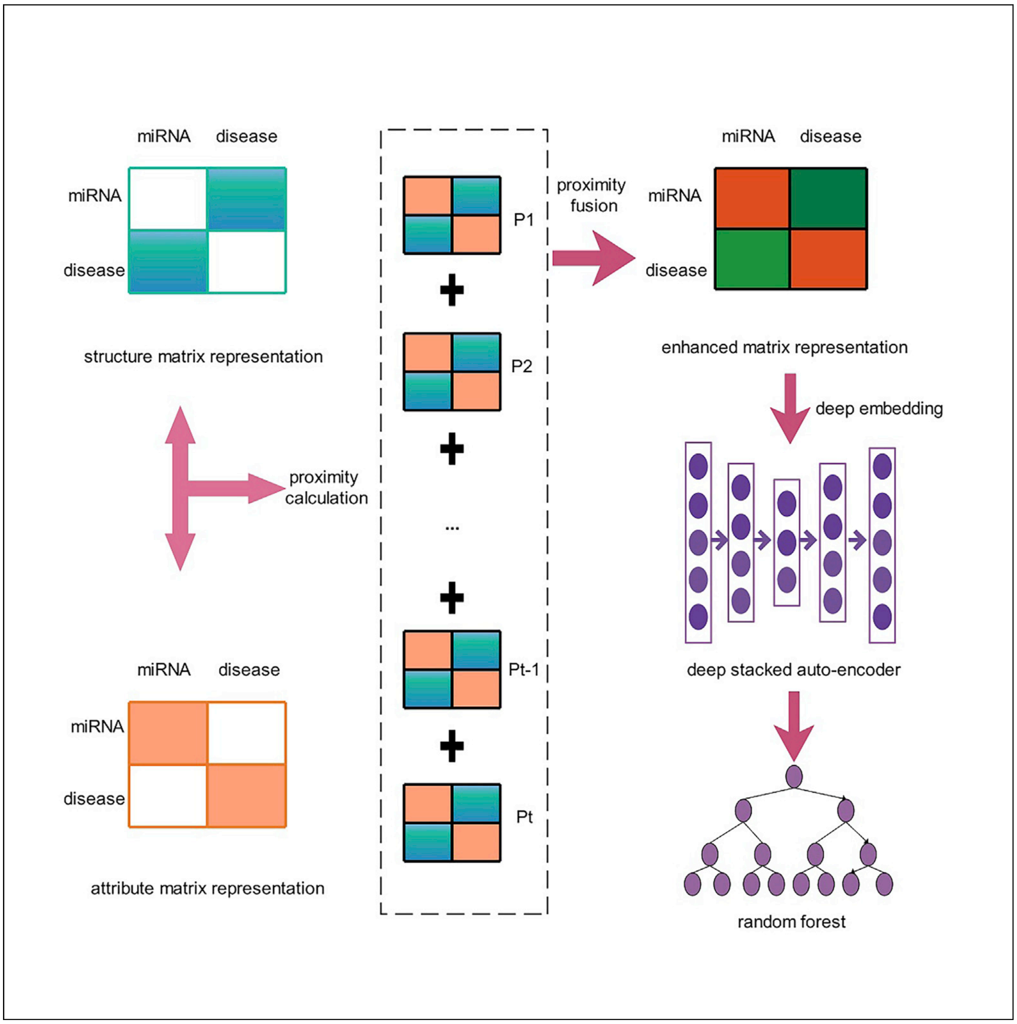
iScience

CellPress OPEN ACCESS

Article

DANE-MDA: Predicting microRNA-disease associations via deep attributed network embedding



Bo-Ya Ji, Zhu-Hong You, Yi Wang, Zheng-Wei Li, Leon Wong

zhuhongyou@ms.xjb.ac.cn

Highlights

A computational machine learning-based method for miRNA-disease association prediction

Preserve structure and attribute features via deep attributed network embedding

Capture the interaction between two kinds of features from diverse degrees of proximity

Extract the higher-order features via deep stacked auto-encoder neural network

Ji et al., iScience *24*, 102455
June 25, 2021 © 2021 The Authors.
https://doi.org/10.1016/j.isci.2021.102455

iScience

CellPress OPEN ACCESS

Article

DANE-MDA: Predicting microRNA-disease associations via deep attributed network embedding

Bo-Ya Ji,[1,2,3] Zhu-Hong You,[1,2,3,5,*] Yi Wang,[1,3] Zheng-Wei Li,[4] and Leon Wong[1,2,3]

SUMMARY

Predicting the microRNA-disease associations by using computational methods is conductive to the efficiency of costly and laborious traditional bio-experiments. In this study, we propose a computational machine learning-based method (DANE-MDA) that preserves integrated structure and attribute features via deep attributed network embedding to predict potential miRNA-disease associations. Specifically, the integrated features are extracted by using deep stacked auto-encoder on the diverse orders of matrixes containing structure and attribute information and are then trained by using random forest classifier. Under 5-fold cross-validation experiments, DANE-MDA yielded average accuracy, sensitivity, and AUC at 85.59%, 84.23%, and 0.9264 in term of HMDD v3.0 dataset, and 83.21%, 80.39%, and 0.9113 in term of HMDD v2.0 dataset, respectively. Additionally, case studies on breast, colon, and lung neoplasms related disease show that 47, 47, and 46 of the top 50 miRNAs can be predicted and retrieved in the other database.

INTRODUCTION

The human genomes have various endogenous ''non-messenger'' or ''non-coding'' RNAs, including a large number of single-stranded microRNAs (miRNAs) containing about 22 nucleotides (Ambros, 2001, 2004). miRNAs play a significant function in various human life processes, including virus defense, tissue development, cell metabolism, and organ formation, and participate in the regulation of post-transcriptional gene expression (Cui et al., 2006; Karp and Ambros, 2005; Lu et al., 2005; Rupaimoole and Slack, 2017; Xu et al., 2004). Furthermore, miRNAs also have a particular therapeutic impact as a regulator for several genes (Ling et al., 2013; Matsui and Corey, 2017). A cascade of studies have shown that miRNAs can become drug targets for human disease treatments (Mishra et al., 2020), hence it is not surprising that predicting and identifying potential miRNAs related to corresponding diseases have been the focus of researchers. For example, Jeong et al. (Jeong et al., 2011) stated that let-7a is under-expressed in the tissues and cells of patients with NSCLC (non-small cell lung cancer) compared with the normal control group. Bang et al. (2012) found that the miR-23/27/24 cluster is related to retinal vascular development and endothelial cell apoptosis and angiogenesis in cardiac ischemia. In recent years, massive miRNA-disease associations have been acquired through traditional biological experiments and stored in public databases. These biological experimental methods usually have high prediction accuracy; nevertheless, their processes are complex, expensive, and time-consuming (Liang et al., 2019). To this end, to accelerate the verification process, and reduce the time consumption and blindness of biological experiments, it is significant to establish computational methods for quickly and effectively predicting possible miRNA-disease associations (Wong et al., 2020; Yi et al., 2020).

Taking advantage of the hypothesis that functionally related miRNAs are more likely to be related to diseases with similar phenotypes, some score function-based computational models have been proposed for predicting miRNA-disease associations, which commonly leverage methods such as random walk to calculate the likelihood of potential associations on the constructed miRNA-disease association network. For example, Chen et al. (2012) first incorporated known miRNA-disease associations and large-scale miRNA-miRNA functional similarity information and then utilized the random walk and global network similarity measure methods to obtain superior performance than previous models. Luo et al. (2017) assessed the similarity between diseases or miRNAs by incorporating several relevant heterogeneous information.

