frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 24 August 2022
DOI 10.3389/fgene.2022.995532

OPEN ACCESS

EDITED BY
Rui Yin,
Harvard Medical School, United States

REVIEWED BY
Jin-Xing Liu,
Qufu Normal University, China
Cheng Liang,
Shandong Normal University, China
Guoxian Yu,
Shandong University, China
Cunmei Ji,
Qufu Normal University, China

*CORRESPONDENCE
Dengju Yao,
ydkvictory@hrbust.edu.cn

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 16 July 2022
ACCEPTED 01 August 2022
PUBLISHED 24 August 2022

CITATION
Yao D, Zhang T, Zhan X, Zhang S, Zhan X and Zhang C (2022), Geometric complement heterogeneous information and random forest for predicting lncRNA-disease associations. *Front. Genet.* 13:995532. doi: 10.3389/fgene.2022.995532

Geometric complement heterogeneous information and random forest for predicting lncRNA-disease associations

Dengju Yao[1]*, Tao Zhang[1], Xiaojuan Zhan[1,2], Shuli Zhang[1], Xiaorong Zhan[3] and Chao Zhang[4]

[1]School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China, [2]College of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin, China, [3]Department of Endocrinology and Metabolism, Hospital of South University of Science and Technology, Shenzhen, China, [4]Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, China

More and more evidences have showed that the unnatural expression of long non-coding RNA (lncRNA) is relevant to varieties of human diseases. Therefore, accurate identification of disease-related lncRNAs can help to understand lncRNA expression at the molecular level and to explore more effective treatments for diseases. Plenty of lncRNA-disease association prediction models have been raised but it is still a challenge to recognize unknown lncRNA-disease associations. In this work, we have proposed a computational model for predicting lncRNA-disease associations based on geometric complement heterogeneous information and random forest. Firstly, geometric complement heterogeneous information was used to integrate lncRNA-miRNA interactions and miRNA-disease associations verified by experiments. Secondly, lncRNA and disease features consisted of their respective similarity coefficients were fused into input feature space. Thirdly, an autoencoder was adopted to project raw high-dimensional features into low-dimension space to learn representation for lncRNAs and diseases. Finally, the low-dimensional lncRNA and disease features were fused into input feature space to train a random forest classifier for lncRNA-disease association prediction. Under five-fold cross-validation, the AUC (area under the receiver operating characteristic curve) is 0.9897 and the AUPR (area under the precision-recall curve) is 0.7040, indicating that the performance of our model is better than several state-of-the-art lncRNA-disease association prediction models. In addition, case studies on colon and stomach cancer indicate that our model has a good ability to predict disease-related lncRNAs.

KEYWORDS

lncRNA-disease association prediction, geometric complement heterogeneous information, random forest, autoencoder, machine learning

1 Introduction

Long non-coding RNA (lncRNA) is a kind of non-coding RNA with a length of more than 200 nucleotides, which have received increasing attention from researchers. LncRNAs have now been proved to play a key role in transcriptional and posttranslational regulation (Taft et al., 2010; Mathieu et al., 2014; Sun et al., 2018; Xie et al., 2018). The pathogenesis of a series of diseases is significantly associated with mutations and dysregulation of lncRNAs (Washietl et al., 2014; Chen et al., 2017). For example, MALAT1 was discovered to be overexpressed in many entity tumors such as lung cancer (Cheetham et al., 2013). It was shown that clonogenic and anchorage-dependent growth of lung cancer cells would be significantly decreased when H19 was down-regulated (Barsyte-Lovejoy et al., 2006). Confirming the associations between lncRNAs and diseases by biological experiments is time-consuming, labor-intensive and challenging, so using computational method to predict the associations not only provides a more efficient way for biological experiments but also reduces a lot of unnecessary human and material resources. Currently, dozens of computational models have been proposed to identify disease-associated lncRNAs based on various biological data. We can broadly classify the current computational models for lncRNA-disease association (LDA) prediction into three categories.

The first class of LDA prediction models is based on biological networks. Sun et al. implemented random walk and restart on lncRNA functional similarity network (Sun et al., 2014). Zhou et al. integrated the LDA network, disease similarity network and lncRNA-miRNA interaction network into a heterogeneous network and applied random walk on the network (Zhou et al., 2015). Chen et al. integrated the known LDAs, lncRNA expression profiles, lncRNA functional similarity, disease semantic similarity and Gaussian interaction profile kernel similarity to predict potential LDAs (Chen, 2015a). Ping et al. (2019) constructed a model based on the known LDA network. However, these models need the known LDA network. Thus, Liu et al. (2014) conceived a model by integrating the known human expression profiles of lncRNA and disease genes, which is the first computational model without relying on the known LDAs. Chen et al. combined miRNA-disease association and lncRNA-miRNA interactions to form a model called HGLDA (Chen, 2015c). Zhou et al. developed a computational method by integrating association among lncRNA, protein, disease, miRNA, drug and high-order proximity preserved embedding for predicting LDAs (Zhou et al., 2021). Sumathipala et al. used the topology of a multi-level network consisting of lncRNA-protein, protein-protein interactions and protein-disease associations to identify LDAs (Sumathipala et al., 2019). Yu et al. used Bi-Random Walks on the lncRNA functional similarity network and disease network to predict LDAs (Yu et al., 2017). Yu et al. (2020) constructed a data fusion model called Attributed Heterogeneous Network Fusion for LDA prediction (AHNF).

The second class of LDA prediction model is based on matrix factorization. Fu et al. proposed a LDA prediction model called MFLDA. MFLDA factored data from heterogeneous data sources into low-rank matrices based on matrix trivialization to discover and explore its intrinsic and shared structure (Fu et al., 2018). Wu et al. constructed a GAMCLDA model by encoding local graph structures and features. The graph convolution network was used to encode the features of this map structure and nodes to learn the potential factorial vectors of lncRNAs and diseases. In addition, the inner product of lncRNA factor vectors and disease factor vectors was used as a decoder to reconstruct the LDA matrix (Wu et al., 2020). Gao et al. (2021) constructed a multi-label fusion collaborative matrix decomposition approach to predict LDAs. Wang et al. (2020) developed a weighted matrix factorization model on multi-relational data to predict LDAs. Liu et al. (2021) introduced a weighted graph regularized collaborative matrix factorization (WGRCMF) method to predict LDAs.

