IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 21, 2020, accepted March 2, 2020, date of publication March 9, 2020, date of current version March 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2979452

Improving the Prediction of Adverse Drug Events Using Feature Fusion-Based Predictive Network Models

JIN LI[1], XIANGMIN JI[1,2], AND LIYAN HUA[1]
[1]College of Automation, Harbin Engineering University, Harbin 150001, China
[2]College of Information Engineering, Ordos Institute of Technology, Ordos 017000, China

Corresponding author: Xiangmin Ji (jixiangmin_hrbeu@foxmail.com)

This work was supported in part by the National Natural Science Foundation, China, under Grant 61773134 and Grant 61803117, in part by the Fundamental Research Funds for the Central Universities at Harbin Engineering University, China, under Grant HEUCFG201824 and Grant 3072019CFM0403, and in part by the Natural Science Foundation of Heilongjiang Province of China under Grant YQ2019F003.

ABSTRACT Computational strategies play a vital role in the prediction of adverse drug events (ADEs) owing to their low cost and increased efficiency. In this study, we used the strengths of the Jaccard and Adamic–Adar indices to build feature fusion-based predictive network models (FFPNMs) with three different machine learning (ML) methods respectively to predict drug–ADE associations. Our FFPNM with the logistic regression (LR) model improved to an area under the receiver operating characteristic curve (AUROC) value of 0.849, while the corresponding AUROC values for the pharmacological network model (PNM) and model based on similarity measures were 0.824 and 0.821, respectively. FFPNM with random forest (RF) is the best model among them with an AUROC value of 0.856, and the performance of FFPNM with SVM is close to that of FFPNM with RF and higher than that of FFPNM with LR. In these models, the bipartite network consisted of 152 drugs and 633 ADEs, which were obtained from the FDA Adverse Event Reporting System (FAERS) 2010 dataset. To better evaluate the performance of FFPNMs, we performed model predictions by different network consisting of 1177 drugs and 97 ADEs which were from the data of the first 120 days of FAERS 2004. FFPNM with RF achieved the best predictive result with AUROC value of 0.913. The results show that FFPNMs with ML methods, specially RF, have a superior prediction performance and robustness using only the topology features of the drug–ADE network. From our findings, the optimal, concise, and efficient models as computational methods for drug-ADE association predictions, were revealed. Source codes of this paper are available on https://github.com/Coderljl/FFPNM.

INDEX TERMS Adverse drug event, prediction, complex network, machine learning, local-information-based similarity measure, feature fusion-based predictive network model.

I. INTRODUCTION

Predicting adverse drug events (ADEs) accurately and earlier is a significant challenge for pharmacovigilance studies. In the United States, millions of people are hospitalized every year owing to ADEs [1]. In some cases, severe deaths have been reported, and ADEs have become the fourth leading cause of death after cancer and heart disease [2]. Various measures have been enforced to avoid increased morbidity and mortality rates due to ADEs. Unfortunately, small-scale clinical trials cannot detect rare events or ADEs during the pre-approval stage. Meanwhile, a lot of information is collected during the post-market phase in an effort to construct a variety of databases, which contain information on ADEs and support the post-market safety surveillance program. Therefore, the development of database-based computational methods to predict ADEs is urgent and necessary. Accordingly, the computational strategy complements this effort due to its low cost and increased efficiency.

Recently, many effective computational methods based on the various databases have been proposed for ADEs prediction, such as the well-known system pharmacology and signal detection algorithms. Signal detection algorithms have been developed to detect drug–ADE associations using FAERS or other similar databases [3]–[7]. Disproportionality analysis is one of the well-known signal detection algorithms based on methods such as the frequentist and Bayesian methods. At the same time, many system pharmacological

The associate editor coordinating the review of this manuscript and approving it for publication was Xiangtao Li.

methods have been proposed to predict ADEs [8]–[16]. System pharmacology is considered a promising approach that uses drug pharmacology attributes derived from multiple databases to build a predictive model using the pharmacological framework. In general, data is used and integrated by the network pharmacology model [17], [18].

Additionally, complex network-based methods have been extensively used in the field of bioinformatics, e.g., protein–protein relationships [19], the relationship between drugs and targets [20]–[22], and the relationship between drugs and diseases [23]–[25], which can be predicted based on the reconfiguration and topological properties of the network. For example, Cami *et al.* developed a pharmacological network model (PNM) that integrated the network structure with safety, taxonomic, and biological data to predict unknown drug–ADE associations [9]. Liu *et al.* proposed a machine-learning-based approach to predict 1385 ADEs of 832 approved drugs [11]. Cheng *et al.* published a drug side-effect similarity inference (DSESI) to predict the drug–target interaction with 621 drugs and 893 targets based on the drug side-effect database known as MetaADEDB, which included 1,330 drugs, 13,200 ADEs, and more than 520,000 drug-ADE associations [13]. Lin *et al.* published a network-based external link prediction method to predict the unknown drug adverse reactions only based on the use of topological features of complex networks [15]. Davazdahemami and Dursun proposed to combine network analytics approach with ML methods to predict ADEs [26]. Therefore, network-based models and machine learning (ML) methods have demonstrated the potential value and application in the predictions of ADEs.

Feature extraction and combination of extracted features are two important factors used for the prediction performance of models. Our goal is to design a simple and efficient network-based method to predict drug–ADE associations. Herein, we use a modified algorithm that uses the strengths of the Jaccard and the Adamic–Adar (AA) indices to extract improved features to build feature fusion-based predictive network models (FFPNMs) to improve the ADE predictions based on a bipartite network. In this network, nodes denote drugs or ADEs, and edges denote the drug–ADE associations. In our FFPNMs, the improved features, i.e., the Jaccard and AA drug fusion (JADF) and Jaccard and AA–ADE fusion (JAAF), are trained based on the use of the logistical regression (LR) model, support vector machines (SVM) and random forest (RF), respectively. As a comparison, a PNM and other models based on five different classical similarity measures are discussed in relation to the prediction of ADEs with LR based on the FAERS 2010 data of 152 drugs and 633 ADEs. Our FFPNM with LR achieved a superior prediction performance among the three tested models. Moreover, we investigated the performance of FFPNMs with SVM and RF respectively, and compared their performance with FFPNM with LR based on the above data. FFPNM with RF was evidently the best model among them, and the performance of FFPNM with SVM is close to that of FFPNM with RF and higher than that of FFPNM with LR. Finally, to further prove the superior performance of FFPNMs we used different data.

Our study differs from previous studies in that: (i) the similarity measures based on local information defined are introduced as the features for ADE predictions, (ii) we define the improved features as JADF and JAAF based on the modified algorithm, and propose the concise, efficient FFPNMs, in which different ML methods are used as the classification algorithms, to optimize the drug–ADE predictions.