[1] Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China

[2] University of the Chinese Academy of Sciences, Beijing 100049, China

[3] Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi 830011, China

[4] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

[5] Lead contact

*Correspondence: zhuhongyou@ms.xjb.ac.cn

https://doi.org/10.1016/j.isci.2021.102455

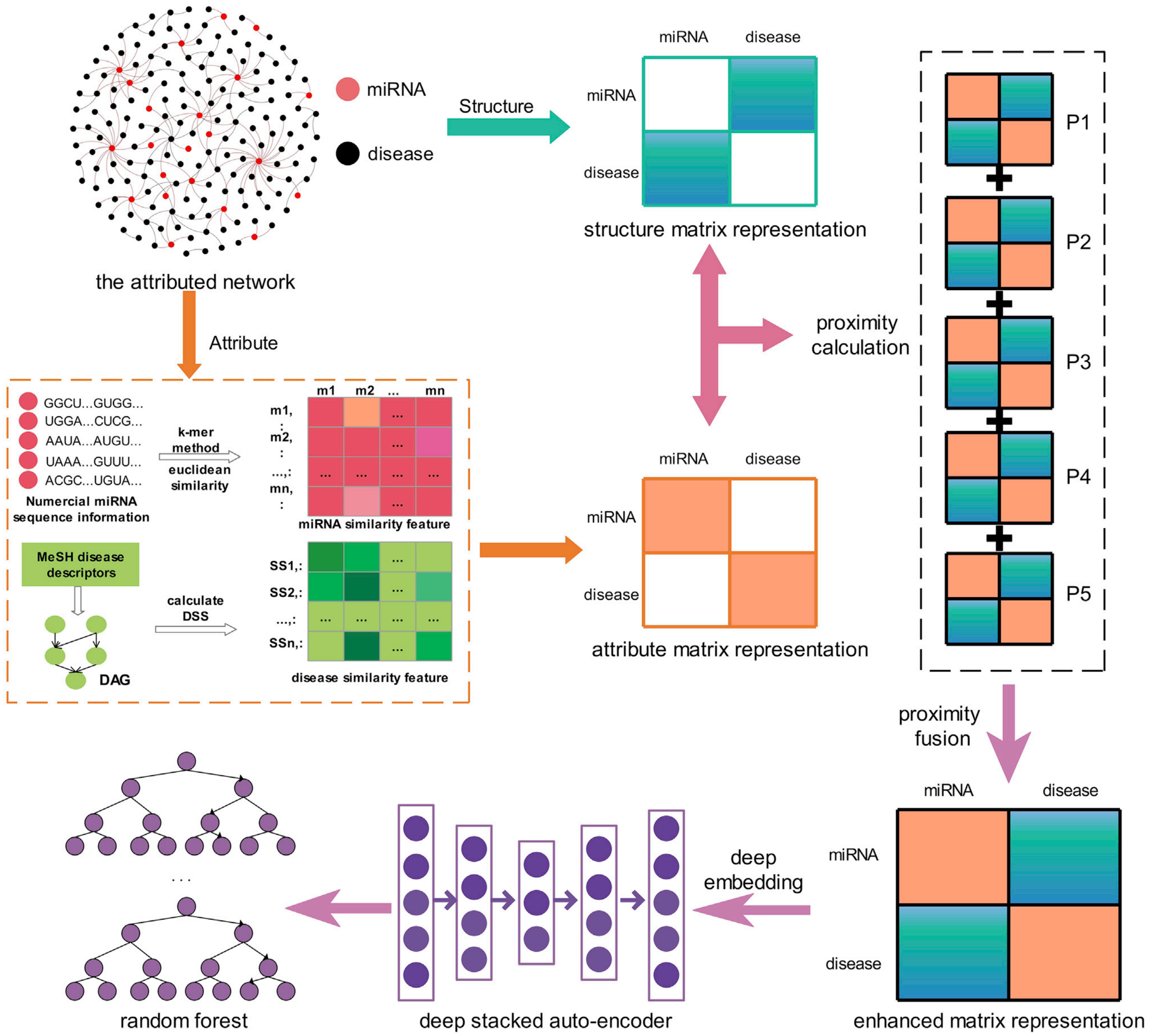


Figure 1. Illustration of the overall framework of DANE-MDA (DAG: directed acyclic graph; DSS: disease semantic similarity)

Then, a semi-supervised mechanics of Kronecker regularized least squares was employed to predict possible miRNAs related to diseases. Wang et al. (2019) utilized the logical trees classifier and fused the known miRNA-disease association, miRNA functional similarity and sequence information, and disease semantic similarity to predict miRNA-disease associations. Empirical results of cross-validation experiments and case studies both demonstrated the reliability and effectiveness of their model. Alaimo et al. (2014) adopted a recommendation algorithm to predict novel associations between miRNAs and diseases based on a tripartite network composed of miRNAs, targets, and diseases, where the targets act as intermediate nodes between miRNAs and diseases. On this basis, a multi-level resource transfer method was employed to compute the correlation degree between each miRNA-disease pair.

Recently, machine learning and deep learning also have been utilized for predicting possible associations between miRNAs and diseases with the growth of known miRNA-disease association data. For example, Xu et al. (2011) calculated four topological features of miRNAs and then trained the gold-standard miRNA dataset using the support vector machine (SVM) for predicting possible miRNA-disease associations. To

Table 1. The results of DANE-MDA under 5-fold cross-validation based on the HMDD v3.0 dataset

Fold	ACC.(%)	AUC(%)	Sen.(%)	Prec.(%)	Spec.(%)	MCC(%)
0	85.10	92.56	83.32	86.40	86.88	70.25
1	85.94	92.89	84.57	86.95	87.31	71.91
2	85.38	92.32	83.48	86.78	87.28	70.81
3	85.59	92.80	84.88	86.11	86.31	71.19
4	85.96	92.66	84.89	86.74	87.02	71.93
Average	**85.59 ± 0.37**	**92.64 ± 0.22**	**84.23 ± 0.77**	**86.60 ± 0.34**	**86.96 ± 0.41**	**71.22 ± 0.72**

The last line represents the average and standard deviation of each indicator.

break the restriction of previous models that cannot be applied for diseases without any known associated miRNAs, Chen and Yan (2014) exploited the least-squares regularization and semi-supervised learning method to reveal the miRNA-disease associations and obtain reliable performance. These existing models almost utilized miRNA functional similarity, miRNA-family associations, disease semantic similarity, miRNA-target associations, and known miRNA-disease associations. However, the known miRNA-disease associations are not well mined. These known miRNA-disease associations can be constructed as a graph or network, but the node features in the graph are rarely calculated. Therefore, some of the recent techniques in graph embedding are used for predicting miRNA-disease associations, such as graph convolutional networks (Kipf and Welling, 2016), matrix factorization (He et al., 2018, 2019), and Bayesian learning (Hu et al., 2019). For example, Xuan et al. (2019) utilized convolutional neural networks and network representation learning to design a computational model to predict miRNA-disease associations. Zheng et al. (2020a) exploited the graph embedding method and random forest classifier to reveal novel miRNA and disease associations. Their method gained good performance by combining the behavior and attribute features of diseases and miRNAs.