The third class of LDA prediction model is based on machine algorithms. Machine learning methods focus on gaining insights into features and imbalanced labels. Chen et al. formulated Laplace regularized least squares method to predict LDAs (called LRLSLDA) in a semi-supervised learning framework, which is the first machine learning-based methods to predict LDAs (Chen et al., 2015). However, for LRLSLDA, parameter optimization is a challenge. Later, Chen et al. combined lncRNA functional similarity with the LRLSLDA-LNCSIM prediction model and enhanced its performance by introducing similarity scores for predicting gene-disease associations (Huang et al., 2016). In addition, Lan et al. implemented a LDAP model based on SVM bagging by combining disease similarity and lncRNA similarity (Lan et al., 2017). Yao et al. constructed a computational model called RFLDA to identify associations based on feature selection by integrating the experiment-supported associations among lncRNA, miRNA, disease, disease semantic similarity and lncRNA functional similarity (Yao et al., 2020). Xuan et al. have developed a collection of convolutional neural networks-based lncRNA-disease prediction models, including CNNLDA (Xuan et al., 2019a), LDAPred (Xuan et al., 2019b), GCNLDA (Xuan et al., 2019c) and CNNDLP (Xuan et al., 2019d). The CNNLDA developed an analysis of the associations between lncRNA and disease using convolutional neural networks that combined semantic and functional similarity as well as lncRNA-disease associations, miRNA-disease associations and lncRNA-miRNA interactions (Xuan et al., 2019a). The LDAPred integrated a convolutional neural network and information flow propagation, combining associations, interactions, similarity structures and topological structures between lncRNAs, miRNAs and diseases (Xuan et al., 2019b). The GCNLDA is based on the graph convolutional network and convolutional neural network to obtain locally

integrated topological information within the lncRNA-disease-microRNA networks (Xuan et al., 2019c). By combining disease similarity, lncRNA similarity, miRNA-disease association and lncRNA-miRNA interactions, CNNDLP learned the attention and the low-dimensional network representation of the lncRNA-disease pairs (Xuan et al., 2019d). Wei et al. developed a method (LDICDL) that denoised lncRNA and disease features with an autoencoder, and used the matrix decomposition algorithm to test for potential disease-lncRNA association (Lan et al., 2022). Fan et al. proposed an lncRNA-disease prediction method that implemented convolutional matrices with conditional random fields and attention mechanisms for learning the embeddings of nodes for scoring latent associations between lncRNAs and diseases (Fan et al., 2022). Wu et al. proposed a method that combined extra trees with multi-layer graph embedding aggregation to predict LDAs (Wu Q. W. et al., 2021). Cui et al. proposed a novel model based on bipartite local model with nearest profile-based association inferring to predict LDAs (Cui et al., 2020).

These methods described above have achieved good prediction performance, but they also have some limitations. The biological network-based approach was affected by the scarcity of known LDA data; For the matrix factorization-based approach, the combination of model parameters is a very complex and necessary procedure; For the machine learning-based approach, feature processing and the impact of imbalanced data is a challenge. In this paper, we proposed a novel LDA prediction model based on geometric complement heterogeneous information and random forest (GCHIRFLDA in short). Firstly, the geometric complementation of LDA matrix was implemented by integrating the information of lncRNA-miRNA and miRNA-disease association information. Secondly, a low-dimensional feature space was extracted from the obtained LDA matrix by using an autoencoder, which combined Jaccard similarity coefficient and Gaussian interaction profile kernel similarity. Finally, a random forest classifier was trained on the constructed sample set to score potential lncRNA-disease associations. The AUC and AURP under five-fold cross-validation demonstrated that the GCHIRFLDA had a better performance than several state-of-the-art LDA prediction models, and the case studies on stomach cancer and colon cancer indicated that the GCHIRFLDA had excellent ability in identifying disease-associated lncRNAs.

2 Materials and methods

2.1 Representation of lncRNA-disease associations , miRNA-disease associations and lncRNA-miRNA interactions

LncRNA-disease associations (LDA), miRNA-disease associations (MDA) and lncRNA-miRNA interactions (LMI) were obtained from previous reports (Fu et al., 2018). The following l, d and m denote the number of lncRNA, disease and miRNA, respectively. The LDAs are represented by a 240× × 412 adjacency matrix $LD_{i\times j} \in LD^{l\times d}$, l is rows represent lncRNAs and d is columns represent diseases. For each element $LD_{i,j}$, its value is equal to one if lncRNA i is related to disease j; otherwise, its value is equal to 0. Similarly, the MDAs are represented by a 495× × 412 adjacency matrix $MD_{i\times j} \in MD^{m\times d}$, m is rows represent miRNAs and d is columns represent diseases. For each element $MD_{i,j}$, its value is equal to one if miRNA i is related to disease j; otherwise, its value is equal to 0. The LMIs are represented by a 240 × ×495 adjacency matrix $LM_{i\times j} \in LM^{l\times m}$, l is rows represent lncRNAs and m is columns represent diseases. For each element $LM_{i,j}$, its value is equal to one if lncRNA i is related to miRNA j; otherwise, its value is equal to 0.

2.2 Calculation of jaccard similarity of disease and lncRNA

Calculation of similarity of disease and lncRNA is an important step in LDAs predicting process. So far, there are many ways to calculate similarity, such as disease semantic similarity, disease cosine similarity, lncRNA functional similarity and lncRNA cosine similarity. In this work, we combine the Jaccard similarity coefficient which is complementary to the binary matrix and the Gaussian interaction profile kernel similarity which encodes the non-linear vectors in the LDA matrix. By experimental research on different similarity measures, we found that the fusion of these two kinds of similarity can greatly improve the performance of the LDA prediction model. Therefore, we chose Jaccard similarity and Gaussian interaction profile kernel similarity for LDA prediction in this work. Thank you again for your comment. The Jaccard similarity coefficient (Jaccard, 1908) of disease was calculated by LDA matrix by Eq. 1:

$$JDS(i,j) = \frac{LD(:,i) \cap LD(:,j)}{LD(:,i) \cup LD(:,j)} \tag{1}$$

In Eq. 1, $LD(:,i)$ is the i-th column vector of the LDA matrix, which represents the association feature of disease i; $LD(:,i) \cap LD(:,j)$ represents the number of lncRNAs that are associated with both disease i and disease j; $LD(:,i) \cup LD(:,j)$ represents the sum of the number of lncRNAs associated with the disease i and disease j.

Similarly to disease, the Jaccard similarity of lncRNA can be calculated by LDA matrix by Eq. 2:

$$JFS(i,j) = \frac{LD(i,:) \cap LD(j,:)}{LD(i,:) \cup LD(j,:)} \tag{2}$$

In Eq. 2, $LD(i,:)$ is the i-th row vector of the LDA matrix, which represents the association feature of lncRNA i;

$LD(i,:) \cap LD(j,:)$ represents the number of diseases that are associated with both lncRNA i and lncRNA j; $LD(i,:) \cup LD(j,:)$ represents the sum of the number of diseases associated with the lncRNA i and lncRNA j.