II. MATERIALS AND METHODS

A. DATA DESCRIPTION

The US Food Drug Administration's (FDA) adverse event reporting system (FAERS) is the largest spontaneous reporting system that collects data from clinicians, individuals, and pharmaceutical companies, and is updated quarterly [27]. ADEs in FAERS are annotated in the Medical Dictionary for Regulatory Activities (MedDRA) [28]. Herein, we chose two sets of data for analysis, which included the FAERS data from 2010 to 2015 and the FAERS data from 2004 to 2009, respectively. Firstly, FAERS data from 2010 to 2015 that contained 152 cancer drugs and 633 ADEs was chosen for analysis. Drug Bank identities (IDs) were used to normalize the drug names [29], and ADEs with the preferred ADE terms (PTs) were mapped to their high-level terms (HLTs). Finally, 33947 drug-ADE associations were applied to construct a bipartite network with FAERS 2010 data, which was also used as the training set. Moreover, FAERS data from 2011 to 2015 were used to construct the validation set containing 21065 new drug–ADE associations, which were not in the training set.

Conversely, the first 120 days data of FAERS 2004 was selected to further evaluate our proposed models. The purpose of choosing this period was the early prediction of drug–ADE associations from the perspective of the FDA due to the collection of the FAERS data started in the first quarter of 2004. All drugs were selected during this period and were also standardized with the Drug Bank IDs. From these data, 1177 distinctive drugs were obtained as the training set. Furthermore, we selected ADEs that the hospitals focused on, including 97 ADEs (at the PT level). Finally, 10307 drug-ADE associations constructed the drug–ADE network with 1177 drugs and 97 ADEs, and the drug–ADE association network, as shown in Fig. 1. FAERS data from the rest of the period until 2009 was used as the validation set, which included 22358 new drug–ADE associations that were not present in the first 120 days of 2004, and an additional 1148 drug–ADE associations in the intersection of FAERS and side effects resource (SIDER) [30]. In summary, the training and the validation sets described above were all selected in chronological order.

B. PHARMACOLOGICAL NETWORK MODEL FEATURES

A PNM is proposed to predict the unknown ADEs based on the drug–ADE bipartite network. Using this network,

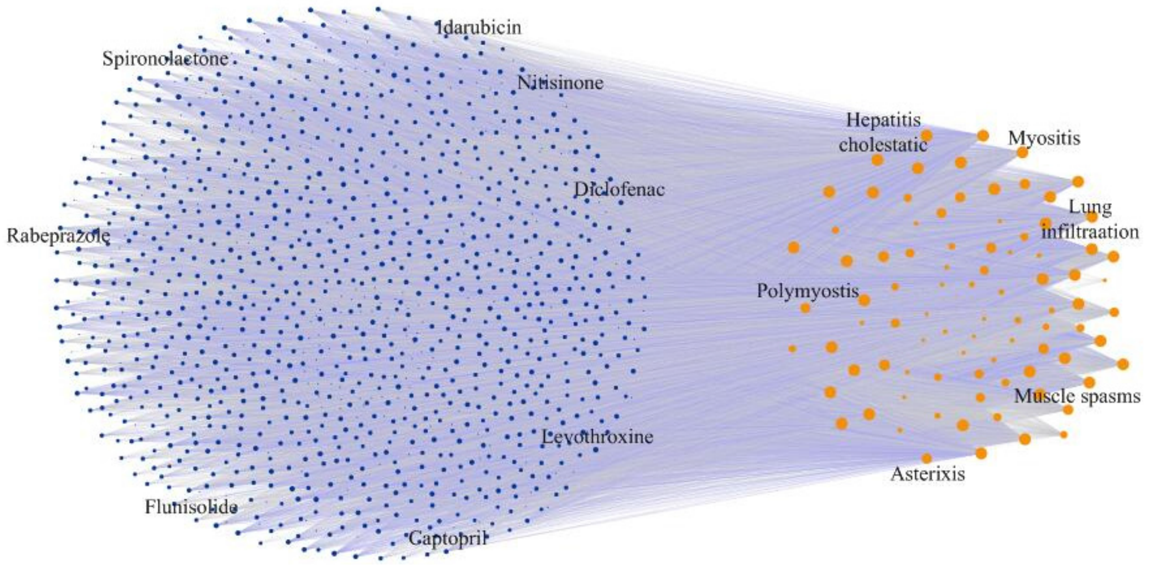


FIGURE 1. Visualization of drug-ADE bipartite network with 1177 drugs and 97 ADEs produced with Cytoscape (http://www.cytoscape.org). The drug nodes are colored in blue, and ADE nodes in orange. The size of an ADE node is proportional to its degree. The edge in the training set is purple, and edge in the validation set is gray. Only a few drugs and ADEs are labeled for illustration.

14 features, including eight network features, four taxonomic features, and two intrinsic features, were generated. An overview of a PNM is shown in Fig. 2.

Each feature is described in detail as follows: first, eight network features are based on the topological nature of the complex network. Among them, the degree-prod is the preferential attachment (PA), and denotes the similarity measure based on local structural information [31]. Degree-prod $X_1(i,j)$ and degree-sum $X_2(i,j)$ are given by the following expressions:

$$X_1(i,j) = D(i) \times D(j) \tag{1}$$

$$X_2(i,j) = D(i) + D(j) \tag{2}$$

where node i denotes drug, node j denotes the ADEs, and $D(i)$ and $D(j)$ denote the degrees of nodes i and j, respectively.

The degree ratio $X_3(i,j)$ and the degree absdiff $X_4(i,j)$ are the same as the defined above, as described by Cami *et al.* [9]. The aim of these four features is whether high-degree drugs tend to be associated with high-degree ADEs or low-degree ADEs. The Jaccard-drug-max and Jaccard-ADE-max indices are given by

$$X_5(i,j) = \max_{k \in N(j)-\{i\}} \frac{|N(i) \cap N(k)|}{|N(i) \cup N(k)|} \tag{3}$$

$$X_6(i,j) = \max_{k \in N(i)-\{j\}} \frac{|N(j) \cap N(k)|}{|N(j) \cup N(k)|} \tag{4}$$

Herein, $J(i,k) = |N(i) \cap N(k)| \big/ |N(i) \cup N(k)|$ is also known as the Jaccard index [32], and represents a similarity measure based on local information, and $N(i)$ denotes the set of neighbors of node i. The definitions of $N(j)$ and $N(k)$ are similar to those of $N(i)$. These features are used to quantify structural similarities between drug and ADE pairs.

