TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 05/18 pp58–67
DOI: 10.26599/TST.2021.9010003
Volume 27, Number 1, February 2022

Metabolite-Disease Association Prediction Algorithm Combining DeepWalk and Random Forest

Jiaojiao Tie, Xiujuan Lei*, and Yi Pan*

Abstract: Identifying the association between metabolites and diseases will help us understand the pathogenesis of diseases, which has great significance in diagnosing and treating diseases. However, traditional biometric methods are time consuming and expensive. Accordingly, we propose a new metabolite-disease association prediction algorithm based on DeepWalk and random forest (DWRF), which consists of the following key steps: First, the semantic similarity and information entropy similarity of diseases are integrated as the final disease similarity. Similarly, molecular fingerprint similarity and information entropy similarity of metabolites are integrated as the final metabolite similarity. Then, DeepWalk is used to extract metabolite features based on the network of metabolite-gene associations. Finally, a random forest algorithm is employed to infer metabolite-disease associations. The experimental results show that DWRF has good performances in terms of the area under the curve value, leave-one-out cross-validation, and five-fold cross-validation. Case studies also indicate that DWRF has a reliable performance in metabolite-disease association prediction.

Key words: DeepWalk; random forest; metabolite-disease associations; molecular fingerprint similarity of metabolites

1 Introduction

Metabolism is the power source that drives all life activities of organisms[1]. Substances produced or consumed during metabolism are called metabolites. The levels of metabolites can directly reflect the physiological state of the body, and sufficient evidence shows that disease is always accompanied with changes in metabolites[2]. Therefore, the recognition of abnormal and disease-related metabolites is of great significance not only to improve the level of clinical diagnosis, but also to better understand the pathological metabolic process.

Over the years, many biologists have obtained considerable achievements in the study of metabolites[3–6]. Moats et al.[7] used quantitative 1H magnetic resonance spectroscopy to analyze 10 patients with Alzheimers disease (AD) and seven normal elderly patients, and the findings showed abnormal metabolite concentrations in the patients with AD. Unschuld et al.[8] used high-field-intensity MRS technology to identify the relationship between brain metabolites and cognitive function within the 7-Tesla range in patients with Huntingtons disease. Hori et al.[9] used gas chromatography-mass spectrometry for the metabolomics analysis of patients with lung cancer (LC). Cheng et al.[10] described changes in the lipid metabolism that regulate metabolic diseases, such as nonalcoholic fatty liver disease, obesity, and cancer, and suggested that natural compounds may provide potential therapeutic agents for the treatment or prevention of metabolic disorders with abnormal lipid metabolism. To organize metabolite data more intuitively, some metabolite databases were created[2, 11, 12]. The traditional biological methods are time consuming and labor intensive. Therefore, it is necessary to develop an effective computational method

• Jiaojiao Tie and Xiujuan Lei are with the School of Computer Science, Shaanxi Normal University, Xi'an 710119, China. E-mail: xjlei@snnu.edu.cn.
• Yi Pan is with the the Department of Computer Science, Georgia State University, Atlanta, GA 30302-3994, USA. E-mail: yipan@gsu.edu.
∗ To whom correspondence should be addressed.
Manuscript received: 2020-12-21; accepted: 2021-01-13

to predict metabolite-disease associations. Based on the functional relationship between metabolites in metabolic pathways, Shang et al.[13] proposed a path-based random walk method to identify candidate metabolites for diseases. Metabolites with similar functions are often associated with the same or similar diseases. Yang et al.[14] used a random walk method to predict disease-related metabolites based on the similarity of metabolites. Considering a large number of metabolic markers in diseases, Wang et al.[15] used the collaborative filtering strategy to construct a reliable metabolic network based on the literature scores and functional similarity of metabolites to further predict the relationship between metabolites and diseases. These methods are based on disease and metabolite similarities and use common network computing methods to predict the metabolite-disease associations. Moreover, only metabolite-disease association data were used instead of their respective topological features. Intuitively, metabolites do not exist independently in the human body, and other behaviors of life activities will also lead to abnormal metabolites, which will lead to the occurrence of diseases. Thus, new metabolite features or disease features can be obtained by combining other omics data. In recent years, machine learning methods have been widely used in computational biology. In this article, we introduce the relationship between metabolites and diseases and use DeepWalk (DW) method to extract new metabolite features. The random forest (RF) algorithm has been widely applied in bioinformatics and has achieved good results. Qi[16] demonstrated that the RF has unique advantages in dealing with small sample sizes, high-dimensional feature spaces, and complex data structures, and its application in computational biology is increasing. Chen et al.[17] found that the RF has a good performance in dealing with unbalanced problems.

In this article, we propose a novel method called DWRF (combination of DW and RF) to identify potential metabolite-disease associations. First, we calculate disease semantic similarity, molecular fingerprint similarity of metabolites, and the information entropy similarity of metabolites and diseases, and integrate the similarity of diseases and metabolites. Second, we extract metabolite features from a metabolite-gene network using the DW method. Finally, the RF algorithm is used to predict disease-related metabolites. The results of our evaluation show that DWRF has a good performance in metabolite-disease association prediction.

2 Material and Method

2.1 Data

The Human Metabolome Database (HMDB) records detailed information about small molecule metabolites found in the human body[12]. By removing redundant and missing data downloaded from the HMBD, we extract 3460 metabolite-disease associations, including 1478 metabolites and 237 diseases. In the HMDB, genes associated with metabolites were recorded. We also extract the association between metabolites and genes, including 4903 genes, 1478 metabolites, and 67295 metabolite-gene associations. The adjacency matrix $Y_{nm\times nd}$ can be utilized to describe the associations between metabolites and diseases, where nm and nd indicate the number of metabolites and diseases, respectively. If metabolite m_i is related to disease d_j, then $Y(i, j)$ is equal to 1; otherwise 0.