In this study, we propose a computational machine learning-based method (DANE-MDA) that attempts to preserve both the diverse degrees of network structure and attribute feature of miRNAs and diseases via deep attributed network embedding to predict potential miRNA-disease associations. DANE-MDA includes four steps. First, we constructed an attributed network by connecting the known miRNA-disease associations in the Human MicroRNA Disease Database (HMDD) and, respectively, calculated the attribute and network structure feature of miRNAs and diseases, where the attribute feature includes miRNA sequence similarity and disease semantic similarity and the network structure feature includes the probability of direct transition between each miRNA-disease association pair. Second, we captured the interactions between network structure and attribute information of miRNAs and diseases from diverse degrees of proximity by utilizing a personalized random walk-based method. Third, we fused the various degrees of proximity to build an enhanced matrix representation, which contains both the attribute feature, as well as the local and global network structure feature of miRNAs and diseases and then exploited the deep stacked auto-encoder to learn the complex and nonlinear information in the enhanced matrix to represent miRNAs and diseases. Finally, the Random Forest classifier is selected to construct the prediction model. The illustration of the DANE-MDA overall framework is shown in Figure 1. As a result, the 5-fold cross-validation experiment was applied to examine the performance of DANE-MDA, which obtained an average 85.59% accuracy, 84.23% sensitivity, and 0.9264 area under the

Table 2. The results of DANE-MDA under 5-fold cross-validation based on the HMDD v2.0 dataset

Fold	ACC.(%)	AUC(%)	Sen.(%)	Prec.(%)	Spec.(%)	MCC(%)
0	84.53	92.22	79.65	88.27	89.41	69.39
1	81.86	90.17	79.56	83.40	84.16	63.79
2	83.89	91.48	80.02	86.73	87.75	67.98
3	83.93	91.17	81.49	85.67	86.37	67.94
4	81.86	90.61	81.22	82.28	82.50	63.73
Average	**83.21 ± 1.26**	**91.13 ± 0.79**	**80.39 ± 0.90**	**85.27 ± 2.44**	**86.04 ± 2.76**	**66.57 ± 2.63**

The last line represents the average and standard deviation of each indicator.

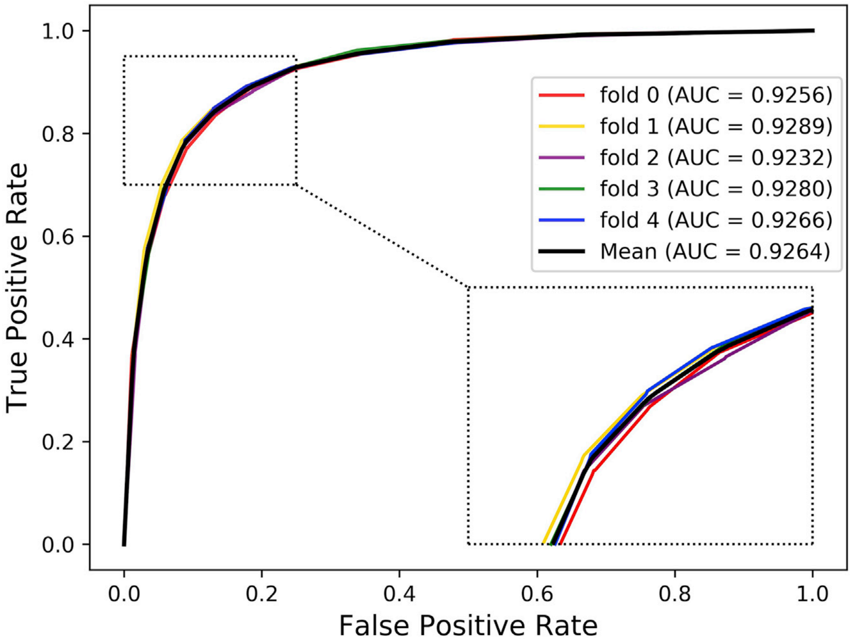


Figure 2. The ROC curves of DANE-MDA under 5-fold cross validation based on HMDD v3.0 dataset

receiver operating characteristic (ROC) curve (AUC) on the HMDD v3.0 dataset, and an average 83.21% accuracy, 80.39% sensitivity, and 0.9113 AUC on the HMDD v2.0 dataset. What's more, we also conducted case studies on three common human diseases, including breast, colon, and lung neoplasms, to verify the performance of DANE-MDA in practical applications. Additionally, we also compared the influence of model parameters and classifiers on prediction results. In summary, the proposed DANE-MDA model has a promising performance for predicting novel miRNA-disease associations and is anticipated to be an effective supplement tool in the field of bioinformatics research.

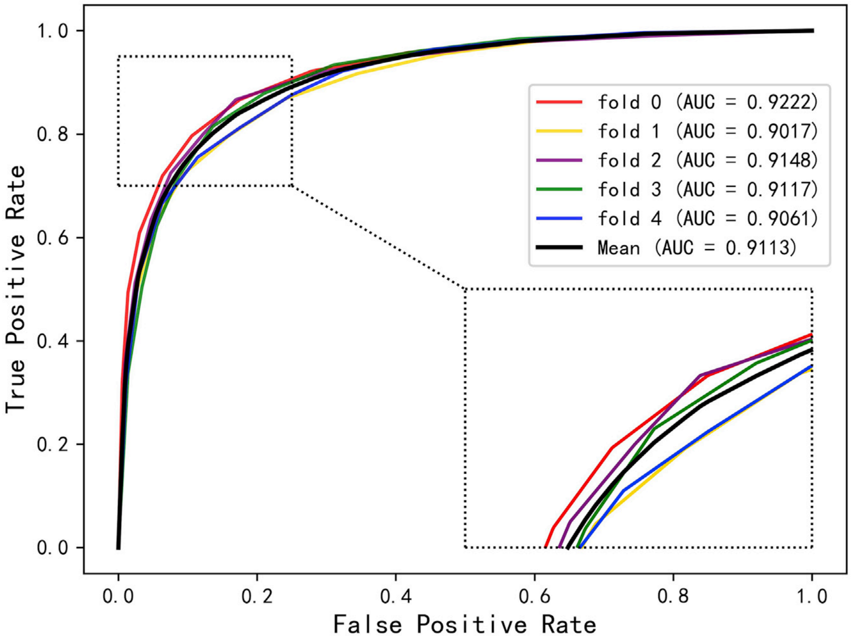


Figure 3. The ROC curves of DANE-MDA under 5-fold cross validation based on HMDD v2.0 dataset

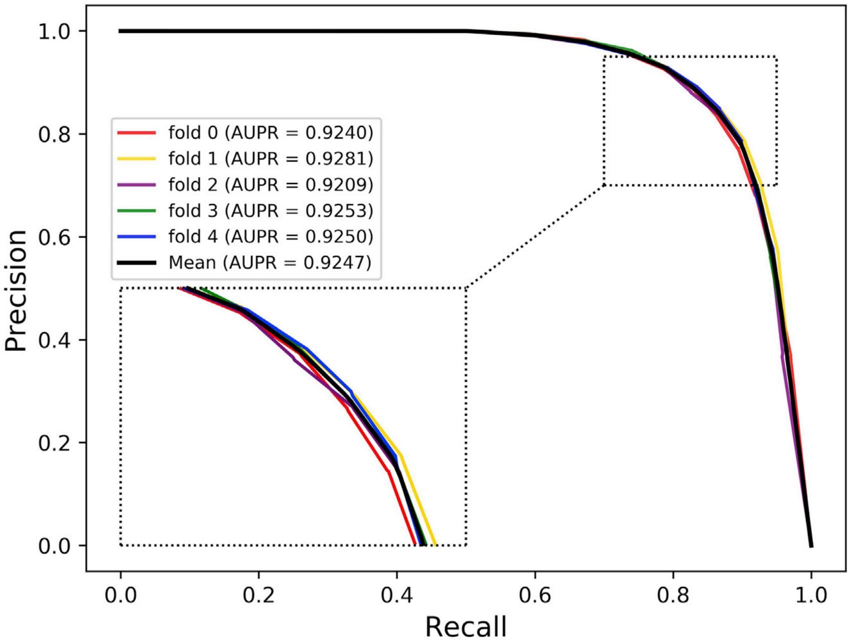


Figure 4. The PR curves of DANE-MDA under 5-fold cross validation based on HMDD v3.0 dataset