The Jaccard–drug–Kullback–Leiber (KL) $[X_7(i,j)]$ and Jaccard–ADE–KL $[X_8(i,j)]$ features are used to take advantage of the full distribution of similarities in local neighborhoods between drugs and ADEs. These eight features described above are all the network features that use the topological features of drug–ADE network.

Taxonomic and intrinsic features consider the network structure and attributes of nodes. The attributes are the chemical biology codes, biochemistry, which could be obtained from PubChem [33], Anatomical Therapeutic Chemical classification system (ATC), and MedDRA codes. These features are listed in Table 1.

C. SIMILARITY MEASURES BASED ON LOCAL INFORMATION

For nodes *i* and *j*, similarity measures based on local structural information of complex network are as follows:

(1) Common Neighbors (CN) [34]: This measure assumes that if two nodes have more neighbors, they tend to have associations. It is defined as

$$S_{ij} = |N(i) \cap N(j)| \tag{5}$$

(2) Salton Index [35]: This measure is also called the cosine similarity, and it is defined as

$$S_{ij} = \frac{|N(i) \cap N(j)|}{\sqrt{D(i)\,D(j)}} \tag{6}$$

(3) Hub Promoted Index (HPI) [36]: This index considers nodes with larger degrees of associations that are more likely to be connections. It is defined as

$$S_{ij} = \frac{|N(i) \cap N(j)|}{min\left(D_i, D_j\right)} \tag{7}$$

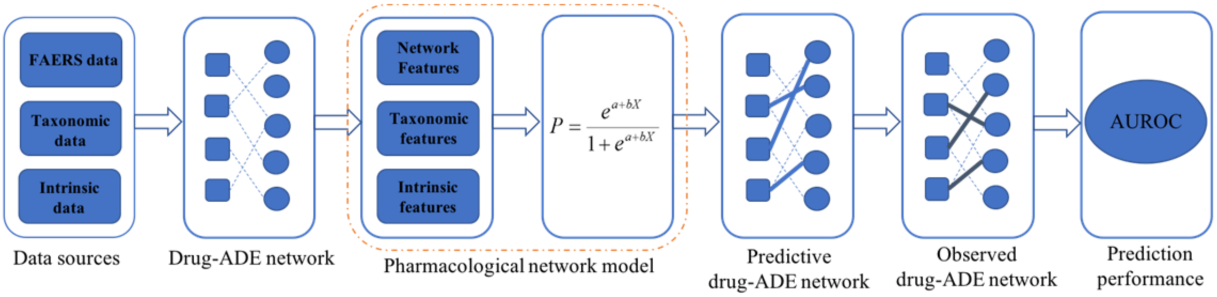


FIGURE 2. Overview of PNM. First, three types of data are integrated, including FAERS, taxonomic, and intrinsic data. Second, the drug–ADE network is constructed based on the drug–ADE associations. Next, the network features, taxonomic features, and intrinsic features are generated based on the drug–ADE network. An LR model was then trained using the training set (FAERS 2010 data). Finally, the prediction performance of PNM was evaluated using the new drug–ADE associations in the validation dataset.

TABLE 1. Definitions of taxonomic and intrinsic features.

Feature Name	Feature Definition	Supplementary Information
Taxonomic atc-min atc-KL	$X_9(i,j) = \min_{k\in N(j)-\{i\}}\{d_{ATC}(i,k)\}$ $X_{10}(i,j)$: KL divergence between the distribution $D_{ATC}(i,j)$ of the variable $D_{ATC}(i,k), k \in N(j)-\{i\}$ and its reference distribution.	The reference distribution was computed as the mean of distributions D_{ATC}(i, j) over the training edges (i, j).
meddra-min meddra-KL	$X_{11}(i,j) = \min_{k\in N(i)-\{j\}}\{d_{MedDRA}(j,k)\}$ $X_{12}(i,j)$: KL divergence between the distribution $D_{MedDRA}(i,j)$ of the variable $D_{MedDRA}(j,k), k \in N(i)-\{j\}$ and its reference distribution.	The reference distribution was computed as the mean of distributions D_{MedDRA}(i, j) over the training edges (i, j).
Intrinsic euclid-min euclid-KL	$X_{13}(i,j) = \min_{k\in N(j)-\{i\}}\{d_{INT}(i,k)\}$ X_{14}(i, j): KL divergence between the distribution $\mathrm{D}_{\mathrm{INT}}$(i, j) of the variable $D_{INT}(i,k), k \in N(j)-\{i\}$ and its reference distribution.	The reference distribution was computed as the mean of distributions D_{INT}(i, j) over the training edges (i, j).

(4) Hub Depressed Index (HDI) [37]: This index has the opposite effect compared with the hub promoted index, and is defined as

$$S_{ij} = \frac{|N(i) \cap N(j)|}{max\left(D_i, D_j\right)} \tag{8}$$

(5) Adamic–Adar (AA) index [38]: This measure considers that a smaller degree is more conducive to the connection with a common neighbor, and is defined as

$$S_{ij} = \sum\nolimits_{k\in N(i)\cap N(j)} \frac{1}{logD\,(k)} \tag{9}$$

(6) Resource Allocation (RA) index [37]: This index is similar to AA in form, and is mainly inspired by the resource allocation model in the network. It is expressed as

$$S_{ij} = \sum\nolimits_{k\in N(i)\cap N(j)} \frac{1}{D\,(k)} \tag{10}$$

All the above indices can be defined as features of the drug–ADE network. These features can be trained by a logistical regression model for the prediction of ADEs, and then compared with the features in the PNM.

D. FEATURES BASED ON THE MODIFIED ALGORITHM

In this section, we propose a modified algorithm named Jaccard and AA Fusion as the improved feature pairs for the ADE predictions using the drug–ADE network, which is a bipartite network consisted of known drug–ADE associations. The modified algorithm of the Jaccard and AA Fusion is given by

$$S\,(i,j) = \sum\nolimits_{k\in N(i)\cap N(j)} \left(\frac{\frac{|N(i)\cap N(j)|}{|N(i)\cup N(j)|}}{logD_k} + \frac{1}{logD_k} \right) \tag{11}$$

The improved features are the JADF and JAAF based on the modified algorithm. These are expressed as follows:

$$X_{JADF}\,(i,j) = \sum\nolimits_{k\in N(j)-\{i\}} \left(\frac{\frac{|N(i)\cap N(k)|}{|N(i)\cup N(k)|}}{logD_k} + \frac{1}{logD_k} \right) \tag{12}$$

$$X_{JAAF}\,(i,j) = \sum\nolimits_{k\in N(i)-\{j\}} \left(\frac{\frac{|N(j)\cap N(k)|}{|N(j)\cup N(k)|}}{logD_k} + \frac{1}{logD_k} \right) \tag{13}$$