2.2 Disease semantic similarity

According to the Medical Subject Headings descriptors of a disease[18], the topology of each disease is visualized as a directed acyclic graph (DAG), in which the nodes represent the disease terms and edges represent the links from the parent disease term nodes to the child disease term nodes. Let $DAG(d) = (d, T(d), E(d))$, where d indicates disease d, $T(d)$ indicates the set of diseases that includes disease d and the ancestors of disease d, and $E(d)$ indicates the set of edges. The semantic contribution of disease t in DAG(d) to disease d can be calculated as follows:

$$D_d(t) = \begin{cases} 1, & \text{if } t = d; \\ \max\{\Delta \times D_d(t) \mid t \in \text{children of } t\}, & \text{if } t \neq d \end{cases} \tag{1}$$

where the disease $t \in T(d)$ and Δ is the semantic contribution decay factor and we set $\Delta = 0.5$[19].

The semantic similarity between d_i and d_j can be calculated as follows:

$$DS(d_i, d_j) = \frac{\sum_{t \in T(d_i) \cap T(d_j)} \left(D_{d_i}(t) + D_{d_j}(t)\right)}{\sum_{t \in T(d_i)} D_{d_i}(t) + \sum_{t \in T(d_j)} D_{d_j}(t)} \tag{2}$$

where $T(d_i) \cap T(d_j)$ indicates the common diseases between disease d_i and d_j and DS represents the disease semantic similarity with dimension $nd \times nd$.

2.3 Molecular fingerprint similarity of metabolites

The HMDB database records the molecular structure of the metabolites, so we calculate the similarity of metabolites by converting the molecular structure of the

metabolite into a series of binary fingerprint sequences. In this study, the Morgan fingerprints are used to measure the similarities between metabolites. The Morgan fingerprints are round fingerprints obtained by modifying the standard Morgan algorithm. Each element in the fingerprint represents a specific substructure that can be easily analyzed and interpreted and used for similarity calculations. The molecular fingerprint similarity of metabolites m_i and m_j is calculated as follows:

$$MS(m_i, m_j) = \frac{c}{a + b - c} \tag{3}$$

where MS represents the molecular fingerprint similarity of metabolites, a and b indicate the number of 1 in the molecular fingerprint of metabolite m_i and metabolite m_j, respectively, and c represents the number of 1 in the fingerprint sequences of metabolite m_i and m_j.

2.4 Calculation of metabolites and disease similarity based on information entropy

In a previous study, information entropy and mutual information were used to calculate the miRNA similarity[20]. In the present study, we use information entropy and mutual information to calculate the metabolite similarity and disease similarity based on metabolite-disease associations. The disease sets of metabolites A and B are $T_m^A = \left\{T_m^A(1), T_m^A(2), \ldots, T_m^A(n_{ma})\right\}$ and $T_m^B = \left\{T_m^B(1), T_m^B(2), \ldots, T_m^B(n_{mb})\right\}$, where n_{ma} and n_{mb} represent the number of diseases associated with metabolites A and B, respectively. The information entropy of metabolite A can be calculated as

$$H(T_m^A) = -\sum_{i=1}^{n_{ma}} p(T_m^A(i)) \log_2(p(T_m^A(i))) \tag{4}$$

where $p(T_m^A(i)) = n(T_m^A(i))/N$ represents the ratio of i-th disease associated with metabolite A in the disease set of metabolite A, N is the total number of known metabolite-disease interactions, and $n(T_m^A(i))$ is the number of known associations between the i-th disease and all metabolites in the disease set associated with metabolite A.

The similarity of metabolite A and metabolite B can be calculated by using the mutual information of them:

$$MI_m = \frac{2 \times H(T_m^A \cap T_m^B)}{H(T_m^A) + H(T_m^B)} \tag{5}$$

where $H(T_m^A \cap T_m^B)$ represents the information entropy of the common disease between metabolites A and B. The higher the probability of the metabolite-related disease, the less information it carries, and vice versa.

Similarly, the similarity of diseases can be obtained through information entropy and mutual information between diseases, which is represented by MI_d.

2.5 Integrated similarity for diseases and metabolites

Based on disease terms, we obtain the disease semantic similarity, which has many empty values. Hence, we use the disease similarity based on the information entropy to supplement the empty value in the disease semantic similarity matrix DS. The integrated disease similarity of diseases i and j can be calculated as

$$S^d(i, j) = \begin{cases} MI_d(i, j), & DS(i, j) = 0; \\ DS(i, j), & DS(i, j) \neq 0 \end{cases} \tag{6}$$

Similarly, the integrated metabolite similarity can be calculated as

$$S^m(i, j) = \alpha MS(i, j) + (1 - \alpha) MI_{\mathrm{m}} \tag{7}$$

where MI_m represents the information entropy similarity of metabolite and α is used to balance the molecular fingerprint similarity of metabolites and metabolite similarity based on information entropy. In this study, we set $\alpha = 0.5$, which means that the two similarities are equally important.

2.6 DWRF prediction method

2.6.1 Feature extraction of metabolites based on the DW method

In view of the sparsity of network representation learning, DW is proposed to learn the social representation of graph vertices[21]. To extract metabolite-gene association data, we establish the association network between them and use the adjacency matrix MG to represent the metabolite-gene network. In this paper, DW is used to extract the feature of metabolites from MG. Chris et al.[22] integrated some deep learning algorithms into a library called deeplearning4j library, from which we could obtain the DW algorithm. Subsequently, the extracted metabolite features are represented by S^{NE}.