RESULTS

The results of DANE-MDA under 5-fold cross-validation experiment

Cross-validation is a common method for building models and verifying model parameters in machine learning (Cooil et al., 1987). In this study, the 5-fold cross-validation experiment is implemented to evaluate the ability of DANE-MDA for predicting novel miRNA-disease associations. Specifically, the positive and negative samples are, respectively, separated into five folds, one fold is the test dataset and the rest

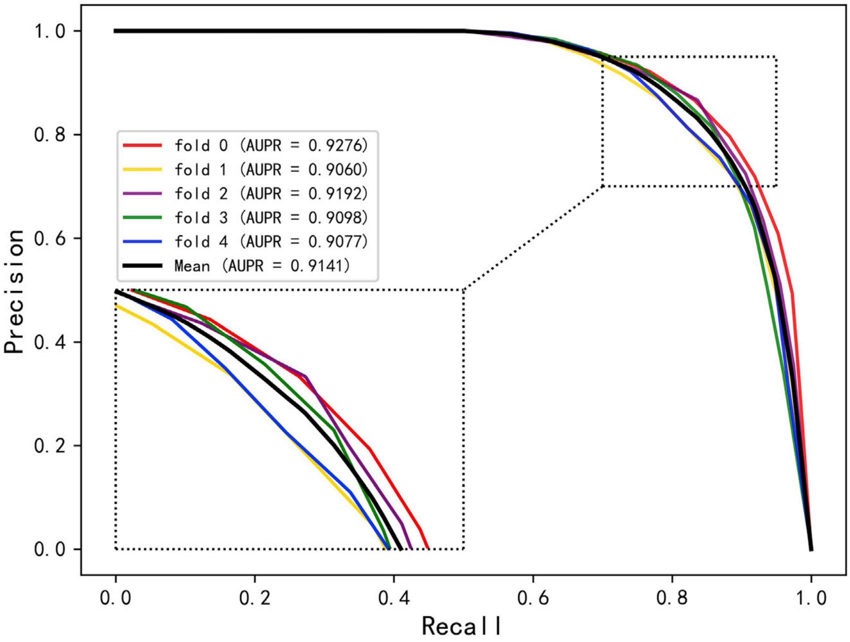


Figure 5. The PR curves of DANE-MDA under 5-fold cross validation based on HMDD v2.0 dataset

Table 3. The AUC values of parameter α under each fold cross-validation (β = 0.94, t = 5)

Fold α	0	1	2	3	4	Average
1	0.9169	0.9224	0.9149	0.9223	0.9171	0.9187 ± 0.34
0.95	0.9242	0.9263	0.9206	0.9269	0.9252	0.9246 ± 0.25
0.90	0.9211	0.9272	0.9230	0.9286	0.9215	0.9243 ± 0.34
0.85	**0.9256**	**0.9289**	**0.9232**	**0.9280**	**0.9266**	**0.9264 ± 0.22**
0.80	0.9271	0.9277	0.9243	0.9270	0.9241	0.9261 ± 0.17
0.75	0.9262	0.9299	0.9224	0.9250	0.9261	0.9259 ± 0.27
0	0.8774	0.8849	0.8776	0.8791	0.8746	0.8787 ± 0.38

four folds are the training dataset. On this basis, five experiments are respectively performed in sequence. In the results, six evaluation indicators in each fold experiment including Accuracy (Acc.), Precision (Prec.), Matthews Correlation Coefficient (MCC), Specificity (Spec.), Sensitivity (Sen.), and the AUC based on the HMDD v3.0 and v2.0 dataset are, respectively, recorded in Tables 1 and 2. Furthermore, the ROC and precision-recall (PR) curve is further selected to verify the prediction ability of DANE-MDA. Figures 2, 3, 4, and 5 respectively show the 5-fold cross-validation ROC and PR curves of DANE-MDA based on the HMDD v3.0 and v2.0, which, respectively, draws the sensitivity (true positive rate) against the specificity (false positive rate) and the precision against the recall under various score thresholds.

The impact of model parameters on prediction results

In this part, we quantitatively analyzed the influence of the parameters in DANE-MDA on the prediction performance, including α, β, and t. Respectively, to fuse the network structure feature and attribute information of miRNAs and diseases, we introduced the weight parameter α to represent the preference ratio between attribute and structural information, with a value between 0 and 1. When $\alpha = 1$, the predictive ability of DANE-MDA entirely depends on the structure information, and when $\alpha = 0$, the predictive ability of DANE-MDA entirely depends on the attribute information. Moreover, the parameter t is introduced to capture global network structure information. Intuitively, the larger the value of t, the more global structure information will be obtained. However, when t gradually increases, the global information obtained gradually becomes weaker, and excess noise information will cause the prediction results to decrease. Last, because the low-order network structure feature is more influential than the high-order ones, we introduced the parameter β to control the downtrend of higher-order information, with a value between 0 and 1. On this basis, we, respectively, selected the following parameters to perform 5-fold cross-validation:$\alpha \in \{1, 0.95, 0.90, 0.85, 0.80, 0.75, 0\}$, $\beta \in \{0.98, 0.96, 0.94, 0.92, 0.90\}$,$t \in \{1, 3, 5, 7, 9\}$ and used the AUC value as the evaluation indicator. For each parameter, other parameters and the experimental environment are controlled to be consistent. Tables 3, 4, and 5, respectively, show the distribution of the AUC values for each cross-validation. Additionally, the line curve of the mean AUC value was shown in Figures 6, 7, and 8. In the results, for parameter α, when $\alpha = 0.85$ (fusion of 85% network structure and 15% attribute feature), DANE-MDA obtains the best performance. For parameter β, when $\beta = 0.94$, DANE-MDA has the best control over the downward trend of high-order features. For parameter t, when $t = 5$, DANE-MDA obtains the optimal global structural features.

Furthermore, to further describe the effectiveness of our feature fusion strategy, we displayed the performance of DANE-MDA with three different feature combinations under the 5-fold cross-validation: only

Table 4. The AUC values of parameter β under each fold cross-validation (α = 0.85, t = 5)

Fold β	0	1	2	3	4	Average
0.98	0.9274	0.9253	0.9208	0.9275	0.9222	0.9246 ± 0.30
0.96	0.9249	0.9312	0.9252	0.9279	0.9222	0.9263 ± 0.34
0.94	**0.9256**	**0.9289**	**0.9232**	**0.9280**	**0.9266**	**0.9264 ± 0.22**
0.92	0.9249	0.9252	0.9221	0.9291	0.9243	0.9251 ± 0.25
0.90	0.9234	0.9268	0.9238	0.9279	0.9224	0.9249 ± 0.24

Table 5. The AUC values of parameter *t* under each fold cross-validation ()

Fold						
t	0	1	2	3	4	Average
1	0.9247	0.9260	0.9210	0.9290	0.9193	0.9240 ± 0.39
3	0.9255	0.9286	0.9236	0.9250	0.9249	0.9255 ± 0.19
5	**0.9256**	**0.9289**	**0.9232**	**0.9280**	**0.9266**	**0.9264 ± 0.22**
7	0.9234	0.9282	0.9213	0.9307	0.9223	0.9252 ± 0.41
9	0.9264	0.9277	0.9202	0.9292	0.9234	0.9254 ± 0.36

attribute features of miRNAs and diseases (α = 0), only network structure features of miRNAs and diseases (α = 1), and the fusion feature of attribute and structure information (α = 0.85). The detailed average prediction results were shown in Table 6. Additionally, Figure 9 showed the ROC and PR curves of the comparative experiment. The empirical results further proved the better performance of our feature fusion strategy.