JADF combines the strengths of both the Jaccard index and the AA index based on the drug pairs. This feature not only considers the structural similarity, but also takes advantage

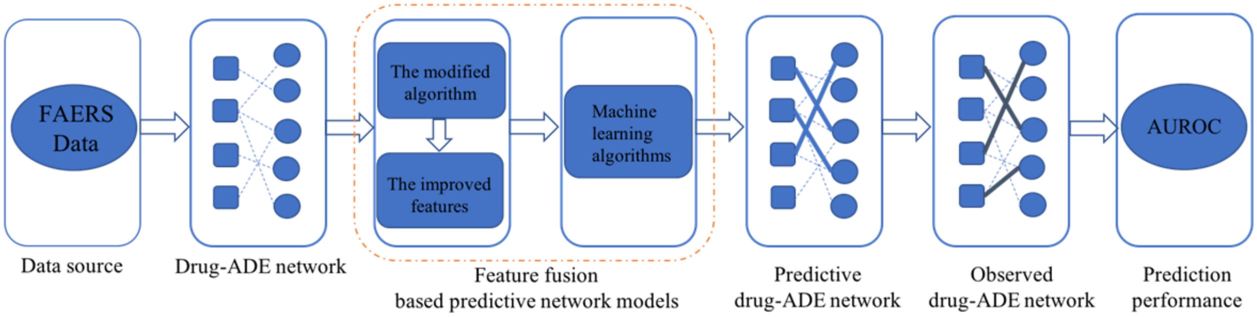

FAERS Data
Data source
Drug-ADE network
The modified algorithm
The improved features
Machine learning algorithms
Feature fusion based predictive network models
Predictive drug-ADE network
Observed drug-ADE network
AUROC
Prediction performance

FIGURE 3. Overview of FFPNMs. First, a drug–ADE network was constructed with the use of the drug–ADE associations from the FAERS data. Second, improved features were generated based on the drug–ADE network. Three different ML methods as classification algorithms were then employed with the use of the training set. Finally, its performance was evaluated with the use of the new drug–ADE associations in the validation set.

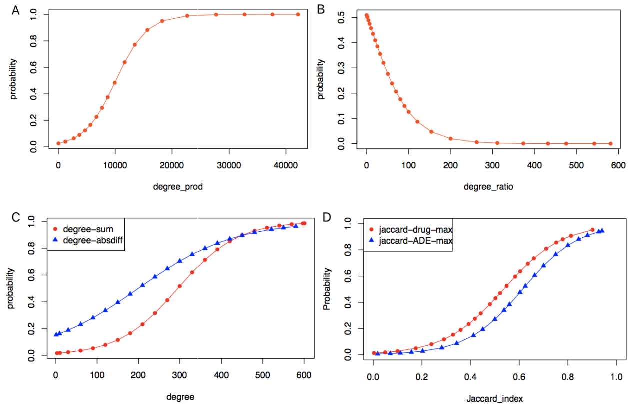


FIGURE 4. Illustration of selected feature effects. Probability of the existence of an association as a function of features degree-prod (A), degree-ratio (B), degree-sum (C), degree-absdiff (C), Jaccard-drug-max (D), and the Jaccard-ADE-max (D) in the PNM.

of the degrees of the nodes. JAAF is the abbreviation for the fusion of the Jaccard and AA–ADE, and is the same as JADF. It is defined based on the ADE pairs.

E. PREDICTION MODEL

We used the training data to train and build our prediction models named FFPNMs based on the defined features. Three different ML methods were employed as the classification algorithms, namely, LR, SVM and RF. Finally, an overview of FFPNMs is shown in Fig. 3.

In the LR model, the probability that the drug–ADE pair is true is given by

$$p_{ij} = exp\left(\sum\nolimits_s q_s x_s(i,j)\right) \Big/ \left[1 + exp\left(\sum\nolimits_s q_s x_s(i,j)\right)\right] \qquad (14)$$

Herein, q_s denotes the regression parameter and x_s denotes the features. The model fit by the training data is determined using the Akaike Information Criterion (AIC) through 10-fold cross validation, where optimal parameters are obtained through the optimal model that has the lowest AIC.

Once we have the fully trained model, the probability of each drug–ADE association in the validation set is predicted as follows:

$$pr_{ij} = 1 \Big/ \left[1 + exp\left(\sum\nolimits_s q_s x_s(i,j)\right)\right] \qquad (15)$$

FFPNM with SVM and FFPNM with RF were trained through 10-fold cross validation using the training set respectively, to obtain the optimal models with optimal parameters. Subsequently, the performance of the prediction models was evaluated using the validation set.

III. RESULTS

A. COMPARISONS OF PREDICTIVE RESULTS OF FEATURES IN PNM AND FEATURES BASED ON SIMILARITY MEASURES

The drug–ADE network is built with 152 drugs and 633 ADEs from the FAERS 2010 data. Correspondingly, 21065 new associations from FAERS 2011–2015 data are used as the validation set. Table 2 presents the analyzed results of the features in the PNM and the features based on the similarity measures, and lists the features trained by the LR model. AUROC values were determined on the validation data.

In the univariate analysis of the network features of PNM (Table 2), the degree-prod, which is also called PA, has the best performance (AUROC = 0.79). The multivariate analysis result of the combination of the Jaccard-drug-max and Jaccard-ADE-max is higher (AUROC = 0.825) than the multivariate analysis result of the combination of degree-prod, degree-sum, degree-ratio, and degree-absdiff (AUROC = 0.798). It should be noted that the univariate analysis results of the Jaccard-drug-max and Jaccard-ADE-max features are not high (AUROC = 0.732, AUROC = 0.704, respectively), but their combination yields a better predictive outcome. The univariate analysis outcomes of the Jaccard-drug-KL and Jaccard-ADE-KL indices are AUROC = 0.654 and AUROC = 0.551, respectively, and their multivariate analysis outcomes are AUROC = 0.621. Finally, the result of the network model is 0.824. When the Jaccard-drug-KL and Jaccard-ADE-KL are eliminated, the performance of the network model is AUROC = 0.826, which

TABLE 2. Comparisons of predictive results of different features based on AUROC.