The main idea of DW is to use the co-occurrence relation between nodes in the graph to learn the vector representation of nodes, which can be divided into two main steps: (1) The random walk algorithm is used to sample the nodes in the graph. (2) The skip-gram algorithm is used to learn the embedding of each node based on the generated node sequence.

Suppose W_{vi} is the result of a random walk starting with vertex v_i, and then traverse the random walk vertex sequence W_{vi} by using the sliding window, and the size of the window is w. In each window, the representation

of the current vertex $\Phi(v_i)$ is updated by maximizing the possibility of the vertex appearing in the window, so as to learn the representation of each vertex Φ. The skip-gram algorithm uses the independence assumption, and the conditional probabilities are approximated as follows:

$$\Pr\left(\{v_i - w, \ldots, v_i + w\} / v_i \mid \Phi(V_i)\right) = \prod_{j=i-w, j\neq w} \left(\Pr(v_j | \Phi(v_i))\right) \tag{8}$$

The above description can be represented by Fig.1.

2.6.2 Construction of feature vectors for metabolite-disease pairs

Three types feature vectors are utilized to describe the feature of metabolite-disease pairs: vectors based on metabolite similarity S^m, vectors based on extracted metabolite feature S^{NE} from metabolite-gene associations and vectors based on disease similarity S^d. Therefore, the feature vector of metabolite m_i and disease d_i can be described as follows:

$$F_{i,j} = (S^{NE}_{[i\times \mathrm{dim}]}, S^{m}_{[i\times nm]}, S^{d}_{[j\times nd]}) \tag{9}$$

where $F_{i,j}$ is the feature vector of metabolite m_i and disease d_i; dim indicates the dimension of S^{NE} (i.e., 128); nm and nd indicate the number of metabolites and diseases, respectively; and F indicates the feature matrix of all metabolite-disease pairs, whose dimension is $nm \times (dim + nm + nd)$. Then we normalize F to F^{final} as follows:

$$F^{final} = \frac{F - F_{min}}{F_{max} - F_{min}} \tag{10}$$

where F_{min} and F_{max} are the minimum and maximum values in F, respectively.

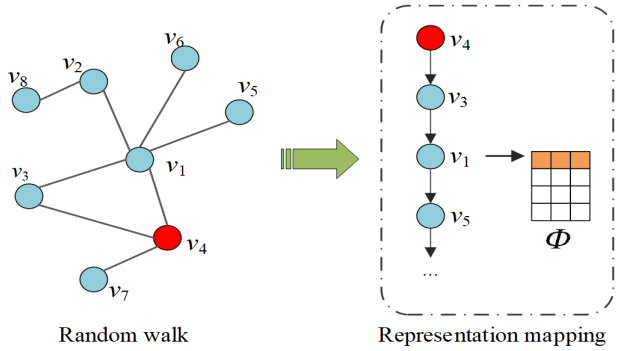

Fig. 1 Main steps of DW. Node v_4 is the initial node, the random walk is used to obtain the node sequence, and then the node sequence is inputted into the skip-gram to obtain the representation vector of the node.

2.6.3 Prediction of metabolite-disease associations by RF

RF is an algorithm that integrates multiple trees through the idea of ensemble learning, which depends on the classification of most decision trees to determine the final classification results[23, 24]. In metabolite-disease association data, positive and negative samples are unbalanced. Considering the wide application of RF and its good performance on unbalanced samples, we build a prediction model of metabolite-disease pairs based on RF. The framework of our method is shown in Fig. 2.

RF is a mature algorithm that has been integrated into the machine library in Python. We classified metabolite-disease pairs by feeding the final feature into RF. In the experiment, we use the grid search method to select

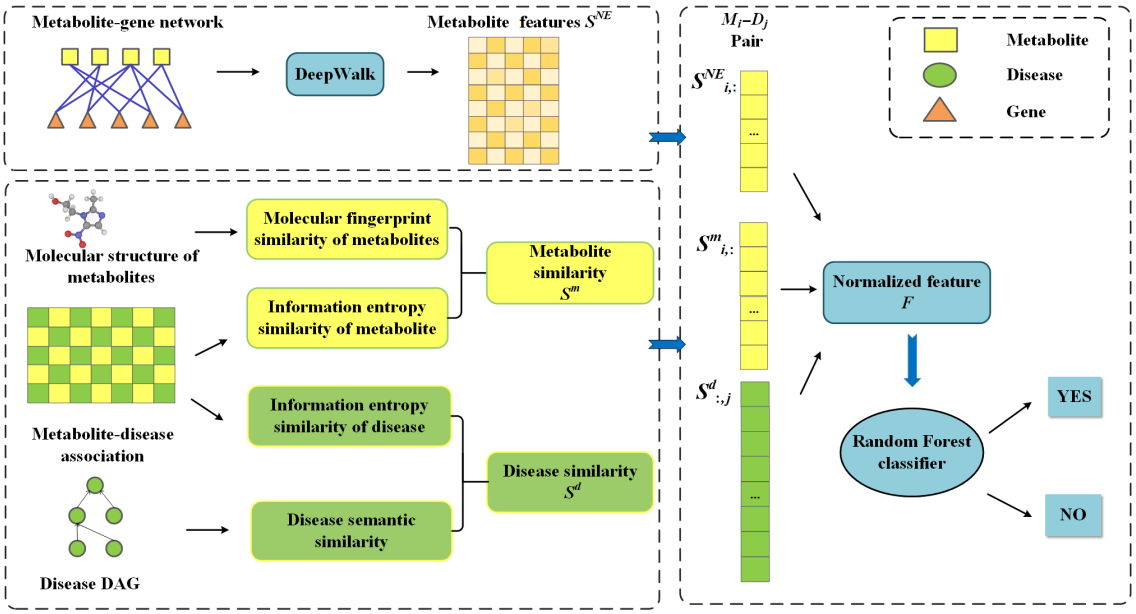

Fig. 2 Framework of DWRF.

the parameters. Finally, the main parameters of the RF classifier, namely, the *max*_features (the range is 0 to 1 with a step size of 0.1), *n*_estimators (the range is 10 to 50 with a step size of 10), and *min*_samples_ leaf (the range is 10 to 50 with a step size of 10), are set to 0.2, 30, and 10, respectively.