2.3 Calculation of Gaussian interaction profile kernel similarity of disease and lncRNA

The Gaussian interaction profile kernel similarity (Chen, 2015b) $GIP_{lnc}(l_i, l_j)$ between lncRNA l_i and lncRNA l_j was calculated by Eq. 3:

$$\begin{cases} GIP_{lnc}(l_i, l_j) = exp(-\lambda \| LD(i,:) - LD(j,:) \|^2) \\ \lambda = \tilde{\lambda} \Big/ \left(\frac{1}{l} \sum_{i=1}^{l} \| l_i \|^2 \right) \end{cases} \tag{3}$$

From the above equation, the Gaussian interaction profile kernel similarity matrix of lncRNA can be obtained. $LD(i,:)$ and $LD(j,:)$ represents i-th and j-th row of LDA matrix respectively, $\tilde{\lambda}$ controls the kernel bandwidth, in this work, we set $\tilde{\lambda}$ to 1.

Similarly, the Gaussian interaction profile kernel similarity matrix of disease $GIP_{dis}(d_i, d_j)$ can be obtained by Eq. 4.

$$\begin{cases} GIP_{dis}(d_i, d_j) = exp(-\lambda \| LD(:,i) - LD(:,j) \|^2) \\ \lambda = \tilde{\lambda} \Big/ \left(\frac{1}{d} \sum_{i=1}^{d} \| d_i \|^2 \right) \end{cases} \tag{4}$$

In Eq. 4, $LD(:,i)$ and $LD(:,j)$ represents i-th and j-th column of LDA matrix respectively, $\tilde{\lambda}$ controls the kernel bandwidth, in this work, we set $\tilde{\lambda}$ to 1.

2.4 Fusing different similarities for lncRNA and disease

In this paper, we used the maximum value method to merge lncRNA Gaussian interaction profile kernel similarity and lncRNA Jaccard similarity into LFJ similarity and fuse disease Gaussian interaction profile kernel similarity and disease Jaccard similarity into DSJ similarity by Eqs. 5, 6, respectively.

$$LFJ\ similarity = \begin{cases} GIP_{lnc}(l_i, l_j) & if\ GIP_{lnc}(l_i, l_j) \geq JFS(i,j) \\ JFS(i,j) & otherwise \end{cases} \tag{5}$$

$$DSJ\ similarity = \begin{cases} GIP_{dis}(d_i, d_j) & if\ GIP_{dis}(d_i, d_j) \geq JDS(i,j) \\ JDS(i,j) & otherwise \end{cases} \tag{6}$$

2.5 Geometric complement for lncRNA-disease associations matrix

The process of constructing the GCHIRFLDA model is divided into three steps (see Figure 1): 1) geometric complement for LDA matrix; 2) feature representation and extraction; 3) random forest classifier training and LDA prediction. Next, we will introduce the process of constructing the GCHIRFLDA model in detail.

Inspired by Francesco et al.'s and Yin et al.'s method (Wang et al., 2021; Yin et al., 2022), from the previous data source, we multiplied the LMI matrix with the MDA matrix and then divided the [i, j]-th element of the result by the i-th row of the LMI matrix and the j-th column of the MDA matrix to represent the potential LDA matrix by Eq. 7:

$$\mathrm{LMD}(\mathrm{i},\mathrm{j}) = \frac{LM(i,:) \cdot MD(:,j)}{\| LM(i,:) \|_1 + \| MD(:,j) \|_1} \tag{7}$$

The fusion matrix of LDA was obtained by taking the maximum value of the potential LDAs computed above and the original LDA matrix in the i-th row and j-th column by Eq. 8.

$$LD_{\mathrm{new}}(i,j) = \max(LD(i,j), \mathrm{LMD}(\mathrm{i},\mathrm{j})) \tag{8}$$

In this way, the original LDA matrix can be geometrically complemented.

2.6 Feature representation and extraction

For the obtained geometric complement matrix, each row represents the feature vector of lncRNA and each column represents the feature vector of disease. We combine the i-th row of the geometric complement matrix and the i-th row of the similarity fusion matrix of lncRNA to form a new feature vector of the i-th lncRNA. Similarly, we combine the j-th column of the geometric complement matrix and the j-th column of the similarity fusion matrix of disease to form a new feature vector of the j-th disease. Finally, each lncRNA and disease is represented as a 652-dimensional feature vector.

Autoencoder is an unsupervised neural network model and has a good performance in data denoising and dimensionality reduction. In the GCHIRFLDA model, we employee autoencoder to compress feature space of lncRNA and disease. We set hidden layer to learn the high-dimensional feature space of the input data so that the hidden layer can reconstruct the original input data (Schmidhuber, 2015; Ji et al., 2021).

In this work, we use an autoencoder with an input layer, a dense layer, an output layer and a fully-connected layer with an activation function sigmoid. The learning process of the noise-reducing encoder is to minimize the error between the reconstructed data and the original data. As a result, each lncRNA, which is originally represented by a 652-dimensional