Predictive results of PNM		Predictive results of similarity measures		
Feature Name	AUROC	Feature Name	Feature Definition	AUROC
$X_1(i,j)$	0.79	Salton-drug-max	$S_{sdm} = \max_{k\in N(j)-\{i\}} \frac{\lvert N(i)\cap N(k)\rvert}{\sqrt{D(i)D(k)}}$	0.722
$X_2(i,j)$	0.737			
$X_3(i,j)$	0.524	Salton-ADE-max	$S_{sam} = \max_{k\in N(i)-\{j\}} \frac{\lvert N(j)\cap N(k)\rvert}{\sqrt{D(j)D(k)}}$	0.616
$X_4(i,j)$	0.649			
X_1.2.3.4	**0.798**	**Salton-drug-ADE-max**		**0.815**
$X_5(i,j)$	0.732			
$X_6(i,j)$	0.704	HPI-drug	$S_{hpid} = \frac{\lvert N(i)\cap N(k)\rvert}{\min(D(i),D(k))}$ $k \in N(j)-\{i\}$	0.6
X_5.6	**0.825**			
		HPI-ADE	$S_{hpia} = \frac{\lvert N(j)\cap N(k)\rvert}{\min(D(j),D(k))}$ $k \in N(i)-\{j\}$	0.546
X_1.2.3.4.5.6	**0.826**			
$X_7(i,j)$	0.654	**HPI-drug-ADE**		**0.605**
$X_8(i,j)$	0.551			
X_7.8	**0.621**	HDI-drug	$S_{hdid} = \frac{\lvert N(i)\cap N(k)\rvert}{\max(D(i),D(k))}$ $k \in N(j)-\{i\}$	0.549
NETWORK	**0.824**			
		HDI-ADE	$S_{hdia} = \frac{\lvert N(j)\cap N(k)\rvert}{\max(D(j),D(k))}$ $k \in N(i)-\{j\}$	0.551
$X_9(i,j)$	0.628			
$X_{10}(i,j)$	0.639			
$X_{11}(i,j)$	0.643	**HDI-drug-ADE**		**0.544**
$X_{12}(i,j)$	0.609			
TAXONOMIC	**0.667**	AA-drug	$S_{aad} = \sum_{k\in(N(j)-i)} \frac{1}{\log D(k)}$	0.692
		AA-ADE	$S_{aaa} = \sum_{k\in(N(i)-j)} \frac{1}{\log D(k)}$	0.675
$X_{13}(i,j)$	0.602			
$X_{14}(i,j)$	0.579	**AA-drug-ADE**		**0.821**
INTRINSIC	**0.601**			
		RA-drug	$S_{rad} = \sum_{k\in(N(j)-i)} \frac{1}{D(k)}$	0.689
TAX+INT	**0.707**	RA-ADE	$S_{raa} = \sum_{k\in(N(i)-j)} \frac{1}{D(k)}$	0.675
NET+TAX+INT	**0.824**	**RA-drug-ADE**		**0.799**

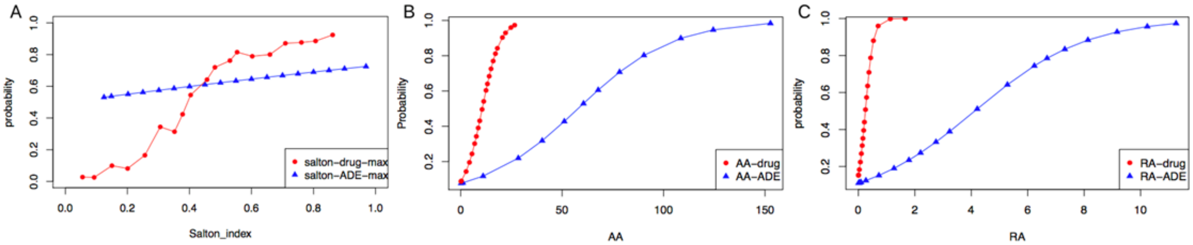


FIGURE 5. Illustration of the effects of selected similar measures. Probability of the existence of associations as functions of Salton-drug-max and Salton-ADE-max (A), AA-drug and AA-ADE (B), and RA-drug and RA-ADE (C).

is the highest. Moreover, the performance of the combination of all network features does not improve the ADE predictions in the network model, and the best multivariate fitting is the combination of degree-prod, degree-sum, degree-ratio, degree-absdiff, Jaccard-drug-max and Jaccard-ADE-max. Their statistics are listed in Table 3, and the illustration of their effects is presented in Fig. 4.

In the univariate analysis of taxonomic features and intrinsic features in the PNM, the outcomes of each of the features were too low, and the multivariate results were higher than their univariate analysis results. Compared with the network model (AUROC = 0.824), the performances of the taxonomic and intrinsic model were much lower (AUROC = 0.707). The performance of the combination of the three types of features yield no improvement (AUROC = 0.824), and it is still smaller than the result yielded by the combination of some features in the network model (AUROC = 0.825 and 0.826). Above all, the prediction of ADEs could be measured using only network features, and yields a superior performance compared to the prediction of the effects of the pharmacological features of PNM.

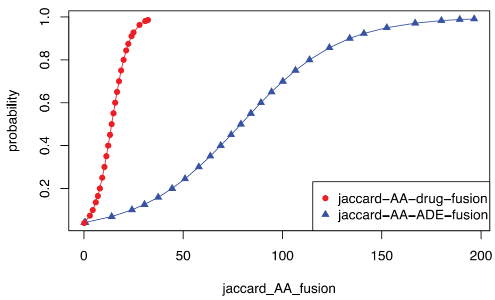


FIGURE 6. Illustration of the improved feature effects. Probability of existence of associations as functions of JADF and JAAF in the FFPNM.

Meanwhile, the predictive results of the local-information-based similarity measures are also presented in Table 2. The univariate analysis results of the Salton-drug-max and

TABLE 3. Multivariate model feature outcomes of PNM.

Feature Name	Regression Coefficients	Standard Error	P-value
degree-prod	6.735e-05	6.387e-06	< 2e-16
degree-sum	3.397e-03	9.873e-04	0.000581
degree-ratio	2.581e-03	1.042e-03	0.289829
degree-absdiff	-1.102e-03	6.088e-04	0.50278
Jaccard-drug-max	9.851e+00	1.976e-01	< 2e-16
Jaccard-ADE-max	8.073e+00	3.111e-01	< 2e-16

TABLE 4. Results of univariate and multivariate analyses of similarity measures.