3 Result

3.1 Evaluation metrics

To illustrate the performance of our method at predicting latent metabolite-disease associations, we utilized five-fold cross-validation (FFCV) and leave-one-out cross-validation (LOOCV) for evaluation. The receiver operating characteristic curve (ROC) is a common standard for evaluating models. The area under the ROC curve is AUC. Moreover, a series of broader assessment criteria[25, 26], including the area under the precise-recall curve (AUPR), F1-measure (F1), accuracy (ACC), specificity (SPE), recall (REC), and precision (PRE), are used for more comprehensive and equitable evaluation of proposed models.

3.2 Performance of DWRF

Because the data of metabolite-disease associations are unbalanced, we randomly choose equal amounts of negative and positive samples. Then, FFCV and LOOCV are conducted to test the performance of our method. Figure 3 shows the ROC of our method under the FFCV and LOOCV. The AUPR, F1, ACC, SPE, REC, and PRE values of FFCV are shown in Table 1. DW is used to extract metabolite features from the metabolite-gene association network. The effect of the value of the DW dimension on the AUC is shown in Fig. 4, where the AUC value achieves the highest when the dimension is 128.

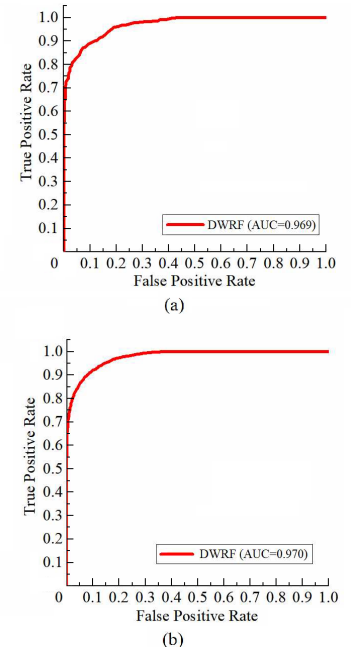

Fig. 3 ROC of our method. (a) ROC curve of DWRF under FFCV, (b) ROC curve of DWRF under LOOCV.

3.3 Comparison with other classifiers

To estimate the performance of a classifier, we compare the RF with common machine learning algorithms, including logistic regression, support vector machines (SVMs), Gaussian naive bayes (GNB), K-nearest neighbor, and AdaBoost under FFCV. The AUC comparison with the different classification algorithms is shown in Fig. 5, and the results of other evaluation criteria with different classification algorithms are shown

Table 1 Values of the AUPR, F1, ACC, REC, and PRE under FFCV.

Fold	AUPR	F1	ACC	SPE	REC	PRE
1	0.971	0.899	0.901	0.908	0.894	0.904
2	0.956	0.883	0.884	0.896	0.872	0.895
3	0.963	0.901	0.905	0.922	0.887	0.912
4	0.971	0.904	0.886	0.855	0.934	0.875
5	0.972	0.895	0.899	0.908	0.869	0.922
Average	0.966	0.896	0.895	0.897	0.891	0.901

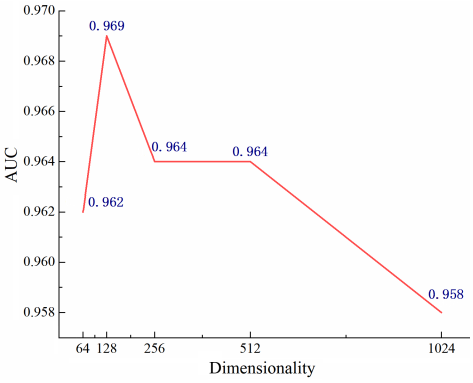

Fig. 4 Effect of the DW dimension.

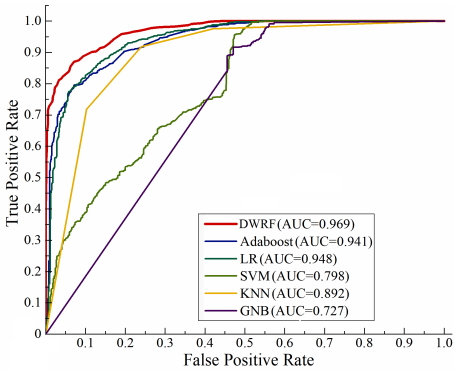

Fig. 5 AUC comparison with different classification algorithms.

in Table 2. It's worth saying that we input the features obtained in Eq. (10) into different classifiers.

Boxplot, also known as box-whisker plot, is a statistical plot used to display information about the dispersion of a set of data[27]. It is mainly used to reflect the distribution characteristics of the original data, and can also be used to compare the distribution characteristics of multiple groups of data. Because the number of negative samples is much more than the number of positive samples, we randomly take as many negative samples as the positive samples for 20 times and then perform FFCV under different classifiers to obtain AUC. Finally, the boxplots of 20 AUC values of different classifiers are drawn as shown in Fig. 6. In six different classifiers, DWRF has the highest median.