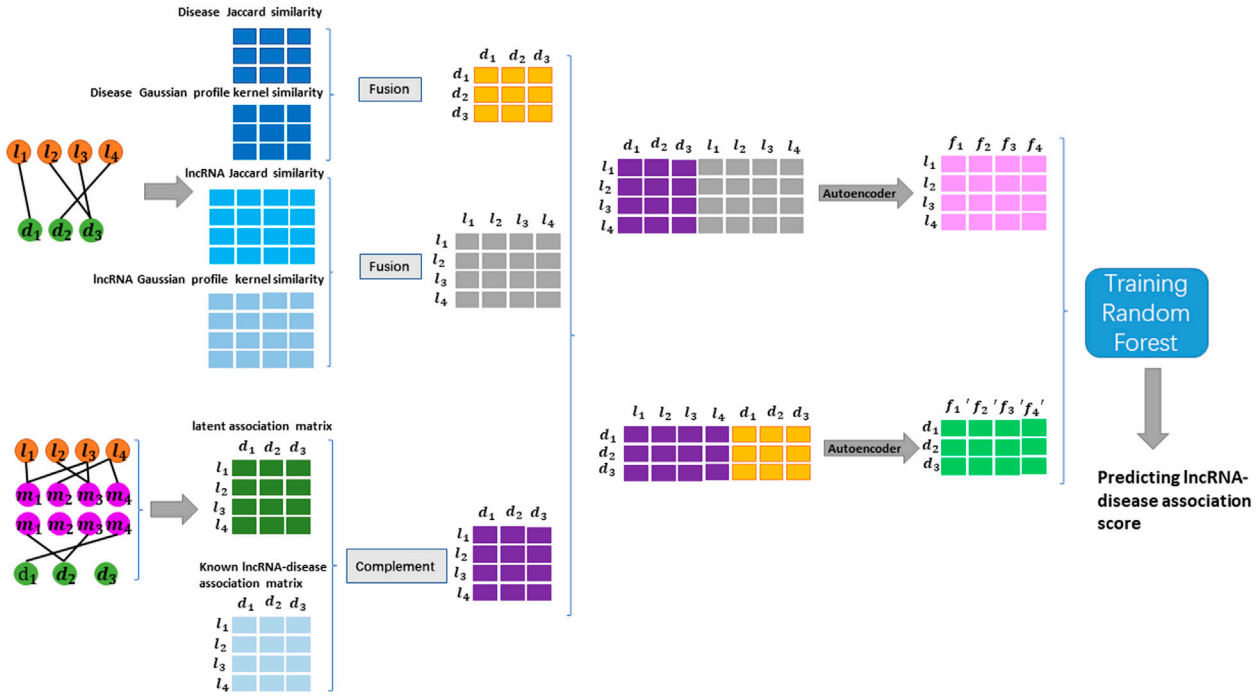

FIGURE 1
The flowchart of constructing the GCHIRFLDA model.

feature vector, is finally compressed into 256-dimensional by autoencoder. Similarly, each disease, which is originally represented by a 652-dimensional feature vector, is finally compressed into 256-dimensional by autoencoder. MSE (mean squared error) is used as model loss evaluation by Eq. 9:

$$\text{loss} = \frac{1}{n}\sum\left(Y_{input} - Y_{output}\right)^2 \tag{9}$$

In Eq. 9, Y_{input} is the original input data, and Y_{output} is the decoded and reconstructed data.

2.7 Random forest classifier training and lncRNA-disease associations prediction

To train the GCHIRFLDA model, the experiment-supported 2697 LDAs in the original LDA matrix were used as positive samples; the remaining lncRNA-disease pairs that were not validated by biological experiments were used as unlabeled samples. To maintain the balance of the training set, an equal number of unlabeled samples were randomly selected from the unlabeled samples as negative samples. The negative samples and the positive samples were combined into the training sample set which consisted of 5394 samples.

For accurately predicting potential LDAs, we employed random forest (RF) for LDA prediction in the GCHIRFLDA model. Random forest is an ensemble machine learning model which combines bagging and random features to add extra diversity of the decision tree model and finally uses a voting method to combine the prediction results of multiple base classifiers (Breiman, 2001). RF has many advantages: 1) it can process a variety of data types, including qualitative data or quantitative data; 2) it has high classification accuracy; 3) it has good robustness for noise data and data with missing values; 4) it has ability to analyze complex interactions between features. In recent years, RF has been widely used in a variety of classification and prediction problems, including differential expression analysis of microarray data, miRNA-disease association prediction, etc. In this work, we have carried out experimental research on six different classifiers, including SVM and Xgboost. Considering AUC, AUPR, Recall and other indicators, the performance of RF classifier is the best. Therefore, RF was chosen as the final classifier in our prediction model. RF has two important parameters, namely the number of randomly selected features (*mtry*) and the number of trees (*ntree*). These parameters have a great impact on the performance of random forest classification model. Here, we set *mtry* and *ntree* by the default value. Then, by the obtained prediction model, all unconfirmed lncRNA-disease pairs are scored, and the closer the score is to 1, the more likely it is that lncRNA is associated with the disease.

TABLE 1 The AUCs under different lncRNA/disease feature dimension.

Dimension	16	32	64	128	256	512
16	0.9576	0.9724	0.9768	0.9782	0.9750	0.9724
32	0.9492	0.9753	0.9775	0.9809	0.9804	0.9788
64	0.9577	0.9760	0.9791	0.9833	0.9842	0.9826
128	0.9561	0.9764	0.9808	0.9872	0.9884	0.9877
256	0.9539	0.9736	0.9804	0.9874	**0.9897**	0.9889
512	0.9109	0.9711	0.9793	0.9880	0.9891	0.9890

TABLE 2 The performance comparison of different classifiers in the GCHIRFLDA model.

Classifier	AUC	AUPR	Recall	Accuracy	F1-score
Xgboost	0.9815	0.4544	0.9523	0.9182	0.9523
RF	**0.9897**	**0.7040**	**0.9673**	**0.9317**	**0.9597**
C50	0.9513	0.1517	0.9340	0.8724	0.9265
GBDT	0.9497	0.2348	0.8942	0.8701	0.9253
SVM	0.9832	0.5826	0.9243	0.9313	0.9595
LightGBM	0.9832	0.5250	0.9428	0.9215	0.9541

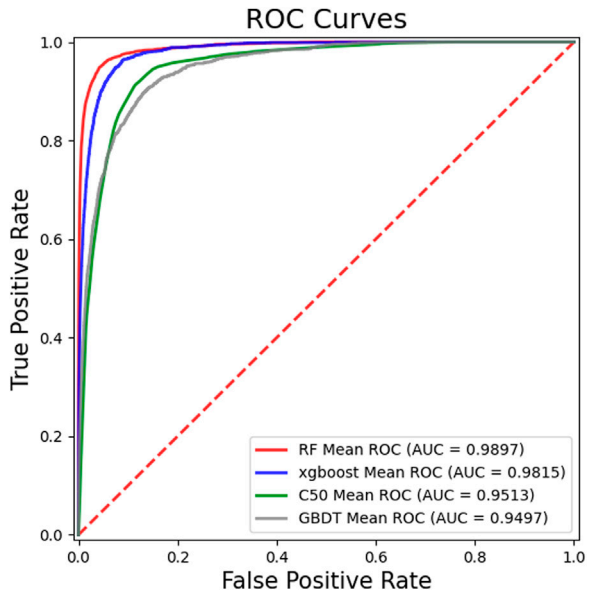

FIGURE 2
The ROC Curves of different classifiers in the GCHIRFLDA model.

3 Results

3.1 Feature dimension analysis of lncRNA and disease

For LDA prediction, the dimensionality of the training sample set has an obvious impact on the accuracy of the prediction model. On the one hand, for a smaller number of features of lncRNAs and diseases, more features are not learned, which leads to under-fitting of the model. On the other hand, for a larger number of features, more time is spent and the model performance will not yet be greatly improved or even over-fitting will occur. Therefore, we used the experimental method to determine the appropriate feature dimension. Specifically, we use autoencoder to compress the dimensions of feature space into 16, 32, 64, 128, 256, and 512 respectively, and the feature dimension that makes the prediction performance of the model the highest is adopted. Table 1 shows the AUC obtained under five-fold cross-validation by different dimensional features, from which one can see that the maximum of AUC is reached when the feature dimension of both lncRNAs and diseases is 256, so we set the feature dimension of extracted lncRNAs and diseases by autoencoder to be 256.

