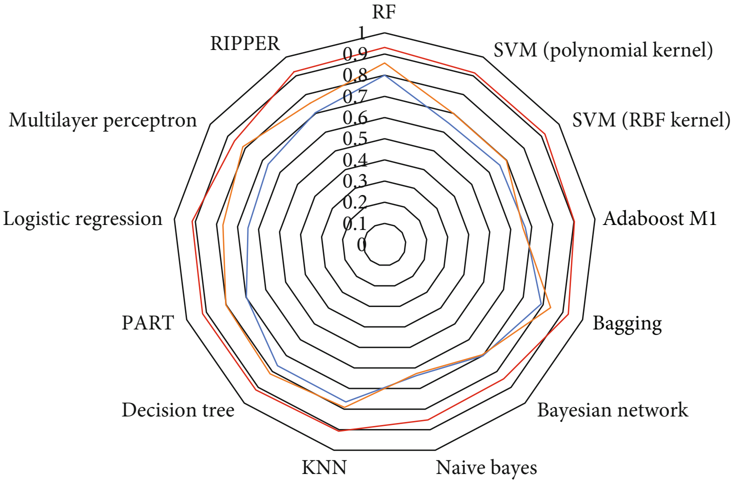
Classification algorithm	Feature extraction scheme		ACC	MCC	$F1$-measure
SVM (polynomial kernel)	Single sampling		0.6487	0.2997	0.6252
	Multiple sampling	Discrete strategy	0.6989	0.4240	0.6357
		Continuous strategy	0.9152	0.8356	0.9101
SVM (RBF kernel)	Single sampling		0.6608	0.3276	0.6251
	Multiple sampling	Discrete strategy	0.6987	0.4188	0.6415
		Continuous strategy	0.9191	0.8428	0.9147
Adaboost M1	Single sampling		0.6693	0.3435	0.6392
	Multiple sampling	Discrete strategy	0.6574	0.3186	0.6287
		Continuous strategy	0.9024	0.8102	0.8963
Bagging	Single sampling		0.7909	0.5828	0.7848
	Multiple sampling	Discrete strategy	0.8386	0.6799	0.8317
		Continuous strategy	0.9273	0.8580	0.9240
Bayesian network	Single sampling		0.7007	0.4076	0.6722
	Multiple sampling	Discrete strategy	0.6950	0.3980	0.6614
		Continuous strategy	0.8473	0.7236	0.8225
Naive Bayes	Single sampling		0.6368	0.2822	0.5859
	Multiple sampling	Discrete strategy	0.6272	0.2616	0.5782
		Continuous strategy	0.8528	0.7329	0.8296
KNN	Single sampling		0.7652	0.5321	0.7740
	Multiple sampling	Discrete strategy	0.7918	0.5838	0.7931
		Continuous strategy	0.9071	0.8148	0.9054
Decision tree	Single sampling		0.7635	0.5315	0.7471
	Multiple sampling	Discrete strategy	0.8154	0.6333	0.8080
		Continuous strategy	0.9170	0.8359	0.9142
PART	Single sampling		0.6986	0.4015	0.6753
	Multiple sampling	Discrete strategy	0.8022	0.6105	0.7874
		Continuous strategy	0.9192	0.8402	0.9166
Logistic regression	Single sampling		0.6501	0.3008	0.6383
	Multiple sampling	Discrete strategy	0.7690	0.5442	0.7515
		Continuous strategy	0.9157	0.8353	0.9115
Multilayer perceptron	Single sampling		0.6680	0.3438	0.6352
	Multiple sampling	Discrete strategy	0.8139	0.6305	0.8052
		Continuous strategy	0.8616	0.7299	0.8688
RIPPER	Single sampling		0.7037	0.4090	0.6904
	Multiple sampling	Discrete strategy	0.7546	0.5156	0.7382
		Continuous strategy	0.9215	0.8460	0.9181

be obtained for other five measurements. The ROC and PR curves of RF classifier with single-feature sampling scheme were also plotted (Figure 3) and were found to be always under those of RF classifiers with discrete strategy. The AUROC and AUPR of the RF classifier with single-feature sampling scheme were 0.870 and 0.878, respectively, which were also lower than those of the RF classifier with discrete strategy. For the RF classifier with continuous strategy, its AUROCs and AUPRs (Figure 4) were even better than those of the RF classifier with discrete strategy and were also higher than those of the RF classifier with single-feature sampling scheme. All these results implied that the features yielded by the multiple sampling scheme contained more essential information of drug–side effect pairs than those obtained by the single sampling scheme. These features provide RF with improved performance.