Feature Name	Regression Coefficients	Standard Error	P-value
Salton-drug-max	9.93724	0.07481	<2e-16
Salton-ADE-max	11.3827	0.08493	<2e-16
Salton-drug-ADE-max	13.1183	0.1179	<2e-16
	15.5852	0.1417	<2e-16
AA-drug	0.224530	0.001745	<2e-16
AA-ADE	0.0431350	0.0003087	<2e-16
AA-drug-ADE	0.399553	0.003483	<2e-16
	0.072575	0.000606	<2e-16
RA-drug	7.00597	0.06255	<2e-16
RA-ADE	0.506379	0.003761	<2e-16
RA-drug-ADE	10.037796	0.092993	<2 e-16
	0.685885	0.005464	<2e-16

Salton-ADE-max are AUROC = 0.722 and 0.616, respectively. The multivariate analysis outcome yields an AUROC of 0.815, which is slightly smaller than the Jaccard index outcome (AUROC = 0.825). The results of univariate and multivariate analyses of HDI and HPI are all very low, with a maximum AUROC of 0.605. Moreover, the multivariate analysis outcome of AA (AUROC = 0.821) is better than that for RA (AUROC = 0.799). The statistics of univariate and multivariate analyses of the selected similarity measures are detailed in Table 4, and the illustration of the selected similarity measure effects are shown in Fig. 5.

B. PREDICTIVE RESULTS OF JADF AND JAAF IN FFPNM WITH LR

As shown in Table 2, the highest predictive values are those associated with the Jaccard index and AA index, which are defined based on similarity measures with the use of local information. AA index considers the degree of common nodes, and the Jaccard index measures the similarity between the neighborhoods of two nodes. The larger the index value is, the more similar the two neighborhoods are. The proposed improved features based on the modified algorithm not only consider the structural similarity, but also take advantage of the degrees of the nodes. The illustration of the improved features effects is shown in Fig. 6. The univariate and multivariate analysis results of FFPNM first, whereby the univariate analysis results of JADF and JAAF are AUROC = 0.757 and 0.732, respectively, are presented in Table 5. The multivariate analysis outcome of their combination is AUROC = 0.849 (21065 new drug-ADE combinations). This is the highest value in all univariate and multivariate analyses among the three tested methods that used FAERS data from 2010 to 2015.

To evaluate the robustness and performance of the improved features and predict ADEs accurately and earlier, the first 120 days of FAERS 2004 data were used as the training set. These included 10307 drug–ADE associations between 1177 drugs and 97 ADEs. Correspondingly, 22358 new drug–ADE associations from FAERS 2004 to 2009 (not from the first 120 days of data) constituted the validation set. The analysis results of univariate model, and multivariate models with two different drug-ADE networks are listed in Table 5. The comparisons of predicted

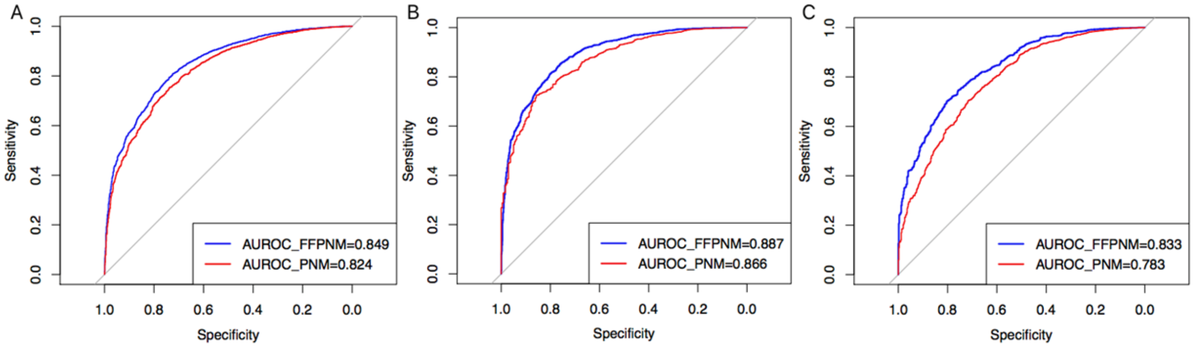


FIGURE 7. **(A) Comparison of performances of FFPNM and PNM for FAERS 2010 drug–ADE network. (B) Comparison of performances of FFPNM and PNM for drug-ADE network of FAERS data of the first 120 days of 2004. (C) Comparison of performances of FFPNM and PNM with the intersection of FAERS and SIDER as the validation set.**

TABLE 5. **Analysis results of univariate model multivariate models in the FFPNM.**

Network	Feature Name	AUROC	Regression Coefficients	Standard Error	P-value
FAERS 2010 data including 152 drugs and 633ADEs	JADF	0.757	0.23185	0.00166	<2e-16
	JAAF	0.732	0.04008	0.00027	<2e-16
	JADF+JAAF	**0.849**	0.26869	0.05654	<2e-16
			0.04674	0.00255	<2e-16
The first 120 days data of FAERS 2004 including 1177 drugs and 97 ADEs	JADF	0.788	0.04063	0.00066	<2e-16
	JAAF	0.809	0.78021	0.012	<2e-16
	JADF+JAAF	**0.887**	0.04966	0.00108	<2e-16
			0.89502	0.01679	<2e-16

TABLE 6. **Predictive results of FFPNMs with SVM, RF, and LR.**

Network	Model	Accuracy	PPV	AUROC	Optimal parameters
FAERS 2010 data including 152 drugs and 633ADEs	SVM	0.853	0.652	0.851	Cost:10, gamma: 10 Estimation error: 0.09056
	RF	0.862	0.667	0.856	Number of trees: 462 Estimation error: 0.09334
	LR	0.851	0.566	0.849	As shown in the fourth row of table 5
The first 120 days data of FAERS 2004 including 1177 drugs and 97 ADEs	SVM	0.907	0.722	0.902	Cost:10, gamma:10 Estimation error: 0.03686
	RF	0.923	0.756	0.913	Number of trees: 482 Estimation error: 0.03782
	LR	0.891	0.675	0.887	As shown in the last row of table 5

outcomes are presented in Fig. 7. Accordingly, AUROC = 0.824 and AUROC = 0.849 are the predicted outcomes of PNM and FFPNM, respectively, in which the network consists of 152 drugs and 633 ADEs from FAERS 2010 data. And, AUROC = 0.887 is the predictive result of the FFPNM compared with the PNM outcomes that yielded a corresponding value of 0.866, whose network consisted of 1177 drugs and 97 ADEs based on the first 120 days of FAERS 2004 data.

Moreover, 1148 drug-ADE associations are in the intersection of FAERS and SIDER based on 22358 new drug-ADE associations. From Fig. 7(C), it is evident that FFPNM achieves an AUROC value of 0.833 compared with an AUROC value of 0.783 for PNM based on the intersection of FAERS and SIDER, which is used as the validation set. In conclusion, the proposed FFPNM with LR has superior performance and robustness for ADE predictions, regardless of the number of drugs and ADEs in the network, and regardless of the time period.