The impact of the classifier on prediction results

For a specific classification problem, it is crucial to choose a suitable classifier. In this part, we selected four commonly used classifiers for comparison, including Naive Bayes (NB) (Rish, 2001), Adaptive Boosting (AdaBoost) (Margineantu and Dietterich, 1997), K-Nearest Neighbors (KNN) (Denoeux, 2008), and Random Forest (RF) (Liaw and Wiener, 2002), and then used the most suitable classification algorithm to build the prediction model according to the final prediction effect. To make the comparison experiment fair and easy to operate, we kept the experimental environment consistent and performed 5-fold cross-validation for different classifiers with default parameters. Finally, the average results and standard deviations of each classifier under 5-fold cross-validation were recorded in Table 7. Moreover, the ROC and PR curves of the classifier comparison experiment are shown in Figure 10. All the experiments proved that the Random Forest classifier achieved better prediction results and was more suitable for our training model.

Comparison of previous related works

In the field of potential miRNA-disease association prediction, a lot of excellent computational methods have been developed. To confirm the superiority of our model, we further compared the prediction performance of DANE-MDA based on the HMDD v3.0 with five previous state-of-the-art computational methods, including WBSMDA (Chen et al., 2016), PBMDA (You et al., 2017), HDMP (Xuan et al., 2013),

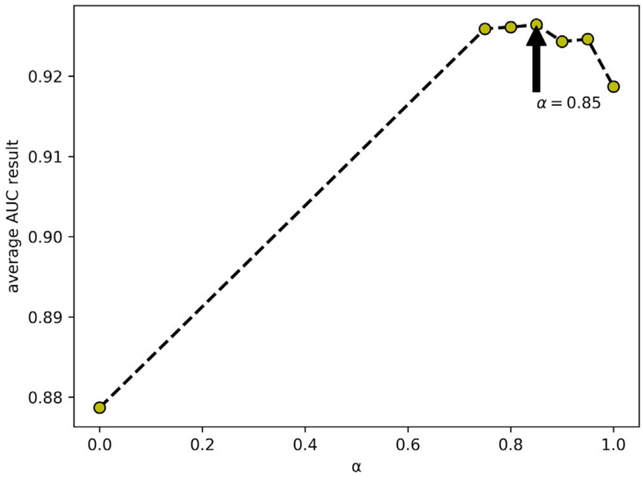


Figure 6. The line graph of average AUC results at different α values of DANE-MDA

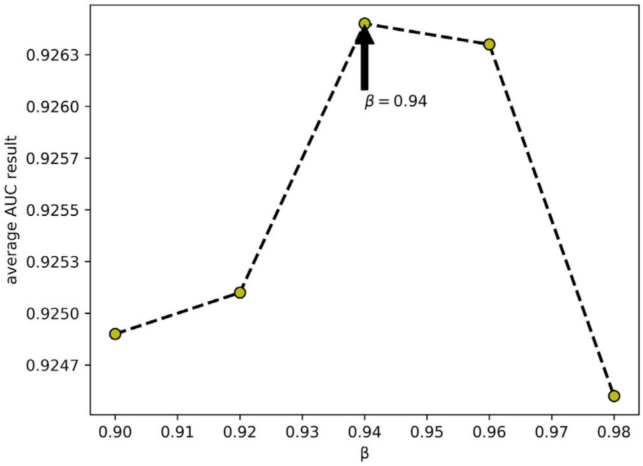


Figure 7. The line graph of average AUC results at different β values of DANE-MDA

RLSMDA (Chen and Yan, 2014), and DBMDA (Zheng et al., 2020b). WBSMDA predicts the potential associations between miRNAs and diseases by utilizing a model of within and between scores. PBMDA is a path-based prediction method by incorporating multiple similarities of miRNAs and diseases. HDMP is a weighted k-most similar neighbors-based miRNA-disease association prediction method, which is a representative method in this field. RLSMDA is a global, semi-supervised, and regularized least squares-based prediction method. DBMDA utilizes the chaos game representation method based on miRNA sequences and infers global similarity from regional distances to predict miRNA-disease associations. All these methods utilized the known miRNA-disease associations in HMDD v3.0 as the dataset and were verified with the 5-fold cross-validation experiment. Hence, we adopted the average AUC value reported in their article as the evaluation index, as shown in Table 8. Moreover, we also compared the prediction performance of DANE-MDA based on the HMDD v2.0 with the following latest four models, which have been confirmed to achieve excellent prediction accuracy, including TLHNMDA (Chen et al., 2018a), NCMCMDA (Chen et al., 2021), RFMDA (Chen et al., 2018b), and MDHGI (Chen et al., 2018c). Here we also computed the average AUC under the 5-fold cross-validation as the evaluative criterion, and greater AUC means the model shows more accurate prediction performance. Table 9 clearly shows that DANE-MDA achieved better AUC performance under the 5-fold cross-validation based on the HMDD v2.0 dataset. In short, we can clearly observe that DANE-MDA performs better than the current model in potential miRNA and disease association predictions under the 5-fold cross-validation based on both the HMDD v3.0 and v2.0 datasets.

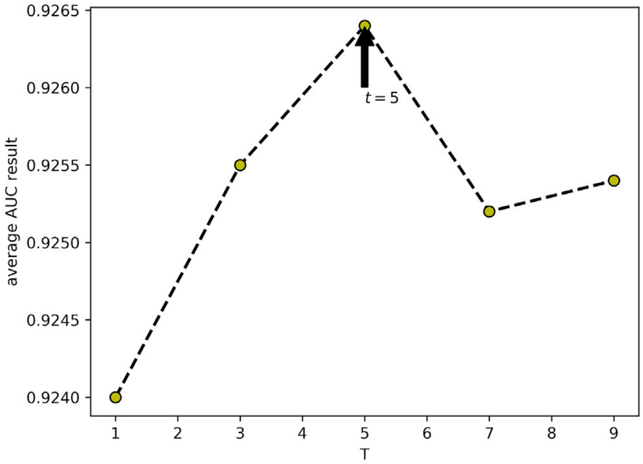


Figure 8. The line graph of average AUC results at different t values of DANE-MDA

Table 6. The average results and standard deviations of DANE-MDA with different feature combinations under 5-fold cross-validation

Feature	Acc.(%)	AUC(%)	Sen.(%)	Prec.(%)	Spec.(%)	MCC(%)
Only attribute	81.01 ± 0.28	87.87 ± 0.38	81.86 ± 0.91	80.49 ± 0.37	80.15 ± 0.63	62.03 ± 0.58
Only structure	84.76 ± 0.21	91.87 ± 0.34	83.39 ± 0.39	85.75 ± 0.31	86.14 ± 0.38	69.55 ± 0.42
Fusion	85.59 ± 0.37	92.64 ± 0.22	84.23 ± 0.77	86.60 ± 0.34	86.96 ± 0.41	71.22 ± 0.72

Case studies

In this part, to evaluate the capability of DANE-MDA for predicting potential miRNA-disease associations in practical applications, case studies were conducted on breast neoplasms, colon neoplasms, and lung neoplasms. First, all known and the same number of randomly constructed unknown miRNA-disease associations were constituted as the training samples. Second, the test samples of miRNA-corresponding disease association pairs were, respectively, constituted. It should be noted that the association pairs that already existed in the training samples have been deleted from the test samples. Finally, DANE-MDA was trained based on the training dataset, and then the association probability of unknown miRNA-disease pairs in the test dataset was predicted. On this basis, we listed the top 50 association pairs according to the prediction scores and confirmed them in two other authoritative databases, miR2Disease (Jiang et al., 2008) and dbDEMC (Yang et al., 2010).