3.2 Performance comparison between random forest and other classifiers

In order to obtain better performance of the GCHIRFLDA model, we compared RF classifier with several classical classifiers, including extreme gradient boosting (Xgboost) (Chen and Guestrin, 2016), C50 (Kuhn, 2013), Gradient Boosting Decision Tree (GBDT) (Ye et al., 2009), SVM (Lan et al., 2017) and LightGBM (Zhang et al., 2021). In this work, we used the average AUC, AUPR, Recall, F1-score and Accuracy based on five-fold cross-validation as evaluation criterion for the six classifiers.

Figure 2 showed the ROC curves and AUCs of different classifiers, from which one can see that the AUC values of RF, Xgboost, C50 and GBDT are 0.9897, 0.9814, 0.98959 and 0.9497, respectively. Figure 3 showed the PR curves and AUPRs of four classifiers, the AUPR values of RF, Xgboost, C50 and GBDT are 0.704, 0.4505, 0.1607 and 0.2336, respectively. Table 2 showed the AUC, AUPR, Recall, F1-score and Accuracy of six classifiers. As one can see from Table 2, all five metrics of RF is the largest among the six classifiers. The results of the experiments suggested that RF outperformed the other five classifiers for LDA prediction. There, RF was finally determined as the final classifier in the GCHIRFLDA model.

3.3 Performance comparison between GCHIRFLDA and other lncRNA-disease associations prediction models

To evaluate the prediction performance of the GCHIRFLDA model, we compared it with seven state-of-the-art LDA

TABLE 3 The AUCs and AUPRs of different LDA prediction models.

Method	AUC	AUPR
GCHIRFLDA	**0.990**	**0.704**
GAERF	0.980	0.491
GCNLDA	0.959	0.223
CNNLDA	0.952	0.251
LDAP	0.863	0.166
MFLDA	0.626	0.066
Ping's Method	0.871	0.219
SIMCLDA	0.746	0.095

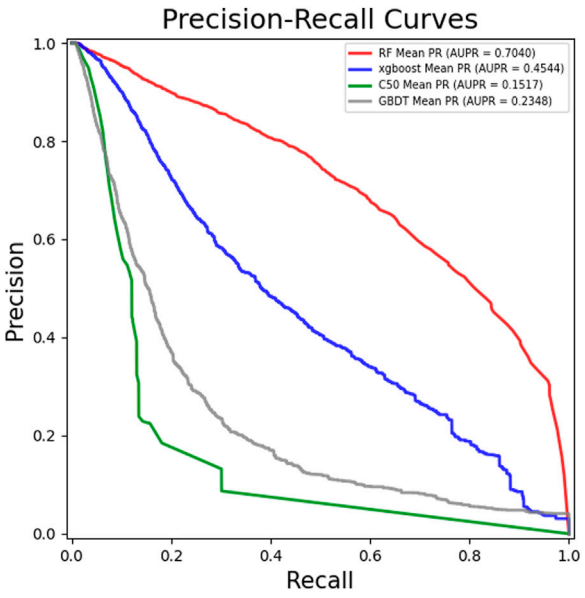

FIGURE 3
The Precision-Recall Curves of different classifiers in the GCHIRFLDA model.

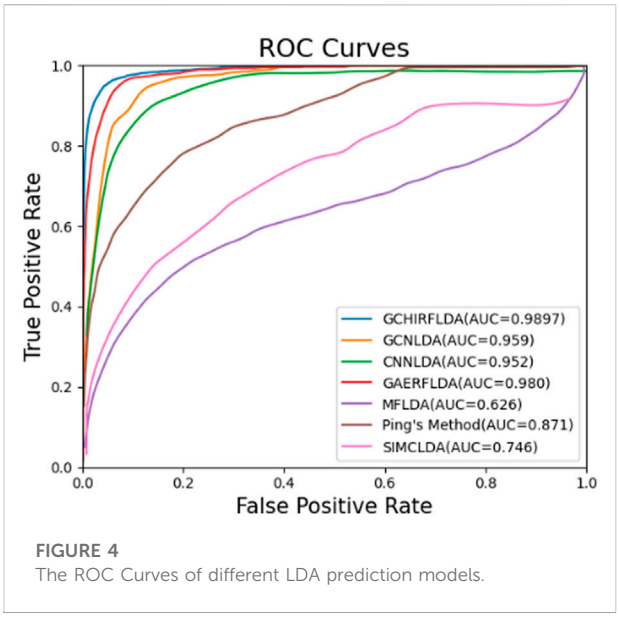

FIGURE 4
The ROC Curves of different LDA prediction models.

prediction models, including GAERF (Wu Q.-W. et al., 2021), CNNLDA (Xuan et al., 2019a), GCNLDA (Xuan et al., 2019c), MFLDA (Fu et al., 2018), Ping's method (Ping et al., 2019) and SIMLDA (Lu et al., 2018). The AUCs and AUPRs of all LDA prediction models are listed in Table 3. Figure 3 showed the ROC curves for these LDA prediction models.

From Table 3 and Figure 4, one can see that the AUC and AUPR of the GCHIRFLDA model are maximal among all LDA prediction models, which achieved 0.990 and 0.704, respectively. In term of AUC, our model achieved 0.990 which was 0.99%, 3.23%, 3.96%, 58.19%, 13.63%, and 32.67% higher than GAERF, GCNLDA, CNNLDA, MFLDA, Ping's method and SIMCLDA, respectively. In term of AUPR, our model achieved 0.704 which was 43.38%, 215.79%, 180.47%, 966.67%, 221.46%, 634.38% higher than GAERF, GCNLDA, CNNLDA, MFLDA, Ping's Method and SIMCLDA, respectively. According to the results of cross validation experiments, our GCHIRFLDA model has better LDA prediction ability.

3.4 Case studies

To further validate the prediction ability of the GCHIRFLDA model, we conducted case studies on two most common cancers, colon cancer and stomach cancer. We used the GCHIRFLDA to score all the unlabeled lncRNA-disease pairs, and selected the top 20 lncRNAs most likely to be associated with stomach cancer and colon cancer respectively according to the score. Finally, the predicted stomach cancer-associated and colon cancer-associated lncRNAs by the GCHIRFLDA model were validated by data from Lnc2Cancer v3.0 (Ning et al., 2016), LncRNADisease v2.0 (Bao et al., 2019) and some published research literature.