3.4. Performance of Other Classifiers with Multiple-Feature Sampling Scheme. The RF classifiers with features yielded by multiple sampling (discrete strategy) were superior to

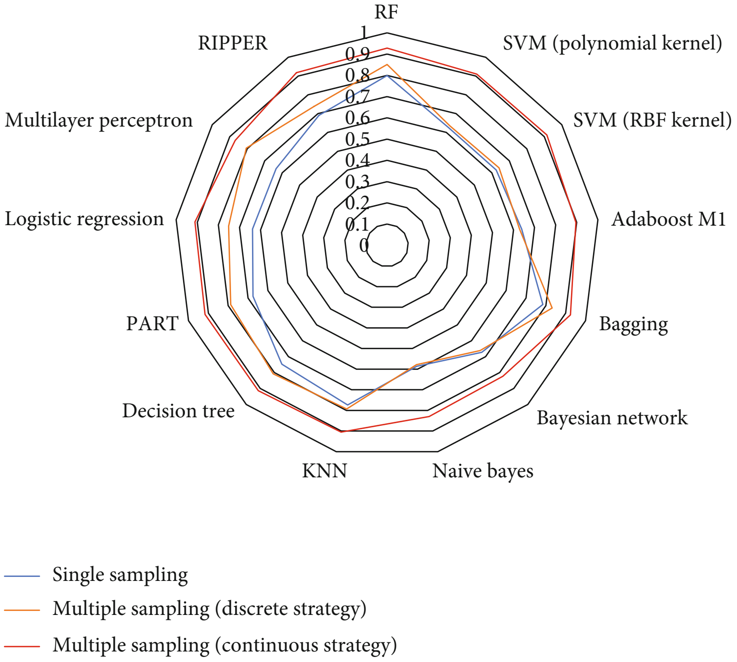


(a)



(b)

Figure 5: Continued.



(c)

Figure 5: Radar graphs to show performance of classifiers with single- and multiple-feature sampling schemes. (a) MCC; (b) ACC; (c) $F1$ -measure. Classifiers with multiple-feature sampling scheme (continuous strategy) provide best performance.

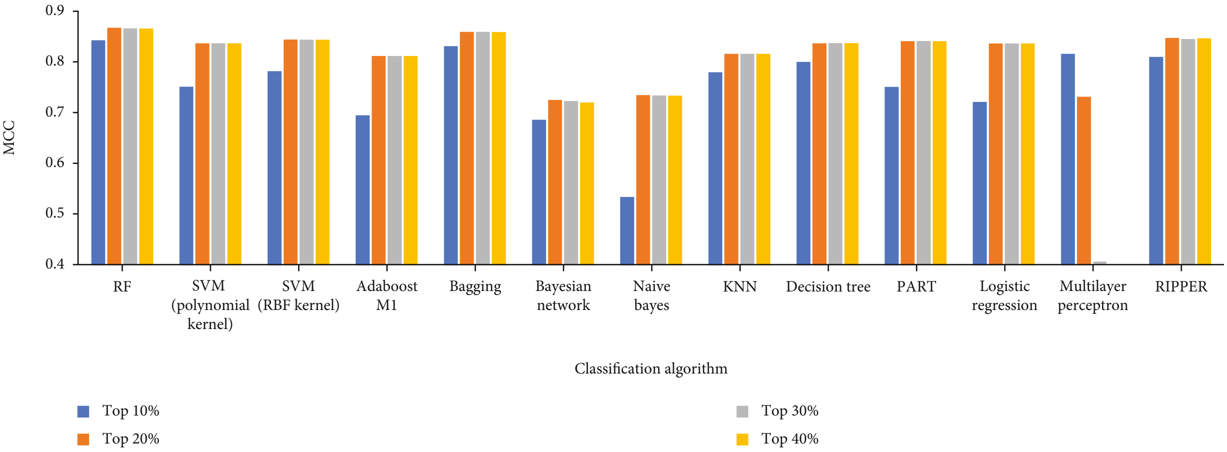


Figure 6: Performance of classifiers with continuous strategy under different parameters.

those with features yielded by single sampling, and the RF classifiers with continuous strategy were better than those with discrete strategy. However, the relevance of this result to the selection of classification algorithms must be explored. In this section, 12 classification algorithms mentioned in Section 2.4 were tested. The classifiers with different algorithms and all feature subsets used for RF were constructed and evaluated by 10-fold cross-validation. The predicted results are listed in Tables S1–S24.