Colon neoplasms are the third leading cause of cancer-related deaths in the United States (Siegel et al., 2016). It is a malignant tumor arising from the inner wall of the large intestine (colon) or rectum. The common risk factors for colon neoplasms include colon polyps, family history, age, African American race, and long-standing ulcerative colitis. miRNAs play an essential part in the carcinogenesis and development of colon neoplasms, and their biomarkers have great advantages in the recurrence prediction, diagnosis, and treatment. In this article, DANE-MDA was used to predict the possible miRNAs related to colon neoplasms, and 47 of the top 50 miRNAs with the highest final prediction score were verified as shown in Table 10.

Breast neoplasms are the most common non-skin malignant tumor in women. In almost all cases it occurs in women, but men can also get breast neoplasms (Bray et al., 2018; Kelsey and Horn-Ross, 1993; Tao et al., 2015). It can begin in different parts of the breast and spread outside the breast through blood and lymph vessels. In addition, more and more studies have shown that miRNAs are a new tool for the prognosis and diagnosis of patients with breast neoplasms. Hence, the prediction of potential breast neoplasms-related miRNAs may identify a novel candidate miRNA for early diagnosis and prevention of breast cancer. In this article, DANE-MDA was used to predict possible miRNAs related to breast neoplasms, and 47 of the top 50 miRNAs with the highest final prediction score were verified as shown in Table 11.

Lung neoplasms are the leading cause of cancer deaths in men and women. It is usually formed in air passage cells or lung tissue. Factors affecting lung neoplasms mainly include smoking, secondhand smoke, family history of lung cancer, air pollution, HIV infection, etc., among which smoking is the most important

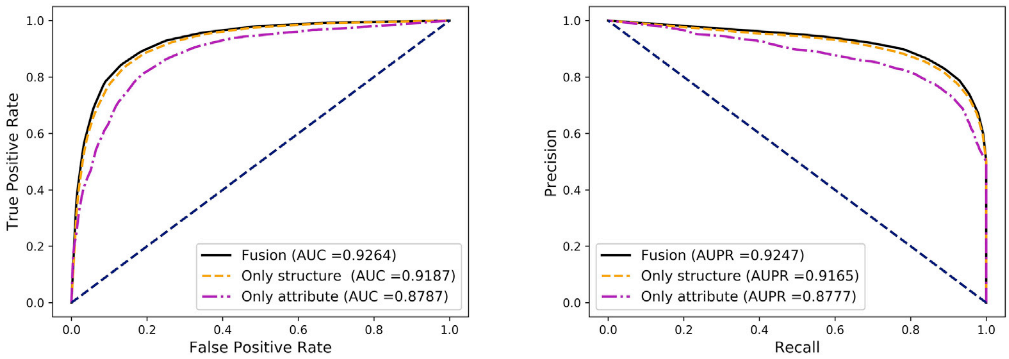


Figure 9. The average ROC and PR curves of DANE-MDA with different feature combinations under 5-fold cross-validation

Table 7. The average results and standard deviations of DANE-MDA with different classifiers under 5-fold cross-validation

Classifier	ACC.(%)	AUC(%)	Sen.(%)	Prec.(%)	Spec.(%)	MCC(%)
KNN	82.69 ± 0.30	89.68 ± 0.39	91.39 ± 0.39	77.85 ± 0.27	74.00 ± 0.35	66.39 ± 0.61
Naive Bayes	78.02 ± 0.44	79.57 ± 0.33	91.77 ± 0.43	71.97 ± 0.35	64.27 ± 0.46	58.28 ± 0.90
AdaBoost	83.56 ± 0.58	91.47 ± 0.22	85.41 ± 0.75	82.36 ± 0.68	81.70 ± 0.83	67.16 ± 1.16
RandomForest	85.59 ± 0.37	92.64 ± 0.22	84.23 ± 0.77	86.60 ± 0.34	86.96 ± 0.41	71.22 ± 0.72

risk factor for lung neoplasms (Torre et al., 2016). miRNAs have been determined to play a key role in the treatment and development of lung neoplasms. Compared with normal tissues, the expression level of miRNA in lung cancer cells and the blood of patients with lung cancer are unregulated. Moreover, the phenotype of lung cancer can be changed by regulating miRNA expression both *in vivo* and *in vitro*. In this article, DANE-MDA was used to predict possible miRNAs related to lung neoplasms, and 46 of the top 50 miRNAs with the highest final prediction score were verified as shown in Table 12.

DISCUSSION

Recently, an increasing number of researches have demonstrated that miRNAs could fulfill a variety of biological functions, and their abnormal expression or function may cause various human diseases. Thus, the prediction of potential miRNA-disease associations will significantly contribute to the treatment and investigation of complex human diseases. Otherwise, traditional biological experiments are generally laborious and expensive, which leads to a very limited number of experimentally verified miRNA-disease associations. In this study, we propose a computational machine learning-based method (DANE-MDA) that preserves integrated structure and attribute features via deep attributed network embedding and the deep stacked auto-encoder neural network to predict potential miRNA-disease associations. Specifically, the DANE-MDA framework is composed of four steps. First, the network structure and attribute feature of diseases and miRNAs is respectively calculated. Second, the interactions between network structure and attribute information of miRNAs and diseases from diverse degrees of proximity are captured by utilizing a personalized random walk-based method. Third, we fuse the diverse degrees of proximity to build an enhanced matrix representation to preserve both the attribute information and the local and global network structure features and then utilized the deep stacked auto-encoder to learn the complex nonlinear information of the enhanced matrix to represent miRNAs and diseases. Finally, the potential miRNA-disease association prediction approach is built based on the Random Forest classifier. The prediction results under 5-fold cross-validation confirmed the excellent capability of DANE-MDA. Moreover, we also discussed the influence of parameters and classifiers on the final prediction results. Last, the case studies performed on three complex human diseases once again demonstrated the good property of DANE-MDA in practical applications.

Limitations of the study

There are still some limitations in the current method that should to be addressed. First, in terms of attribute feature extraction, we hope to make full use of various information in the future, such as miRNA

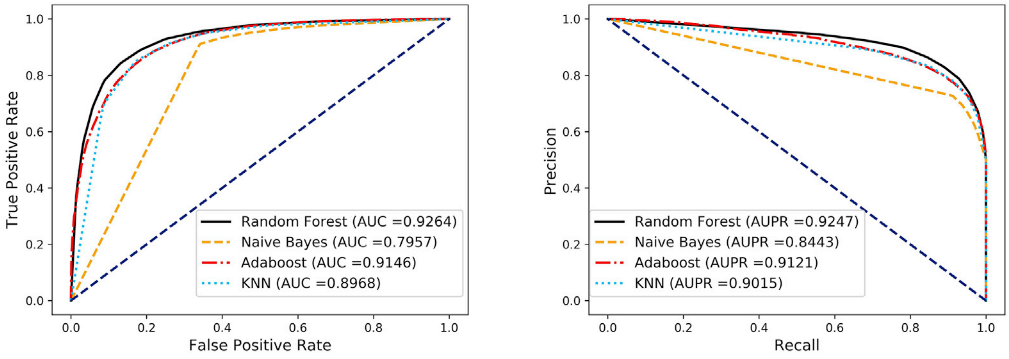


Figure 10. The average ROC and PR curves of DANE-MDA with different classifiers under 5-fold cross-validation

Table 8. Comparison of the average AUC value of DANE-MDA and different models based on HMDD v3.0 dataset

Models	Average AUC (%)
DBMDA	91.29
WBSMDA	81.85
PBMDA	91.72
HDMP	83.42
RLSMDA	85.69
SAE-MDA	**92.64**

functional similarity and Gaussian interaction profile kernel similarity, rather than just the sequence and semantic information of miRNAs and diseases. Second, in terms of advanced feature extraction and avoiding the curse of dimensionality, we hope to compare deep stacked auto-encoder with other deep neural network learning algorithms in the future to achieve better performance. Third, DANE-MDA is a computational machine learning-based prediction model. Hence, a suitable machine learning classifier is essential for our predictive model. We hope to consider other new classifiers to improve prediction ability in the future instead of using the old model such as random forest.