Colon cancer is the third most common cancer worldwide and the fourth leading cause of cancer-related death. The incidence of colon cancer has increased dramatically in China because of a shift in our habits as a society (Xue et al., 2015). In this work, we used the GCHIRFLDA to predict colon cance-associated lncRNAs. As a result, the top 20 predicted lncRNAs associated with colon cancer and the provenances of the evidence are shown in Table 4. As one can see from Table 4, 17 predicted lncRNAs have been confirmed by records included in the Lnc2Cancer (v3.0) or LncRNADisease (v2.0) or published literature. For example, Wan et al. showed that the overexpressing of CDKN2B-AS1 exhibited accelerated proliferation in colon cancer (Wan et al., 2013). Xu et al. reported the tumor

TABLE 4 The top 20 colon cancer-related lncRNA candidates predicted by the GCHIRFLDA model.

lncRNA	Rank	Evidence
CDKN2B-AS1	1	Lnc2Cancer 3.0& LncRNADisease v2.0
PVT1	2	Lnc2Cancer 3.0& LncRNADisease v2.0
UCA1	3	Lnc2Cancer 3.0& LncRNADisease v2.0
NEAT1	4	Lnc2Cancer 3.0& LncRNADisease v2.0
KCNQ1OT1	5	Lnc2Cancer 3.0
XIST	6	Lnc2Cancer 3.0& LncRNADisease v2.0
GAS5	7	Lnc2Cancer 3.0& LncRNADisease v2.0
SPRY4-IT1	8	Lnc2Cancer 3.0& LncRNADisease v2.0
MIR17HG	9	Literature (Xu et al., 2019)
TUG1	10	Lnc2Cancer 3.0& LncRNADisease v2.0
BANCR	11	Lnc2Cancer 3.0& LncRNADisease v2.0
HOTTIP	12	Lnc2Cancer 3.0& LncRNADisease v2.0
BCYRN1	13	LncRNADiseasev2.0
HNF1A-AS1	14	Lnc2Cancer 3.0
AFAP1-AS1	15	Lnc2Cancer 3.0
HULC	16	Lnc2Cancer 3.0
TUSC7	17	Lnc2Cancer 3.0
KIRREL3-AS3	18	unknown
LSINCT5	19	unknown
NPTN-IT1	20	unknown

suppressor B-cell linker (BLNK) was reduced in expression *via* MIR17HG, which resulted in an increase in invasion and migration of colorectal cancer cells (Xu et al., 2019).

In the digestive tract, stomach cancer is one of the most prevalent malignancies (Gu et al., 2017). The identification of new biomolecular markers of stomach cancer is essential for treatment and diagnosis. In this work, we used the GCHIRFLDA to predict stomach cancer-associated lncRNAs. As a result, the top 20 predicted lncRNAs associated with colon cancer and the provenances of the evidence are shown in Table 5. As seen in Table 5, 18 predicted lncRNAs have been confirmed by records included in the Lnc2Cancer (v3.0) or LncRNADisease (v2.0) or published literature. For example, Feng et al. revealed that KCNQ1OT1 inhibited stomach cancer cell progression *via* regulating miR-9 and LMX1A expression (Feng et al., 2020); Wu et al. found the high expression of lncRNA-CCAT2 indicated poor prognosis of stomach cancer and promoted cell proliferation and invasion (Wu et al., 2017). Consequently, the case studies on colon cancer and stomach cancer showed that GCHIRFLDA was an excellent predictor.

4 Conclusion

In this work, we proposed a geometric complement heterogeneous information and random forest-based approach for predicting LDAs (named GCHIRFLDA). Firstly, the potential

TABLE 5 The top 20stomach cancer-related lncRNA candidates predicted by the GCHIRFLDA model.

lncRNA	Rank	Evidence
MALAT1	1	Lnc2Cancer 3.0& LncRNADisease v2.0
XIST	2	Lnc2Cancer 3.0& LncRNADisease v2.0
NEAT1	3	Lnc2Cancer 3.0& LncRNADisease v2.0
CCAT2	4	Lnc2Cancer 3.0& LncRNADisease v2.0
TUG1	5	Lnc2Cancer 3.0& LncRNADisease v2.0
KCNQ1OT1	6	Lnc2Cancer 3.0
HOTTIP	7	Lnc2Cancer 3.0& LncRNADisease v2.0
WT1-AS	8	Lnc2Cancer 3.0& LncRNADisease v2.0
HNF1A-AS1	9	Lnc2Cancer 3.0& LncRNADisease v2.0
HULC	10	Lnc2Cancer 3.0& LncRNADisease v2.0
MIR17HG	11	Literature (Bahari et al., 2015)
CRNDE	12	Lnc2Cancer 3.0& LncRNADisease v2.0
NPTN-IT1	13	Lnc2Cancer 3.0& LncRNADisease v2.0
LINC00675	14	Lnc2Cancer 3.0
KIRREL3-AS3	15	unknown
TP53COR1	16	unknown
BCYRN1	17	Lnc2Cancer 3.0
HOTAIRM1	18	Lnc2Cancer 3.0
AFAP1-AS1	19	LncRNADisease v.2.0
LINC01133	20	Lnc2Cancer 3.0

LDA matrix is constructed by integrating the LMIs and MDAs with the original LDA matrix. Then, the Jaccard similarity and the Gaussian interaction profile similarity of lncRNA and disease are combined to represent features of lncRNA and disease. Next, a low-dimensional feature space is extracted by using autoencoder. Finally, RF is employed as the classifier to predict potential LDAs. In conclusion, the AUC and AUPR comparison with other LDA prediction models based on five-fold cross-validation and the case studies show that our model has better LDA prediction performance.

Although the GCHIRFLDA model has a good performance, it still has some limitations. Firstly, the lack of data verified by biological experimental is a big shortcoming for computational models. Secondly, randomly selecting the unknown lncRNA-disease pairs as negative samples may incorrectly classify potential positive samples as negative samples, which may affect the prediction performance. Finally, only the heterogeneous information of miRNAs is introduced in this work, and in the future, more biological information will be fused to improve the performance of the LDA prediction model.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

TZ conceived and implemented the model, performed the experiments, and wrote the paper. DY directed the research and revised the paper. XjZ and XrZ analyzed the experimental results and revised the paper. CZ performed the experiments. All authors have read and approved the final manuscript.

Funding

This work is supported by the National Natural Science Foundation of China (Grant No. 62172128) and the Postdoctoral Research Start Fund of Heilongjiang Province (LBH-Q20098). The funding body did not play any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bahari, F., Emadi-Baygi, M., and Nikpour, P. (2015). miR-17-92 host gene, uderexpressed in gastric cancer and its expression was negatively correlated with the metastasis. *Indian J. Cancer* 52, 22–25. doi:10.4103/0019-509X.175605

Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., and Dong, D. (2019). LncRNADisease 2.0: An updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res.* 47, D1034–D1037. doi:10.1093/nar/gky905

Barsyte-Lovejoy, D., Lau, S. K., Boutros, P. C., Khosravi, F., Jurisica, I., Andrulis, I. L., et al. (2006). The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis. *Cancer Res.* 66, 5330–5337. doi:10.1158/0008-5472.can-06-0037

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Cheetham, S. W., Gruhl, F., Mattick, J. S., and Dinger, M. E. (2013). Long noncoding RNAs and the genetics of cancer. *Br. J. Cancer* 108, 2419–2425. doi:10.1038/bjc.2013.233

Chen, T. Q., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proc. 22nd Acm Sigkdd Int. Conf. Knowl. Discov. Data Min. Kdd'16*, 785–794.