C. PREDICTIVE RESULTS OF JADF AND JAAF IN FFPNMs WITH ML METHODS SVM AND RF

FFPNM with LR had the superior predictive performance. We further investigated the performance of FFPNMs, in which SVM and RF are as the classification algorithms, respectively. Herein, prediction models were trained by FAERS 2010 data and the first 120 days data of FAERS 2004, respectively.

The predictive results of SVM and RF in FFPNMs are listed in Table 6. Accordingly, SVM and RF provide better predicted results than LR in FFPNM with two different drug-ADE networks. In general, FFPNM with RF turns out to be the best model among them with an AUROC value of 0.856 and an AUROC value of 0.913. The performance of FFPNM with SVM is close to that of FFPNM with RF and higher than that of FFPNM with LR. Furthermore, accuracy and the positive predictive values (PPV) of FFPNM with RF are higher than that of FFPNM with SVM and FFPNM with LR. The values highlight the superior capability of the ensemble model (RF) over the individual models (i.e., SVM and LR), and the proposed FFPNM with SVM and FFPNM with RF have further improved ADE predictions in comparison with that of FFPNM with LR.

IV. CONCLUSION

In this study, we proposed the simple and efficient FFPNMs, in which different ML methods were used as the classification algorithms, to predict ADEs. The FFPNMs used the modified algorithms defined based on the similarity measures as the improved features, which only extracted information from drug-ADE network. The improved features combined the strengths of the Jaccard and AA indices and improved the predictions of ADEs. Our findings provided the crucial information about the influences of different structural features and combinations of features on prediction, and combined with ML methods, to achieve an optimal, concise, and efficient model as the computational method for drug-ADE association predictions.

We investigated the PNM proposed by Cami for comparison. With the exception of the degree-prod, which was derived from similarity measures of complex networks, degree-sum and degree-ratio were generated for completeness from the perspective of PNM. As illustrated, the degree-ratio effect was opposite compared with the other three degree features, while the univariate analysis result (AUROC = 0.524) was the lowest among the four degree features. Thus, in the definition of features, appropriateness should be greater than completeness for a better prediction of the performance. Furthermore, the Jaccard-drug-KL and Jaccard-ADE-KL indices, used as extensions of the network structural features, yielded poor improvements. Accordingly, the contractor prediction performance of the network model based on these indices became worse. It suggested that the performances of the Jaccard-drug-KL and Jaccard-ADE-KL indices may depend on the data, and inappropriate data will have the opposite effect. Moreover, taxonomic and intrinsic features consider the attributes of drugs and ADEs on the basis of the network structure and introduce pharmacological implications to the network model, which mainly make the model more comprehensive, but also more complex to apply. Regardless of univariate or multivariate analyses, the predictive effects of these features are poor, and the AUROC values are approximately equal to 0.6. These two types of features do not play important roles, and only a slight improvement was documented from the perspective of optimization. Finally, the best performance features were based on the similarity measure and PNM only presented a network-based model and did not develop an optimal model.

The predicted outcomes that the five different similarity measures extracted as features indicate that AA and Salton indices yield good performances compared with other similarity measures for ADE predictions. Based on the analysis of PNM and the similarity measures, we extracted the features of Jaccard and Adamic-Adar indices to build the FFPNMs with ML algorithms. The FFPNM with LR generated more concise and efficient predictions, achieved superior prediction performance compared with PNM [10], and were robust. Especially, FFPNM with RF presented the best prediction performance, and the result of FFPNM with SVM is close to RF and higher than LR. We believe that the superior result is simply due to the power of ensemble ML algorithms, which can capture the sophisticated patterns in data.

Our future work will be extended in the following directions: firstly, features with improved performances will be considered as the supplementary features from the attributes of drugs and ADEs for ADEs prediction. In addition, future research can extend our approach by developing features defined based on similarity measures of the path information of complex network, in an effort to achieve better predictions following the considerations of the weight information. Moreover, both PNM and FFPNM with LR can generate probabilities for any drug–ADE association. However, these probabilities have different interpretations for the drug–ADE frequencies estimated using spontaneous reporting system databases. To address this challenging issue, novel integrated methods need to be developed in the future.

REFERENCES

[1] F. T. Bourgeois, M. W. Shannon, C. Valim, and K. D. Mandl, "Adverse drug events in the outpatient setting: An 11-year national analysis," *Pharmacoepidemiology Drug Saf.*, vol. 19, no. 9, pp. 901–910, Sep. 2010.

[2] J. Lazarou, B. H. Pomeranz, and P. N. Corey, "Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies," *Surv. Anesthesiology*, vol. 43, no. 1, pp. 53–54, Feb. 1999.

[3] S. J. W. Evans, P. C. Waller, and S. Davis, "Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports," *Pharmacoepidemiology Drug Saf.*, vol. 10, no. 6, pp. 483–486, Oct. 2001.

[4] E. P. van Puijenbroek, A. Bate, H. G. M. Leufkens, M. Lindquist, R. Orre, and A. C. G. Egberts, "A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions," *Pharmacoepidemiology Drug Saf.*, vol. 11, no. 1, pp. 3–10, Jan. 2002.

[5] A. Bate, M. Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner, and R. M. De Freitas, "A Bayesian neural network method for adverse drug reaction signal generation," *Eur. J. Clin. Pharmacol.*, vol. 54, no. 4, pp. 315–321, Jul. 1998.

[6] W. DuMouchel, "[Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting System]: Reply," *Amer. Statistician*, vol. 53, no. 3, p. 201, Aug. 1999.

[7] P. Zhang, M. Li, C.-W. Chiang, L. Wang, Y. Xiang, L. Cheng, W. Feng, T. K. Schleyer, S. K. Quinney, H.-Y. Wu, D. Zeng, and L. Li, "Three-component mixture model-based adverse drug event signal detection for the adverse event reporting system," *CPT: Pharmacometrics Syst. Pharmacol.*, vol. 7, no. 8, pp. 499–506, Aug. 2018.

[8] N. Atias and R. Sharan, "An algorithmic framework for predicting side effects of drugs," *J. Comput. Biol.*, vol. 18, no. 3, pp. 207–218, Mar. 2011.

[9] A. Cami, A. Arnold, S. Manzi, and B. Reis, "Predicting adverse drug events using pharmacological network models," *Sci. Transl. Med.*, vol. 3, no. 114, Dec. 2011, Art. no. 114ra127.

[10] L.-C. Huang, X. Wu, and J. Y. Chen, "Predicting adverse side effects of drugs," *BMC Genomics*, vol. 12, no. Suppl 5, p. S11, Dec. 2011.