The performances of classifiers with single sampling and the best performance of classifiers with multiple sampling are listed in Table 4. The classifiers with multiple sampling (discrete strategy) were generally better than those with single sampling, and those with continuous strategy were superior to those with discrete strategy and single sampling. For a visualized confirmation, a radar graph was plotted for each value of ACC, MCC, and $F1$-measure as illustrated in Figure 5. For each measurement, the area in the closed curve of classifiers with multiple sampling (continuous strategy) was the largest, followed by the closed curve of classifiers with multiple sampling (discrete strategy); the area in the closed curve of classifiers with single sampling was the smallest. On the basis of these results, multiple sampling scheme is more efficient to capture the essential properties of drug–side effect pairs than single sampling scheme, and continuous strategy is better than discrete strategy.

3.5. Analysis of the Parameter of Continuous Strategy. For the continuous strategy, the parameter q is a key factor that determines the number of selected features from the candidate feature list. Here, its influence on the performance of classifiers was investigated.

For RF classifiers, the highest MCC of 0.8661 was achieved when $q = 20$ (Table 2). For other classifiers with different classification algorithms, $q = 20$ always yields the best performance as shown in Figure 6. Among the 13 classifiers with different classification algorithms, 10 provided the best performance when $q = 20$, occupying 76.92%. Meanwhile, two yielded the best performance when $q = 30$. This phenomenon was reasonable. When q is extremely small, some essential information of drug–side effect pairs cannot be included. When q is large, several noises may be employed. Current investigation revealed that the values of q can be taken in an interval [20, 30].

4. Conclusions

This study prevents a novel investigation on drug side effects. The contributions contained two aspects. One was the multiple-feature sampling scheme that can extract essential features from drug–side effect pairs, and other one was novel computational methods for the identification of drug side effects based on the features yielded by the multiple sampling scheme. Classifiers were built on the basis of different classification algorithms. By comparison, the classifiers using features yielded by the multiple sampling scheme performed better than those using features yielded by the single sampling scheme. The proposed classifiers can be useful tools to identify drug side effects, and the novel feature extraction scheme can be applied to other similar biological or medical problems.

Data Availability

The original data used to support the findings of this study are available at SIDER and in supplementary information files.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Natural Science Foundation of Shanghai (17ZR1412500).

Supplementary Materials

Table S1: performance of SVM (polynomial kernel) classifier with discrete strategy. Table S2: performance of SVM (polynomial kernel) classifier with continuous strategy. Table S3: performance of SVM (RBF kernel) classifier with discrete strategy. Table S4: performance of SVM (RBF kernel) classifier with continuous strategy. Table S5: performance of Adaboost M1 classifier with discrete strategy. Table S6: performance of Adaboost M1 classifier with continuous strategy. Table S7: performance of Bagging classifier with discrete strategy. Table S8: performance of Bagging classifier with continuous strategy. Table S9: performance of Bayesian network classifier with discrete strategy. Table S10: performance of Bayesian network classifier with continuous strategy. Table S11: performance of Naive Bayes classifier with discrete strategy. Table S12: performance of Naive Bayes classifier with continuous strategy. Table S13: performance of KNN classifier with discrete strategy. Table S14: performance of KNN classifier with continuous strategy. Table S15: performance of decision tree classifier with discrete strategy. Table S16: performance of decision tree classifier with continuous strategy. Table S17: performance of PART classifier with discrete strategy. Table S18: performance of PART classifier with continuous strategy. Table S19: performance of logistic regression classifier with discrete strategy. Table S20: performance of logistic regression classifier with continuous strategy. Table S2: performance of multilayer perceptron classifier with discrete strategy. Table S22: performance of multilayer perceptron classifier with continuous strategy. Table S23: performance of RIPPER classifier with discrete strategy. Table S24: performance of RIPPER classifier with continuous strategy. *(Supplementary Materials)*

References

[1] A. Onan, S. Korukoğlu, and H. Bulut, “A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification,” *Expert Systems with Applications*, vol. 62, pp. 1–16, 2016.