Chen, X., Clarence Yan, C. C., Luo, C., Ji, W., Zhang, Y., and Dai, Q. (2015). Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci. Rep.* 5, 11338. doi:10.1038/srep11338

Chen, X. (2015a). Katzlda: KATZ measure for the lncRNA-disease association prediction. *Sci. Rep.* 5, 16840. doi:10.1038/srep16840

Chen, X. (2015b). Katzlda: KATZ measure for the lncRNA-disease association prediction. *Sci. Rep.* 5. doi:10.1038/srep16840

Chen, X. (2015c). Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. *Sci. Rep.* 5, 13186. doi:10.1038/srep13186

Chen, X., Yan, C. C., Zhang, X., and You, Z. H. (2017). Long non-coding RNAs and complex diseases: From experimental results to computational models. *Brief. Bioinform* 18, 558–576. doi:10.1093/bib/bbw060

Cui, Z., Liu, J. X., Gao, Y. L., Zhu, R., and Yuan, S. S. (2020). LncRNA-disease associations prediction using bipartite local model with nearest profile-based association inferring. *IEEE J. Biomed. Health Inf.* 24, 1519–1527. doi:10.1109/jbhi.2019.2937827

Fan, Y., Chen, M., and Pan, X. (2022). Gcrflda: Scoring lncRNA-disease associations using graph convolution matrix completion with conditional random field. *Brief. Bioinform* 23, bbab361. doi:10.1093/bib/bbab361

Feng, L., Li, H., Li, F., Bei, S., and Zhang, X. (2020). LncRNA KCNQ1OT1 regulates microRNA-9-LMX1A expression and inhibits gastric cancer cell progression. *Aging* 12, 707–717. doi:10.18632/aging.102651

Fu, G., Wang, J., Domeniconi, C., and Yu, G. (2018). Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. *Bioinformatics* 34, 1529–1537. doi:10.1093/bioinformatics/btx794

Gao, M. M., Cui, Z., Gao, Y. L., Wang, J., and Liu, J. X. (2021). Multi-label fusion collaborative matrix factorization for predicting LncRNA-disease associations. *IEEE J. Biomed. Health Inf.* 25, 881–890. doi:10.1109/jbhi.2020.2988720

Gu, J., Li, Y., Fan, L., Zhao, Q., Tan, B., Hua, K., et al. (2017). Identification of aberrantly expressed long non-coding RNAs in stomach adenocarcinoma. *Oncotarget* 8, 49201–49216. doi:10.18632/oncotarget.17329

Huang, Y. A., Chen, X., You, Z. H., Huang, D. S., and Chan, K. C. (2016). Ilncsim: Improved lncRNA functional similarity calculation model. *Oncotarget* 7, 25902–25914. doi:10.18632/oncotarget.8296

Jaccard, P. (1908). Nouvelles recherches sur la Distribution florale. *Bull. Soc. Vaudoise Sci. Nat.* 44, 223–270.

Ji, C., Gao, Z., Ma, X., Wu, Q., Ni, J., and Zheng, C. (2021). Aemda: Inferring miRNA-disease associations based on deep autoencoder. *Bioinformatics* 37, 66–72. doi:10.1093/bioinformatics/btaa670

Kuhn, M. (2013). *Classification using C5. 0 UseR! 2013*. CT, USA: Pfizer Global R&D: Groton.

Lan, W., Lai, D., Chen, Q., Wu, X., Chen, B., Liu, J., et al. (2022). Ldicdl: LncRNA-disease association identification based on collaborative deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 19, 1715–1723. doi:10.1109/tcbb.2020.3034910

Lan, W., Li, M., Zhao, K., Liu, J., Wu, F. X., Pan, Y., et al. (2017). Ldap: A web server for lncRNA-disease association prediction. *Bioinformatics* 33, 458–460. doi:10.1093/bioinformatics/btw639

Liu, J. X., Cui, Z., Gao, Y. L., and Kong, X. Z. (2021). Wgrcmf: A weighted graph regularized collaborative matrix factorization method for predicting novel LncRNA-disease associations. *IEEE J. Biomed. Health Inf.* 25, 257–265. doi:10.1109/jbhi.2020.2985703

Liu, M. X., Chen, X., Chen, G., Cui, Q. H., and Yan, G. Y. (2014). A computational framework to infer human disease-associated long noncoding RNAs. *Plos One* 9. doi:10.1371/journal.pone.0084408

Lu, C., Yang, M., Luo, F., Wu, F. X., Li, M., Pan, Y., et al. (2018). Prediction of lncRNA-disease associations based on inductive matrix completion. *Bioinformatics* 34, 3357–3364. doi:10.1093/bioinformatics/bty327

Mathieu, E. L., Belhocine, M., Dao, L. T., Puthier, D., and Spicuglia, S. (2014). Rôle des longs ARN non codants dans le développement normal et pathologique. *Med. Sci. Paris.* 30, 790–796. doi:10.1051/medsci/20143008018

Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., et al. (2016). Lnc2Cancer: A manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.* 44, D980–D985. doi:10.1093/nar/gkv1094

Ping, P., Wang, L., Kuang, L., Ye, S., Iqbal, M. F. B., and Pei, T. (2019). A novel method for LncRNA-disease association prediction based on an lncRNA-disease association network. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 16, 688–693. doi:10.1109/tcbb.2018.2827373

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Netw.* 61, 85–117. doi:10.1016/j.neunet.2014.09.003

Sumathipala, M., Maiorino, E., Weiss, S. T., and Sharma, A. (2019). Network diffusion approach to predict LncRNA disease associations using multi-type biological networks: Lion. *Front. Physiol.* 10, 888. doi:10.3389/fphys.2019.00888

Sun, J., Shi, H. B., Wang, Z. Z., Zhang, C. J., Liu, L., Wang, L. T., et al. (2014). Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. *Mol. Biosyst.* 10, 2074–2081. doi:10.1039/c3mb70608g

Sun, W., Shi, Y., Wang, Z., Zhang, J., Cai, H., Zhang, J., et al. (2018). Interaction of long-chain non-coding RNAs and important signaling pathways on human cancers (Review). *Int. J. Oncol.* 53, 2343–2355. doi:10.3892/ijo.2018.4575