Table 2 Results of other evaluation criteria with different classification algorithms.

Classifier	AUPR	F1	ACC	SPE	REC	PRE
SVM	0.783	0.738	0.807	0.897	0.768	0.711
LR	0.935	0.842	0.876	0.826	0.885	0.844
GBDT	0.916	0.834	0.831	0.815	0.846	0.833
Adaboost	0.932	0.867	0.862	0.823	0.900	0.836
KNN	0.896	0.844	0.839	0.762	0.882	0.792
RF	**0.966**	**0.896**	**0.895**	**0.897**	**0.891**	**0.901**

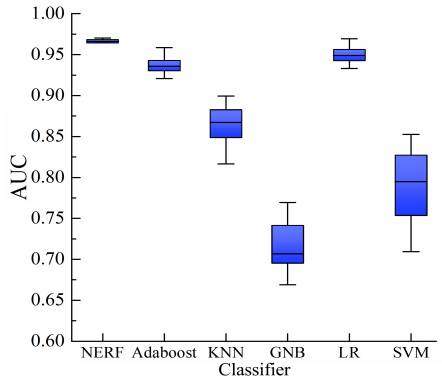

Fig. 6 Boxplot of 20 times FFCV for different classifiers.

This finding shows that DWRF has the best performance in predicting the associations between metabolites and diseases.

To further assess the performance of DWRF, we perform experiments on nine common diseases, namely, Lewy body dementia, Crohns disease, Parkinsons disease, autism, celiac disease, periodontal disease, uremia, prostate cancer, and gout. Figure 7 shows the comparison of these diseases in different classifiers under FFCV. As shown in Fig. 7, DWRF shows good performance as the AUC values are higher than those of other classifiers among the nine diseases, except gout.

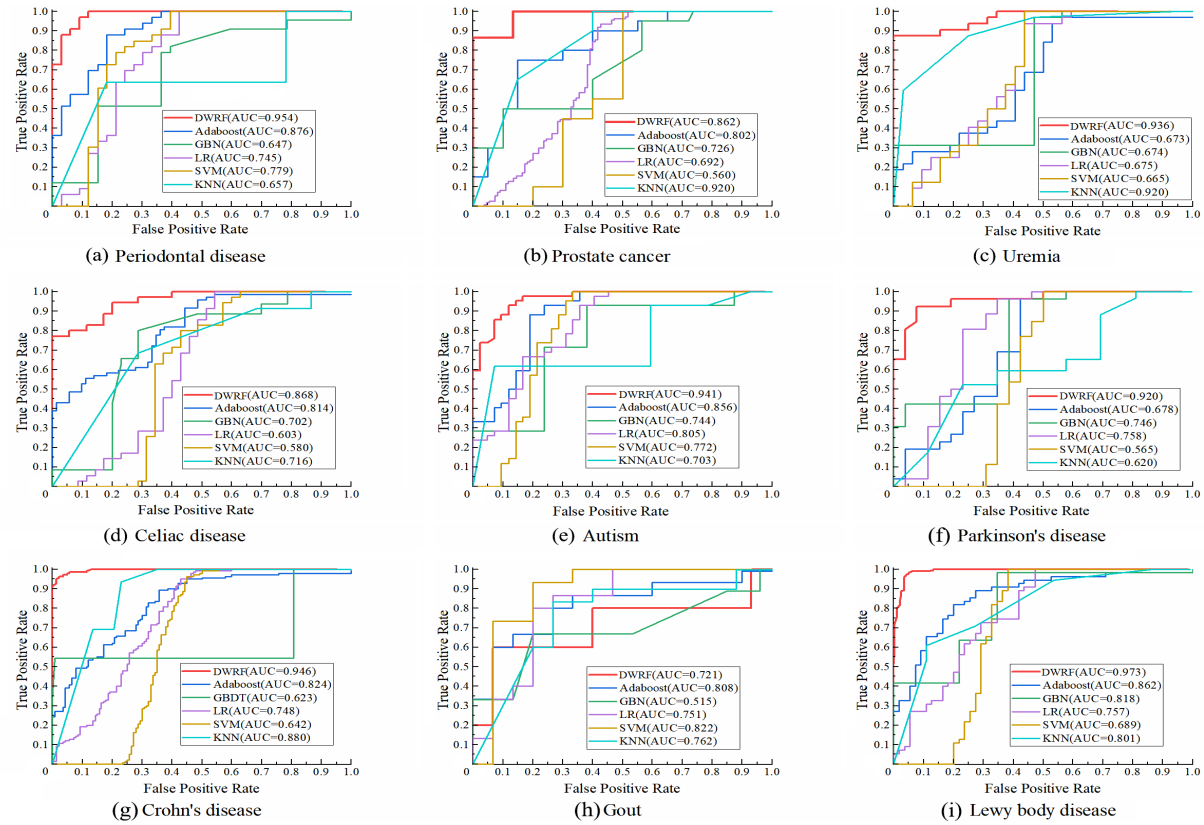

Fig. 7 Comparison of nine diseases in different classifiers under FFCV.

The AUC of SVM and DWRF on the FFCV results of gout were 0.882 and 0.721, respectively. In this study, DW is used to obtain the characteristics of metabolites. The parameters of DW mainly consider most diseases, and the prediction results of a few diseases with less known associations are expected to be lower.