[2] A. Onan, S. Korukoğlu, and H. Bulut, “Ensemble of keyword extraction methods and classifiers in text classification,” *Expert Systems with Applications*, vol. 57, pp. 232–247, 2016.

[3] A. Onan and S. Korukoğlu, “Exploring performance of instance selection methods in text sentiment classification,” in *Artificial Intelligence Perspectives in Intelligent Systems*, pp. 167–179, Springer, 2016.

[4] A. Onan, S. Korukoğlu, and H. Bulut, “A hybrid ensemble pruning approach based on consensus clustering and multiobjective evolutionary algorithm for sentiment classification,” *Information Processing & Management*, vol. 53, no. 4, pp. 814–833, 2017.

[5] A. Onan and M. A. Toçoğlu, “A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification,” *IEEE Access*, vol. 9, pp. 7701–7722, 2021.

[6] E. Pauwels, V. Stoven, and Y. Yamanishi, “Predicting drug side-effect profiles: a chemical fragment-based approach,” *BMC Bioinformatics*, vol. 12, no. 1, p. 169, 2011.

[7] S. Jamal, S. Goyal, A. Shanker, and A. Grover, “Predicting neurological adverse drug reactions based on biological, chemical and phenotypic properties of drugs using machine learning models,” *Scientific Reports*, vol. 7, no. 1, p. 872, 2017.

[8] Y. Zheng, H. Peng, S. Ghosh, C. Lan, and J. Li, “Inverse similarity and reliable negative samples for drug side-effect prediction,” *BMC Bioinformatics*, vol. 19, Suppl 13, p. 554, 2019.

[9] M. Liu, Y. Wu, Y. Chen et al., “Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs,” *Journal of the American Medical Informatics Association*, vol. 19, no. e1, pp. e28–e35, 2012.

[10] S. Dey, H. Luo, A. Fokoue, J. Hu, and P. Zhang, “Predicting adverse drug reactions through interpretable deep learning framework,” *BMC Bioinformatics*, vol. 19, Suppl 21, p. 476, 2018.

[11] L. Chen, T. Huang, J. Zhang et al., “Predicting drugs side effects based on chemical-chemical interactions and protein-chemical interactions,” *BioMed Research International*, vol. 2013, Article ID 485034, 8 pages, 2013.

[12] W. Zhang, F. Liu, L. Luo, and J. Zhang, “Predicting drug side effects by multi-label learning and ensemble learning,” *BMC Bioinformatics*, vol. 16, no. 1, p. 365, 2015.

[13] N. Atias and R. Sharan, “An algorithmic framework for predicting side effects of drugs,” *Journal of Computational Biology*, vol. 18, no. 3, pp. 207–218, 2011.

[14] E. Muñoz, V. Novácek, and P. Y. Vandenbussche, “Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models,” *Briefings in Bioinformatics*, vol. 20, no. 1, pp. 190–202, 2019.

[15] W. Zhang, Y. Chen, S. Tu, F. Liu, and Q. Qu, “Drug side effect prediction through linear neighborhoods and multiple data source integration,” in *IEEE International Conference on Bioinformatics and Biomedicine*, pp. 427–434, Shenzhen, Guangdong, China, 2016.

[16] E. Munoz, V. Novacek, and P. Y. Vandenbussche, “Using drug similarities for discovery of possible adverse reactions,” in *American Medical Informatics Association Annual Symposium Proceedings*, pp. 924–933, USA, 2016.

[17] Y. J. Ding, J. J. Tang, and F. Guo, “Identification of drug-side effect association via multiple information integration with centered kernel alignment,” *Neurocomputing*, vol. 325, pp. 211–224, 2019.

[18] X. Guo, W. Zhou, Y. Yu, Y. Ding, J. Tang, and F. Guo, “A novel triple matrix factorization method for detecting drug-side effect association based on kernel target alignment,” *BioMed Research International*, vol. 2020, Article ID 4675395, 11 pages, 2020.