Taft, R. J., Pang, K. C., Mercer, T. R., Dinger, M., and Mattick, J. S. (2010). Non-coding RNAs: Regulators of disease. *J. Pathol.* 220, 126–139. doi:10.1002/path.2638

Wan, G., Mathur, R., Hu, X., Liu, Y., Zhang, X., Peng, G., et al. (2013). Long non-coding RNA ANRIL (CDKN2B-AS) is induced by the ATM-E2F1 signaling pathway. *Cell. Signal.* 25, 1086–1095. doi:10.1016/j.cellsig.2013.02.006

Wang, B., Zhang, C., Du, X.-X., and Zhang, J.-F. (2021). lncRNA-disease association prediction based on latent factor model and projection. *Sci. Rep.* 11, 19965. doi:10.1038/s41598-021-99493-5

Wang, Y., Yu, G., Wang, J., Fu, G., Guo, M., and Domeniconi, C. (2020). Weighted matrix factorization on multi-relational data for LncRNA-disease association prediction. *Methods* 173, 32–43. doi:10.1016/j.ymeth.2019.06.015

Washietl, S., Kellis, M., and Garber, M. (2014). Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.* 24, 616–628. doi:10.1101/gr.165035.113

Wu, Q.-W., Xia, J.-F., Ni, J.-C., and Zheng, C.-H. (2021a). Gaerf: Predicting lncRNA-disease associations by graph auto-encoder and random forest. *Brief. Bioinform* 22, bbaa391. doi:10.1093/bib/bbaa391

Wu, Q. W., Cao, R. F., Xia, J., Ni, J. C., Zheng, C. H., and Su, Y. (2021b). Extra trees method for predicting LncRNA-disease association based on multi-layer graph embedding aggregation. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi:10.1109/tcbb.2021.3113122

Wu, S. W., Hao, Y. P., Qiu, J. H., Zhang, D. B., Yu, C. G., and Li, W. H. (2017). High expression of long non-coding RNA CCAT2 indicates poor prognosis of gastric cancer and promotes cell proliferation and invasion. *Minerva Med.* 108, 317–323. doi:10.23736/S0026-4806.17.04703-6

Wu, X., Lan, W., Chen, Q., Dong, Y., Liu, J., and Peng, W. (2020). Inferring LncRNA-disease associations based on graph autoencoder matrix completion. *Comput. Biol. Chem.* 87, 107282. doi:10.1016/j.compbiolchem.2020.107282

Xie, H., Ma, B., Gao, Q., Zhan, H., Liu, Y., Chen, Z., et al. (2018). Long non-coding RNA CRNDE in cancer prognosis: Review and meta-analysis. *Clin. Chim. Acta* 485, 262–271. doi:10.1016/j.cca.2018.07.003

Xu, J., Meng, Q., Li, X., Yang, H., Xu, J., Gao, N., et al. (2019). Long noncoding RNA MIR17HG promotes colorectal cancer progression via miR-17-5p. *Cancer Res.* 79, 4882–4895. doi:10.1158/0008-5472.can-18-3880

Xuan, P., Cao, Y., Zhang, T., Kong, R., and Zhang, Z. (2019a). Dual convolutional neural networks with attention mechanisms based method for predicting disease-related lncRNA genes. *Front. Genet.* 10, 416. doi:10.3389/fgene.2019.00416

Xuan, P., Jia, L., Zhang, T., Sheng, N., Li, X., and Li, J. (2019b). LDAPred: A method based on information flow propagation and a convolutional neural network for the prediction of disease-associated lncRNAs. *Int. J. Mol. Sci.* 20. doi:10.3390/ijms20184458

Xuan, P., Pan, S., Zhang, T., Liu, Y., and Sun, H. (2019c). Graph convolutional network and convolutional neural network based method for predicting lncRNA-disease associations. *Cells* 8. doi:10.3390/cells8091012

Xuan, P., Sheng, N., Zhang, T., Liu, Y., and Guo, Y. (2019d). Cnndlp: A method based on convolutional autoencoder and convolutional neural network with adjacent edge attention for predicting lncRNA-disease associations. *Int. J. Mol. Sci.* 20. doi:10.3390/ijms20174260

Xue, Y., Ma, G. X., Gu, D. Y., Zhu, L. J., Hua, Q. H., Du, M. L., et al. (2015). Genome-wide analysis of long noncoding RNA signature in human colorectal cancer. *Gene* 556, 227–234. doi:10.1016/j.gene.2014.11.060

Yao, D., Zhan, X., Zhan, X., Kwoh, C. K., Li, P., and Wang, J. (2020). A random forest based computational model for predicting novel lncRNA-disease associations. *BMC Bioinforma.* 21, 126. doi:10.1186/s12859-020-3458-1

Ye, J., Chow, J.-H., Chen, J., and Zheng, Z. (2009). "Stochastic gradient boosted distributed decision trees," in *Proceedings of the 18th ACM conference on Information and knowledge management* (Hong Kong, China: Association for Computing Machinery). doi:10.1145/1645953.1646301

Yin, M. M., Liu, J. X., Gao, Y. L., Kong, X. Z., and Zheng, C. H. (2022). Ncplp: A novel approach for predicting microbe-associated diseases with network consistency projection and label propagation. *IEEE Trans. Cybern.* 52, 5079–5087. doi:10.1109/tcyb.2020.3026652

Yu, G., Fu, G., Lu, C., Ren, Y., and Wang, J. (2017). Brwlda: Bi-random walks for predicting lncRNA-disease associations. *Oncotarget* 8, 60429–60446. doi:10.18632/oncotarget.19588

Yu, G., Wang, Y., Wang, J., Domeniconi, C., Guo, M., and Zhang, X. (2020). Attributed heterogeneous network fusion via collaborative matrix tri-factorization. *Inf. Fusion* 63, 153–165. doi:10.1016/j.inffus.2020.06.012

Zhang, C., Lei, X. J., and Liu, N. (2021). Predicting metabolite-disease associations based on LightGBM model. *Front. Genet.* 12, 660275. doi:10.3389/fgene.2021.660275

Zhou, J. R., You, Z. H., Cheng, L., and Ji, B. Y. (2021). Prediction of lncRNA-disease associations via an embedding learning HOPE in heterogeneous information networks. *Mol. Ther. - Nucleic Acids* 23, 277–285. doi:10.1016/j.omtn.2020.10.040

Zhou, M., Wang, X. J., Li, J. W., Hao, D. P., Wang, Z. Z., Shi, H. B., et al. (2015). Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. *Mol. Biosyst.* 11, 760–769. doi:10.1039/c4mb00511b