3.4 Comparison with different features

In this study, we integrate the embedding features of metabolites into a biological similarity. We also perform experiments under FFCV: (a) only use the embedded features of metabolites obtained by DW and integrated disease similarity and (b) only use the biological similarity, including integrated disease similarity and integrated metabolite similarity. Figure 8 shows the ROC curve comparison results of the two experiments, where only DW and only BioSIM represent the abbreviation of the first and second experiments, respectively.

3.5 Case study

To testify the reliability of DWRF, we consider three common human diseases, namely, AD, colorectal cancer (CRC), and LC. The predicted scores are presented in a descending order, and we obtain the top 10 metabolites associated with the diseases. We successively take each related metabolite and search for verifiable literature from the National Center for Biotechnology Information (NCBI). The metabolite-disease associations can be verified by the literature. We give the PubMed Unique Identifier to the corresponding literature; otherwise, we set unconfirmed. AD is a neurodegenerative disease with an insidious onset and slow progression that has caused serious public health problems[28]. The top 10 candidate metabolites of AD are shown in Table 3, 9 of which can be found in verifiable literature in NCBI.

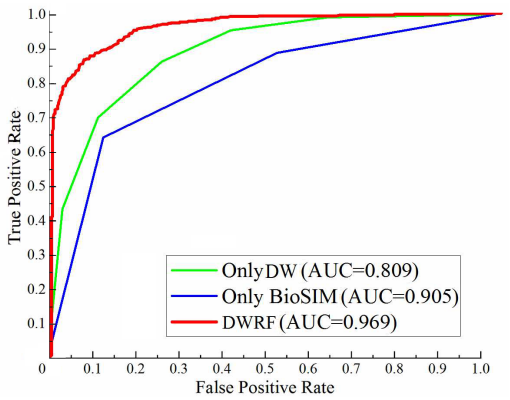

Fig. 8 ROC curve comparison results of the two experiments.

Table 3 Top 10 candidate metabolites associated with AD.

Rank	Metabolite Name	Evidences
1	Betaine	PMID:28671332
2	Adenosine monophosphate	Unconfirmed
3	L-Tyrosine	PMID:24898638
4	L-Phenylalanine	PMID:23857558
5	L-Alanine	PMID:21292280
6	L-Isoleucine	PMID:29519576
7	L-Lysine	PMID:9693263
8	L-Serine	PMID:28929385
9	L-Glutamine	PMID:26402632
10	Creatine	PMID:26402632

CRC has become a common type of cancer and currently ranks among the highest in morbidity and mortality worldwide[29]. The top 10 candidate metabolites of CRC are shown in Table 4, which can all be found in verifiable literature. LC is one of the malignancies with the fastest increase in morbidity and mortality and the greatest threat to human health and life[30]. The top 10 candidate metabolites of LC are shown in Table 5, 9 of which can be found in verifiable literature of NCBI. These results demonstrate that DWRF can effectively predict metabolite-disease associations.

We also draw the top 10 association networks predicted metabolite candidates for AD, CRC, and

Table 4 Top 10 candidate metabolites associated with CRC.

Rank	Metabolite Name	Evidences
1	Acetic acid	PMID:25700314
2	beta-Alanine	PMID:30296444
3	Creatine	PMID:29168152
4	8-hydroxy-Deoxyguanosine	PMID:30932412
5	Choline	PMID:25785727
6	Glycine	PMID:27351202
7	Gentisic acid	PMID:25037050
8	Hypoxanthine	PMID:28640361
9	L-Phenylalanine	PMID:31289671
10	L-Alanine	PMID:28207045

Table 5 Top 10 candidate metabolites associated with LC.

Rank	Metabolite Name	Evidences
1	Taurine	PMID:29552188
2	L-Alanine	PMID:25961003
3	Acetic acid	PMID:22157537
4	L-Threonine	Unconfrimed
5	Glycine	PMID:18953024
6	Betaine	PMID:23383301
7	Creatine	PMID:25961003
8	Trimethylamine N-oxide	PMID:22157537
9	Choline	PMID:25591716
10	L-Serine	PMID:29251665

LC and the genes associated with those metabolites. A disease is related to different metabolites, and a metabolite is associated with different genes. Different diseases can be associated with the same metabolites, so we remove 30 duplicated metabolites associated with the three common human diseases to obtain 20 metabolites and take out the common genes associated with the 20 metabolites from all the genes. In Fig. 9, the rectangle, blue hexagon, and green ellipse represent the disease, gene, and metabolite, respectively. The solid line in cyan indicates the association between CRC and metabolites. The solid line in purple indicates the association between LC and metabolites. The solid line in red indicates the association between AD and metabolites. The association between diseases and metabolites and metabolites and genes are represented by lines of different colors. Figure 9 shows that different diseases are related to the same metabolite. For instance, the metabolite creatine is related to LC, AD, and colon cancer. Additionally, different metabolites are related to the same gene, such that the gene UGT1A1 is related to gentisic acid, acetic acid, and beta-alanine. Based on this information, it is advisable to extract features from the association between genes and metabolites.

4 Conclusion

In this paper, we introduce known metabolite-gene associations. DW is used to extract the new metabolite features. Compared with the previous methods, the biological similarity features of metabolites and the topological features of bipartite networks are integrated to make more reliable prediction results of disease-related metabolites. Combining the biological information features and the features extracted from the metabolite-gene associations, the machine learning method, RF, is used to predict the potential metabolite-disease association. The results of LOOCV, FFCV, and case studies of three human diseases (AD, CRC, and LC) demonstrated that DWRF is a reliable prediction algorithm.

Nonetheless, our method still has limitations. The random selection of negative samples will lead to the deviation of the results, and a relatively good negative-sample selection strategy should be considered in future works. The accuracy of predicting new metabolites and isolated diseases should also be improved.