[11] M. Liu, Y. Wu, Y. Chen, J. Sun, Z. Zhao, X.-W. Chen, M. E. Matheny, and H. Xu, "Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs," *J. Amer. Med. Inform. Assoc.*, vol. 19, no. e1, pp. e28–e35, Jun. 2012.

[12] N. P. Tatonetti, P. P. Ye, R. Daneshjou, and R. B. Altman, "Data-driven prediction of drug effects and interactions," *Sci. Transl. Med.*, vol. 4, no. 125, Mar. 2012, Art. no. 125ra31.

[13] F. Cheng, W. Li, X. Wang, Y. Zhou, Z. Wu, J. Shen, and Y. Tang, "Adverse drug events: Database construction and in silico prediction," *J. Chem. Inf. Model.*, vol. 53, no. 4, pp. 744–752, Apr. 2013.

[14] M. Duran-Frigola and P. Aloy, "Analysis of chemical and biological features yields mechanistic insights into drug side effects," *Chem. Biol.*, vol. 20, no. 4, pp. 594–603, Apr. 2013.

[15] J. Lin, Q. Kuang, Y. Li, Y. Zhang, J. Sun, Z. Ding, and M. Li, "Prediction of adverse drug reactions by a network based external link prediction method," *Anal. Methods*, vol. 5, no. 21, p. 6120, 2013.

[16] X. Ji, L. Wang, L. Hua, X. Wang, P. Zhang, A. Shendre, W. Feng, J. Li, and L. Li, "Improved adverse drug event prediction through information component guided pharmacological network model (IC-PNM)," in *Proc. IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published.

[17] Z. Wang, J. Liu, Y. Yu, Y. Chen, and Y. Wang, "Modular pharmacology: The next paradigm in drug discovery," *Expert Opinion Drug Discovery*, vol. 7, no. 8, pp. 667–677, Aug. 2012.

[18] F. Azuaje, "Drug interaction networks: An introduction to translational and clinical applications," *Cardiovascular Res.*, vol. 97, no. 4, pp. 631–641, Mar. 2013.

[19] C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi, "From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks," *Sci. Rep.*, vol. 3, no. 1, pp. 1613–1626, Dec. 2013.

[20] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: A network-based approach to human disease," *Nature Rev. Genet.*, vol. 12, no. 1, pp. 56–68, Jan. 2011.

[21] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, and Y. Tang, "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Comput. Biol.*, vol. 8, no. 5, 2012, Art. no. e1002503.

[22] Y. Luo, Q. Liu, W. Wu, F. Li, and X. Bo, "Predicting drug side effects based on link prediction in bipartite network," in *Proc. 7th Int. Conf. Biomed. Eng. Informat.*, Oct. 2014, pp. 729–733.

[23] H. Chen, H. Zhang, Z. Zhang, Y. Cao, and W. Tang, "Network-based inference methods for drug repositioning," *Comput. Math. Methods Med.*, vol. 2015, no. 2, pp. 1–7, Feb. 2015.

[24] B. Kaya and M. Poyraz, "Finding relations between diseases by age-series based supervised link prediction," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2015, pp. 1097–1103.

[25] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabasi, "Uncovering disease-disease relationships through the incomplete interactome," *Science*, vol. 347, no. 6224, Feb. 2015, Art. no. 1257601.

[26] B. Davazdahemami and D. Delen, "A chronological pharmacovigilance network analytics approach for predicting adverse drug events," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1311–1321, Oct. 2018.

[27] (2018). *US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS)*. [Online]. Available: https://www.fda.gov/Drugs/InformationOnDrugs/ucm135151.htm

[28] (2018). *Medical Dictionary for Regulatory Activities (MedDRA)*. [Online]. Available: https://www.meddra.org

[29] (2018). *DrugBank*. [Online]. Available: https://www.drugbank.ca

[30] (2019). *Side Effects Resource*. [Online]. Available: http://sideeffects.embl.de/

[31] S. A. Barabá, "Emergence of Scaling in RandomNetworks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[32] P. Jaccard, "Distribution de la flore alpine dans le bassin des Dranses et dans quelques regions voisines," *Bull. de la Societe Vaudoise des Sci. Naturelles*, vol. 37, no. 142, pp. 547–579, 1901.

[33] (2018). *PubChem*. [Online]. Available: https://pubchem.ncbi.nlm.nih.gov

[34] F. Lorrain and H. C. White, "Structural equivalence of individuals in social networks," *J. Math. Sociology*, vol. 1, no. 1, pp. 49–80, Jan. 1971.

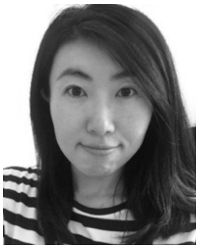
[35] G. Salton and M.J. Mcgill. *Introduction to Modern Information Retrieval*. Auckland, New Zealand: MuGraw-Hill, 1983.

[36] E. Ravasz, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, Aug. 2002.

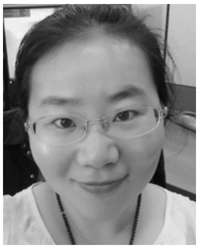
[37] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *Eur. Phys. J. B*, vol. 71, no. 4, pp. 623–630, Oct. 2009.

[38] L. A. Adamic and E. Adar, "Friends and neighbors on the Web," *Social Netw.*, vol. 25, no. 3, pp. 211–230, Jul. 2003.

JIN LI received the B.S. degree in computer science and technology, the M.S. degree in computer application, and the Ph.D. degree in control theory and control engineering from Harbin Engineering University, China, in 1984, 1991, and 2001, respectively. She was a Postdoctoral Fellow with the Harbin Institute of Technology, from 2002 to 2004. She worked as a Senior Visiting Researcher with the Kitami Institute of Technology, Japan, from December 1999 to December 2000, and with the Hong Kong Polytechnic University, from December 2001 to February 2002, January 2003 to March 2003, respectively. She is currently a Professor with the Automation College, Harbin Engineering University. Her research interests include digital image processing and bioinformatics. She has authored or coauthored more than 100 peer-reviewed publications.



XIANGMIN JI received the M.S. degree from Harbin Engineering University, Harbin, China, in 2009, where she is currently pursuing the Ph.D. degree in control science and engineering. Her current research interests include bioinformatics, pharmacovigilance, data mining, and medical records analysis.



LIYAN HUA received the M.S. degree from Harbin Normal University, Harbin, China, in 2012. She is currently pursuing the Ph.D. degree in control science and engineering with Harbin Engineering University, Harbin. Her current research interests include bioinformatics, data mining, and drug metabolic pathway analysis.