[19] Y. Ding, J. Tang, and F. Guo, “Identification of drug-side effect association via semi-supervised model and multiple kernel learning,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2619–2632, 2019.

[20] X. Zhao, L. Chen, Z. H. Guo, and T. Liu, “Predicting drug side effects with compact integration of heterogeneous networks,” *Current Bioinformatics*, vol. 14, no. 8, pp. 709–720, 2019.

[21] X. Zhao, L. Chen, and J. Lu, “A similarity-based method for prediction of drug side effects with heterogeneous information,” *Mathematical Biosciences*, vol. 306, pp. 136–144, 2018.

[22] H. Liang, L. Chen, X. Zhao, and X. Zhang, “Prediction of drug side effects with a refined negative sample selection strategy,” *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 1573543, 16 pages, 2020.

[23] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, “A side effect resource to capture phenotypic effects of drugs,” *Molecular Systems Biology*, vol. 6, no. 1, p. 343, 2010.

[25] L. L. Hu, C. Chen, T. Huang, Y. D. Cai, and K. C. Chou, “Predicting biological functions of compounds based on chemical-chemical interactions,” *PLoS One*, vol. 6, no. 12, article e29491, 2011.

[26] L. Chen, W. M. Zeng, Y. D. Cai, K. Y. Feng, and K. C. Chou, “Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities,” *PLoS One*, vol. 7, no. 4, article e35254, 2012.

[27] L. Chen, J. Lu, N. Zhang, T. Huang, and Y. D. Cai, “A hybrid method for prediction and repositioning of drug anatomical therapeutic chemical classes,” *Molecular BioSystems*, vol. 10, no. 4, pp. 868–877, 2014.

[28] L. Chen, T. Liu, and X. Zhao, “Inferring anatomical therapeutic chemical (ATC) class of drugs using shortest path and random walk with restart algorithms,” *Biochimica et Biophysica Acta-Molecular Basis of Disease*, vol. 1864, no. 6, pp. 2228–2240, 2018.

[29] H. Y. Liang, B. Hu, L. Chen, S. Wang, and Aorigele, “Recognizing novel chemicals/drugs for anatomical therapeutic chemical classes with a heat diffusion algorithm,” *Biochimica et Biophysica Acta-Molecular Basis of Disease*, vol. 1866, no. 11, article 165910, 2020.

[30] D. Weininger, “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules,” *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[31] G. Landrum, “RDKit: open-source cheminformatics,” 2006, http://www.rdkit.org.

[32] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, “KEGG: new perspectives on genomes, pathways, diseases and drugs,” *Nucleic Acids Research*, vol. 45, no. D1, pp. D353–D361, 2017.

[33] M. Kanehisa and S. Goto, “KEGG: Kyoto encyclopedia of genes and genomes,” *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[34] M. Kuhn, D. Szklarczyk, S. Pletscher-Frankild et al., “STITCH 4: integration of protein–chemical interactions with user data,” *Nucleic Acids Research*, vol. 42, no. Database issue, pp. 401–407, 2014.

[35] D. S. Wishart, Y. D. Feunang, A. C. Guo et al., “DrugBank 5.0: a major update to the DrugBank database for 2018,” *Nucleic Acids Research*, vol. 46, no. D1, pp. D1074–D1082, 2018.

[36] M. Carlos, K. Zoran, and S. Juan, “Predicting non-deposition sediment transport in sewer pipes using random forest,” *Water Research*, vol. 189, p. 116639, 2021.

[37] Y. Jia, R. Zhao, and L. Chen, “Similarity-based machine learning model for predicting the metabolic pathways of compounds,” *IEEE Access*, vol. 8, pp. 130687–130696, 2020.

[38] D. V. Urista, D. B. Carrué, I. Otero et al., “Prediction of antimalarial drug-decorated nanoparticle delivery systems with random forest models,” *Biology*, vol. 9, no. 8, p. 198, 2020.

[39] Z. B. Lv, J. Zhang, H. Ding, and Q. Zou, “RF-PseU: a random forest predictor for RNA pseudouridine sites,” *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 10, 2020.