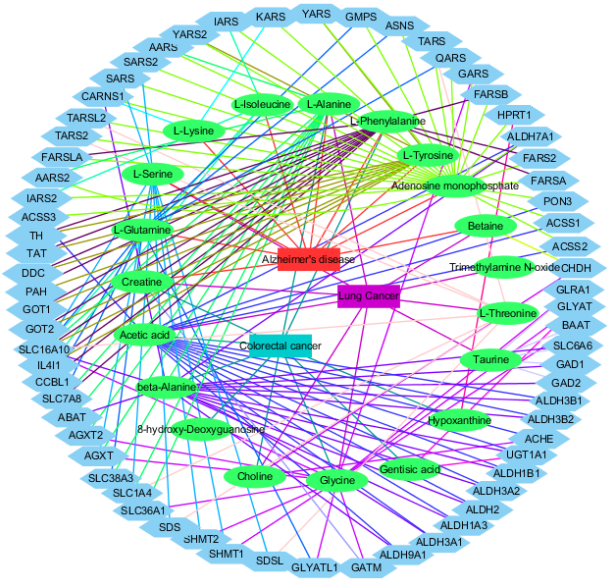

Fig. 9 Top 10 association networks predicted metabolite candidates for AD, CRC, and LC and the genes associated with those metabolites.

References

[1] J. A. Harris and F. G. Benedict, A biometric study of human basal metabolism, *Proc. Natl. Acad. Sci. USA*, vol. 4, no. 12, pp. 370–373, 1918.

[2] L. Cheng, H. X. Yang, H. Q. Zhao, X. Y. Pei, H. B. Shi, J. Sun, Y. P. Zhang, Z. Z. Wang, and M. Zhou, MetSigDis: A manually curated resource for the metabolic signatures of diseases, *Brief. Bioinform.*, vol. 20, no. 1, pp. 203–209, 2019.

[3] Y. M. Chen, Y. Liu, R. F. Zhou, X. L. Chen, C. Wang, X. Y. Tan, L. J. Wang, R. D. Zheng, H. W. Zhang, W. H. Ling, et al., Associations of gut-flora-dependent metabolite trimethylamine-N-oxide, betaine and choline with non-alcoholic fatty liver disease in adults, *Sci. Rep.*, vol. 6, no. 1, p. 19076, 2016.

[4] D. Y. Hui, Intestinal phospholipid and lysophospholipid metabolism in cardiometabolic disease, *Curr. Opin. Lipidol.*, vol. 27, no. 5, pp. 507–512, 2016.

[5] E. T. Oni, R. Kalathiya, E. C. Aneni, S. S. Martin, M. J. Blaha, T. Feldman, A. S. Agatston, R. S. Blumenthal, R. D. Conceiçao, J. A. M. Carvalho, et al., Relation of physical activity to prevalence of nonalcoholic Fatty liver disease independent of cardiometabolic risk, *Am.J. Cardiol.*, vol. 115, no. 1, pp. 34–39, 2015.

[6] A. Budhu, A. Terunuma, G. Zhang, S. P. Hussain, S. Ambs, and X. W. Wang, Metabolic profiles are principally different between cancers of the liver, pancreas and breast, *Int.J. Biol. Sci.*, vol. 10, no. 9, pp. 966–972, 2014.

[7] R. A. Moats, T. Ernst, T. K. Shonk, and B. D. Ross, Abnormal cerebral metabolite concentrations in patients with probable Alzheimer disease, *Magn. Reson. Med.*, vol. 32, no. 1, pp. 110–115, 1994.

[8] P. G. Unschuld, R. A. E. Edden, A. Carass, X. Y. Liu, M. Shanahan, X. Wang, K. Oishi, J. Brandt, S. S. Bassett, G. W. Redgrave, et al., Brain metabolite alterations and cognitive dysfunction in early Huntington's disease, *Mov. Disord.*, vol. 27, no. 7, pp. 895–902, 2012.

[9] S. Hori, S. Nishiumi, K. Kobayashi, M. Shinohara, Y. Hatakeyama, Y. Kotani, N. Hatano, Y. Maniwa, W. Nishio, T. Bamba, et al., A metabolomic approach to lung cancer, *Lung Cancer*, vol. 74, no. 2, pp. 284–292, 2011.

[10] C. Cheng, S. M. Zhuo, B. Zhang, X. Zhao, Y. Liu, C. L. Liao, J. Quan, Z. Z. Li, A. M. Bode, Y. Cao, et al., Treatment implications of natural compounds targeting lipid metabolism in nonalcoholic fatty liver disease, obesity and cancer, *Int. J. Biol. Sci.*, vol. 15, no. 8, pp. 1654–1663, 2019.

[11] Y. J. Xu, H. X. Yang, T. Wu, Q. Dong, Z. G. Sun, D. S. Shang, F. Li, Y. Q. Xu, F. Su, and S. Y. Liu, BioM2MetDisease: A manually curated database for associations between microRNAs, metabolites, small molecules and metabolic diseases, *Database*, vol. 2017, p. bax037, 2017.

[12] D. S. Wishart, Y. D. Feunang, A. Marcu, A. C. Guo, K. Liang, R. Vázquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, et al., HMDB 4.0: The human metabolome database for 2018, *Nucleic Acids Res.*, vol. 46, no. D1, pp. D608–D617, 2018.

[13] D. S. Shang, C. Q. Li, Q. L. Yao, H. X. Yang, Y. J. Xu, J. W. Han, J. Li, F. Su, Y. P. Zhang, C. L. Zhang, et al., Prioritizing candidate disease metabolites based on global functional relationships between metabolites in the context of metabolic pathways, *PLoS One*, vol. 9, no. 8, p. e104934, 2014.