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Zhu-Hong You (zhuhongyou@ms.xjb.ac.cn).

Materials availability

In this study, the known miRNA-disease association dataset was first selected from the Human MicroRNA Disease Database (HMDD) v3.0 (Huang et al., 2019), which is a public online database that contains 32,281 experimentally affirmed miRNA-disease associations from 17,412 papers, containing 850 diseases and 1,102 miRNAs. On this basis, we conducted data preprocessing to eliminate duplicate associations and delete the associations related to certain miRNAs considered unreliable by the public database miRBase (Griffiths-Jones et al., 2006). Finally, 16,427 miRNA-disease associations containing 850 diseases and 901 miRNAs were acquired as the positive samples. Additionally, the Human MicroRNA Disease Database (HMDD) v2.0 dataset was downloaded from the http://www.cuilab.cn/static/hmdd3/data/hmdd2.zip, including 5,430 experimentally verified human miRNA-diseases associations about 383 diseases and 495 miRNAs. For the negative samples, we adopted most previous methods that utilize random selection to generate them with the same number as positive samples (Ben-Hur and Noble, 2005).

Data and code availability

The datasets generated and/or analyzed during this study are available under open licenses in the data repository, https://github.com/jiboya123/DANE-MDA.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

Table 9. Comparison of the average AUC value of DANE-MDA and different models based on HMDD v2.0 dataset

Models	Average AUC (%)
TLHNMDA	87.95
NCMCMDA	89.42
RFMDA	88.18
MDHGI	87.94
SAE-MDA	**91.13**

Table 10. The top 50 miRNA-colon neoplasm associations predicted by DANE-MDA

Rank	miRNA	Evidence	Rank	miRNA	Evidence
1	hsa-miR-29c-5p	dbDemc	26	hsa-miR-199a-5p	dbDemc
2	hsa-miR-99b-5p	dbDemc	27	hsa-miR-19b-3p	dbDemc
3	hsa-miR-144-5p	dbDemc	28	hsa-miR-497-5p	dbDemc
4	hsa-miR-182-5p	dbDemc	29	hsa-miR-30e-5p	dbDemc
5	hsa-miR-92a-2-5p	dbDemce	30	hsa-miR-27b-5p	dbDemc
6	hsa-miR-338-5p	dbDemc	31	hsa-miR-206	dbDemc
7	hsa-miR-422a	dbDemc; miR2Disease	32	hsa-miR-185-5p	dbDemc
8	hsa-miR-199b-5p	dbDemc	33	hsa-miR-425-5p	dbDemc
9	hsa-miR-378a-5p	dbDemc	34	hsa-miR-135a-5p	dbDemc
10	hsa-miR-373-5p	Unconfirmed	35	hsa-miR-491-5p	dbDemc
11	hsa-miR-451a	dbDemc	36	hsa-miR-340-5p	dbDemc
12	hsa-miR-29b-2-5p	dbDemc	37	hsa-miR-149-5p	dbDemc
13	hsa-miR-214-5p	dbDemc	38	hsa-miR-187-5p	dbDemc
14	hsa-miR-503-5p	dbDemc	39	hsa-miR-129-5p	dbDemc
15	hsa-miR-28-5p	dbDemc	40	hsa-miR-184	dbDemc
16	hsa-miR-146b-5p	dbDemc	41	hsa-miR-95-5p	Unconfirmed
17	hsa-miR-590-5p	dbDemc	42	hsa-miR-7-2-3p -7-2-3p	Unconfirmed
18	hsa-miR-342-5p	dbDemc	43	hsa-miR-7-1-3p	dbDemc
19	hsa-miR-193a-5p	dbDemc	44	hsa-miR-582-5p	dbDemc
20	hsa-miR-421	dbDemc	45	hsa-miR-16-5p	dbDemc
21	hsa-miR-186-5p	dbDemc	46	hsa-miR-10a-5p	dbDemc
22	hsa-miR-26a-5p	dbDemc	47	hsa-miR-181a-2-3p	dbDemc
23	hsa-miR-26b-5p	dbDemc	48	hsa-miR-423-5p	dbDemc
24	hsa-miR-124-5p	dbDemc	49	hsa-miR-181c-5p	dbDemc
25	hsa-miR-122-5p	dbDemc	50	hsa-miR-20b-5p	dbDemc

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102455.

ACKNOWLEDGMENTS

Z.-H.Y. was supported by the NSFC Excellent Young Scholars Program, under Grants 61722212 in part by the National Science Foundation of China under Grants 61873212, 61861146002, and 61732012 and in part by the West Light Foundation of the Chinese Academy of Sciences, Grants 2017-XBZG-BR-001. The authors would like to thank the editors and anonymous reviewers for their reviews.

AUTHOR CONTRIBUTION

B.-Y.J. and Z.-H.Y. designed and performed the experiment, Y.W., Z.-W.L., and W.L. prepared data and wrote the article. All the authors contributed to the text of the manuscript.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

Received: December 12, 2020
Revised: March 2, 2021
Accepted: April 19, 2021
Published: June 25, 2021

Table 11. The top 50 miRNA-breast neoplasm associations predicted by DANE-MDA

Rank	miRNA	Evidence	Rank	miRNA	Evidence
1	hsa-miR-15a-5p	dbDemc	26	hsa-miR-582-5p	dbDemc
2	hsa-miR-181d-5p	dbDemc	27	hsa-miR-1271-5p	dbDemc
3	hsa-miR-99b-5p	dbDemc	28	hsa-miR-1231	dbDemc
4	hsa-miR-500a-5p	dbDemc	29	hsa-miR-589-5p	dbDemc
5	hsa-miR-637	dbDemce	30	hsa-miR-650	dbDemc
6	hsa-miR-454-5p	dbDemc	31	hsa-miR-376a-2-5p	Unconfirmed
7	hsa-miR-646	dbDemc	32	hsa-miR-323b-5p	dbDemc
8	hsa-miR-767-5p	dbDemc	33	hsa-miR-384	dbDemc
9	hsa-miR-28-5p	dbDemc	34	hsa-miR-543	dbDemc
10	hsa-miR-382-5p	dbDemc	35	hsa-miR-302e	dbDemc
11	hsa-miR-508-5p	dbDemc	36	hsa-miR-19b-2-5p	dbDemc
12	hsa-miR-211-5p	dbDemc	37	hsa-miR-337-5p	dbDemc
13	hsa-miR-431-5p	dbDemc	38	hsa-miR-557	dbDemc
14	hsa-miR-532-5p	dbDemc	39	hsa-miR-602	dbDemc
15	hsa-miR-483-5p	dbDemc	40	hsa-miR-154-5p	dbDemc
16	hsa-miR-1297	dbDemc	41	hsa-miR-361-5p	dbDemc
17	hsa-miR-519a-5p	Unconfirmed	42	hsa-miR-4732-5p	dbDemc
18	hsa-miR-501-5p	dbDemc	43	hsa-miR-941	dbDemc
19	hsa-miR-628-5p	dbDemc	44	hsa-miR-362-5p	dbDemc
20	hsa-miR-455-5p	dbDemc	45	hsa-miR-297	dbDemc
21	hsa-miR-601	dbDemc	46	hsa-miR-513c-5p	Unconfirmed
22	hsa-miR-622	dbDemc	47	hsa-miR-571	dbDemc
23	hsa-miR-422a	dbDemc	48	hsa-miR-544a	dbDemc
24	hsa-miR-300	dbDemc	49	hsa-miR-636	dbDemc
25	hsa-miR-325	dbDemc	50	hsa-miR-3651	dbDemc

REFERENCES

Alaimo, S., Giugno, R., and Pulvirenti, A. (2014). ncPred: ncRNA-disease association prediction through tripartite network-based inference. Front. Bioeng. Biotechnol. *2*, 71.