[40] M. Baranwal, A. Magner, P. Elvati, J. Saldinger, A. Violi, and A. O. Hero, “A deep learning architecture for metabolic pathway prediction,” *Bioinformatics*, vol. 36, no. 8, pp. 2547–2553, 2020.

[41] Y. Yang and L. Chen, “Identification of drug–disease associations by using multiple drug and disease networks,” *Current Bioinformatics*, vol. 17, no. 1, pp. 48–59, 2022.

[42] I. H. Witten and E. Frank, *Data Mining:Practical Machine Learning Tools and Techniques*, Kaufmann, San Francisco, Morgan, 2nd ed edition, 2005.

[43] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[44] Y. Freund and R. E. Schapire, “Experiments with a new boosting algorithm,” in *Thirteenth International Conference on ML*, Citeseer, 1996.

[45] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[46] S. Lee and S. Shimoji, “BAYESNET: Bayesian Classification Network Based on Biased Random Competition Using Gaussian Kernels,” in *IEEE International Conference on Neural Networks*, San Francisco, CA, USA, 1993.

[47] I. Rish, “An empirical study of the naive Bayes classifier,” in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, IBM New York, USA, 2001.

[48] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[49] R. Quinlan, *C4.5: Programs for Machine Learning.*, Morgan Kaufmann Publishers, San Mateo, CA, USA, 1993.

[50] E. Frank and I. H. Witten, “Generating accurate rule sets without global optimization,” in *15th International Conference on Machine Learning*, pp. 144–151, San Francisco, CA, USA, 1998.

[51] M. Sumner, E. Frank, and M. Hall, “Speeding up logistic model tree induction,” in *European Conference on Principles of Data Mining and Knowledge Discovery*, Springer, 2005.

[52] S. K. Pal and S. Mitra, *Multilayer perceptron, fuzzy sets, classifiaction*, IEEE, 1992.

[53] W. W. Cohen, “Fast effective rule induction,” in *Machine Learning Proceedings 1995*Morgan Kaufmann Publishers, Inc.

[54] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *International joint Conference on artificial intelligence*, Lawrence Erlbaum Associates Ltd., 1995.

[55] Y.-H. Zhang, Z. Li, T. Zeng et al., “Detecting the multiomics signatures of factor-specific inflammatory effects on airway smooth muscles,” *Frontiers in Genetics*, vol. 11, article 599970, 2021.

[56] Y. H. Zhang, H. Li, T. Zeng et al., “Identifying transcriptomic signatures and rules for SARS-CoV-2 infection,” *Frontiers in Cell and Development Biology*, vol. 8, article 627302, 2021.

[57] X. Pan, H. Li, T. Zeng et al., “Identification of protein subcellular localization with network and functional embeddings,” *Frontiers in Genetics*, vol. 11, article 626500, 2021.

[58] Y. Zhu, B. Hu, L. Chen, and Q. Dai, “iMPTCE-Hnetwork: a multi-label classifier for identifying metabolic pathway types of chemicals and enzymes with a heterogeneous network,” *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 6683051, 12 pages, 2021.

[59] J.-P. Zhou, L. Chen, and Z.-H. Guo, “iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs,” *Bioinformatics*, vol. 36, no. 5, pp. 1391–1396, 2020.

[60] B. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” *Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[61] Y.-H. Zhang, T. Zeng, L. Chen, T. Huang, and Y. D. Cai, “Determining protein-protein functional associations by functional rules based on gene ontology and KEGG pathway,” *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1869, no. 6, article 140621, 2021.

[62] L. Chen, C. Chu, Y. H. Zhang et al., “Identification of drug-drug interactions using chemical interactions,” *Current Bioinformatics*, vol. 12, no. 6, pp. 526–534, 2017.

[63] L. Chen, S. Wang, Y. H. Zhang et al., “Identify key sequence features to improve CRISPR sgRNA efficacy,” *IEEE Access*, vol. 5, pp. 26582–26590, 2017.

[64] J. Egan, *Signal Detection Theory and ROC Analysis*, Academic Press, New York, 1975.

[65] Z. Liu, F. Guo, J. Gu et al., “Similarity-based prediction for anatomical therapeutic chemical classification of drugs by integrating multiple data sources,” *Bioinformatics*, vol. 31, no. 11, pp. 1788–1795, 2015.