[14] Y. Hu, T. Y. Zhao, N. Y. Zhang, T. Y. Zang, J. Zhang, and L. Cheng, Identifying diseases-related metabolites using random walk, *BMC Bioinformatics*, vol. 19, no. S5, p. 116, 2018

[15] Y. T. Wang, L. R. Juan, J. J. Peng, T. Y. Zang, and Y. D. Wang, Prioritizing candidate diseases-related metabolites based on literature and functional similarity, *BMC Bioinformatics*, vol. 20, no. 18, p. 574, 2019.

[16] Y. J. Qi, Random forest for bioinformatics, in *Ensemble Machine Learning: Methods and Applications*, C. Zhang and Y. Q. Ma, eds. Boston, MA, USA: Springer, 2012, pp. 307–323.

[17] C. Chen, A. Liaw, and L. Breiman, *Using Random Forest to Learn Imbalanced Data*, Berkeley, CA, USA: University of California, 2004.

[18] H. J. Lowe and G. O. Barnett, Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches, *JAMA*, vol. 271, no. 14, pp. 1103–1108, 1994.

[19] Z. Q. Fang and X. J. Lei, Prediction of miRNA-circRNA associations based on k-NN multi-label with random walk restart on a heterogeneous network, *Big Data Mining and Analytics*, vol. 2, no. 4, pp. 261–272, 2019.

[20] X. Y. Li, Y. P. Lin, C. L. Gu, and J. L. Yang, FCMDAP: Using miRNA family and cluster information to improve the prediction accuracy of disease related miRNAs, *BMC Syst. Biol.*, vol. 13, no. 2, p. 26, 2019.

[21] B. Perozzi, R. Al-Rfou, and S. Skiena, DeepWalk: Online learning of social representations, in *Proc. 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, New York, NY, USA, 2014, pp. 701–710.

[22] DL4J, https://deeplearning4j.org, 2021.

[23] A. Liaw and M. Wiener, Classification and regression by randomForest, *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[24] W. Jiang, J. Y. Lin, H. Q. Wang, and S. C. Zou. Hybrid semantic service matchmaking method based on a random forest, *Tsinghua Sci. Technol.*, vol. 25, no. 6, pp. 798–812, 2020.

[25] G. Y. Wu, X. Guo, and B. H. Xu. BAM: A block-based Bayesian method for detecting genome-wide associations with multiple diseases, *Tsinghua Sci. Technol.*, vol. 25, no. 5, pp. 678–689, 2020.

[26] M. Bouazizi and T. Ohtsuki, Multi-class sentiment analysis on twitter: Classification performance and challenges, *Big Data Mining and Analytics*, vol. 2, no. 3, pp. 181–194, 2019.

[27] P. J. Rousseeuw, I. Ruts, and J. W. Tukey, The bagplot: A bivariate boxplot, *Am. Stat.*, vol. 53, no. 4, pp. 382–387, 1999.

[28] M. Goedert and M. G. Spillantini, A century of Alzheimer's disease, *Science*, vol. 314, no. 5800, pp. 777–781, 2006.

[29] R. L. Siegel, K. D. Miller, S. A. Fedewa, D. J. Ahnen, R. G. S. Meester, A. Barzi, and A. Jemal, Colorectal cancer statistics, 2017, *CA: A Cancer J. Clin.*, vol. 67, no. 3, pp. 177–193, 2017.

[30] C. C. Zhang, L. F. Ma, Y. J. Niu, Z. X. Wang, X. Xu, Y. Li, and Y. C. Yu, Circular RNA in lung cancer research: Biogenesis, functions, and roles, *Int. J. Biol. Sci.*, vol. 16, no. 5, pp. 803–814, 2020.

Xiujuan Lei received the MS and PhD degrees from Northwestern Polytechnical University, China, in 2001 and 2005, respectively. She is currently a professor at the School of Computer Science, Shaanxi Normal University. Her research interests include bioinformatics, swarm intelligent optimization, and deep learning.

Jiaojiao Tie is currently persuing the MS degree in Shaanxi Normal University, China. Her main research interests are biological network analysis and metabolite-disease association prediction.

Yi Pan is currently a Regents' Professor Emeritus and has served as the chair of Computer Science Department at Georgia State University during 2005–2020. He has also served as an interim associate dean and chair of Biology Department during 2013–2017. He joined Georgia State University in 2000, was promoted to full professor in 2004, named a Distinguished University Professor in 2013, and designated a Regents' Professor (the highest recognition given to a faculty member by the University System of Georgia) in 2015. He is a professor of the School of Computer Science and Control Engineering, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences since 2021. His current research interests mainly include bioinformatics and health informatics using big data analytics, cloud computing, and machine learning technologies. He has published more than 450 papers including over 250 journal papers with more than 100 papers published in IEEE/ACM Transactions journals. In addition, he has edited/authored 43 books. His work has been cited more than 16293 times based on Google Scholar and his current h-index is 82. He has served as an editor-in-chief or editorial board member for 20 journals including 7 IEEE Transactions. Currently, he is serving as an associate editor-in-chief of *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. He is the recipient of many awards including one IEEE Transactions Best Paper Award, five IEEE and other international conference or journal Best Paper Awards, 4 IBM Faculty Awards, 2 JSPS Senior Invitation Fellowships, IEEE BIBE Outstanding Achievement Award, IEEE Outstanding Leadership Award, NSF Research Opportunity Award, and AFOSR Summer Faculty Research Fellowship. He has organized numerous international conferences and delivered keynote speeches at over 60 international conferences around the world.