Ambros, V. (2001). microRNAs: tiny regulators with great potential. Cell *107*, 823–826.

Ambros, V. (2004). The functions of animal microRNAs. Nature *431*, 350–355.

Bang, C., Fiedler, J., and Thum, T. (2012). Cardiovascular importance of the microRNA-23/27/24 family. Microcirculation *19*, 208–214.

Ben-Hur, A., and Noble, W.S. (2005). Kernel methods for predicting protein–protein interactions. Bioinformatics *21*, i38–i46.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. *68*, 394–424.

Chen, X., Liu, M.-X., and Yan, G.-Y. (2012). RWRMDA: predicting novel human microRNA–disease associations. Mol. BioSyst. *8*, 2792–2798.

Chen, X., Qu, J., and Yin, J. (2018a). TLHNMDA: triple layer heterogeneous network based inference for MiRNA-disease association prediction. Front. Genet. *9*, 234.

Chen, X., Sun, L.-G., and Zhao, Y. (2021). NCMCMDA: miRNA–disease association prediction through neighborhood constraint matrix completion. Brief. Bioinformatics *22*, 485–496.

Chen, X., Wang, C.-C., Yin, J., and You, Z.-H. (2018b). Novel human miRNA-disease association inference based on random forest. Mol. Ther. Nucleic Acids *13*, 568–579.

Chen, X., Yan, C.C., Zhang, X., You, Z.-H., Deng, L., Liu, Y., Zhang, Y., and Dai, Q. (2016). WBSMDA: within and between score for MiRNA-disease association prediction. Sci. Rep. *6*, 21106.

Chen, X., and Yan, G.-Y. (2014). Semi-supervised learning for potential human microRNA-disease associations inference. Sci. Rep. *4*, 5501.

Chen, X., Yin, J., Qu, J., and Huang, L. (2018c). MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. PLoS Comput. Biol. *14*, e1006418.

Cooil, B., Winer, R.S., and Rados, D.L. (1987). Cross-validation for prediction. J. Marketing Res. *24*, 271–279.

Cui, Q., Yu, Z., Purisima, E.O., and Wang, E. (2006). Principles of microRNA regulation of a human cellular signaling network. Mol. Syst. Biol. *2*, 46.

Denoeux, T. (2008). A k-nearest neighbor classification rule based on Dempster-Shafer theory. In Classic Works of the Dempster-Shafer Theory of Belief Functions (Springer), pp. 737–760.

Griffiths-Jones, S., Grocock, R.J., Van Dongen, S., Bateman, A., and Enright, A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. *34*, D140–D144.

He, T., Hu, L., Chan, K.C., and Hu, P. (2018). Learning latent factors for community identification and summarization. IEEE Access *6*, 30137–30148.

He, T., Liu, Y., Ko, T.H., Chan, K.C., and Ong, Y.-S. (2019). Contextual correlation preserving multiview featured graph clustering. IEEE Trans. Cybern. *50*, 4318–4331.

Table 12. The top 50 miRNA-lung neoplasm associations predicted by DANE-MDA

Rank	miRNA	Evidence	Rank	miRNA	Evidence
1	hsa-miR-15b-5p	dbDemc	26	hsa-miR-16-2-3p	dbDemc
2	hsa-miR-16-1-3p	dbDemc	27	hsa-miR-425-5p	dbDemc; miR2Disease
3	hsa-miR-518b	dbDemc	28	hsa-miR-484	dbDemc
4	hsa-miR-642a-5p	dbDemc	29	hsa-miR-575	dbDemc
5	hsa-miR-429	dbDemc; miR2Disease	30	hsa-miR-452-5p	dbDemc
6	hsa-miR-106b-5p	dbDemc	31	hsa-miR-590-5p	dbDemc
7	hsa-miR-424-5p	dbDemc	32	hsa-miR-625-5p	dbDemc
8	hsa-miR-28-5p	dbDemc	33	hsa-miR-193b-5p	dbDemc
9	hsa-miR-382-5p	dbDemc	34	hsa-miR-302c-5p	Unconfirmed
10	hsa-miR-409-5p	dbDemc	35	hsa-miR-505-5p	dbDemc
11	hsa-miR-421	dbDemc	36	hsa-miR-181b-5p	dbDemc
12	hsa-miR-532-5p	dbDemc	37	hsa-miR-708-5p	dbDemc
13	hsa-miR-483-5p	dbDemc	38	hsa-miR-1246	dbDemc
14	hsa-miR-128-3p	dbDemc	39	hsa-miR-151a-5p	dbDemc
15	hsa-miR-491-5p	dbDemc	40	hsa-miR-376c-5p	dbDemc
16	hsa-miR-885-5p	dbDemc	41	hsa-miR-370-5p	dbDemc
17	hsa-miR-92b-5p	Unconfirmed	42	hsa-miR-298	dbDemc
18	hsa-miR-509-5p	dbDemc	43	hsa-miR-23b-5p	dbDemc
19	hsa-miR-1307-5p	dbDemc	44	hsa-miR-628-5p	dbDemc
20	hsa-miR-455-5p	dbDemc	45	hsa-miR-539-5p	dbDemc
21	hsa-miR-489-5p	Unconfirmed	46	hsa-miR-711	Unconfirmed
22	hsa-miR-422a	dbDemc	47	hsa-miR-1179	dbDemc
23	hsa-miR-1271-5p	dbDemc	48	hsa-miR-1244	dbDemc
24	hsa-miR-125b-2-3p	dbDemc	49	hsa-miR-339-5p	dbDemc
25	hsa-miR-181d-5p	dbDemc	50	hsa-miR-3613-5p	dbDemc

Hu, L., Chan, K.C., Yuan, X., and Xiong, S. (2019). A variational Bayesian framework for cluster analysis in a complex network. IEEE Trans. Knowledge Data Eng. *32*, 2115–2128.

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., Zhou, Y., and Cui, Q. (2019). HMDD v3. 0: a database for experimentally supported human microRNA–disease associations. Nucleic Acids Res. *47*, D1013–D1017.

Jeong, H.C., Kim, E.K., Lee, J.H., Yoo, H.N., and Kim, J.K. (2011). Aberrant expression of let-7a miRNA in the blood of non-small cell lung cancer patients. Mol. Med. Rep. *4*, 383–387.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. (2008). miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. *37*, D98–D104.

Karp, X., and Ambros, V. (2005). Encountering microRNAs in cell fate signaling. Science *310*, 1288–1289.

Kelsey, J.L., and Horn-Ross, P.L. (1993). Breast cancer: magnitude of the problem and descriptive epidemiology. Epidemiol. Rev. *15*, 7.

Kipf, T.N., and Welling, M. (2016). Semi-supervised Classification with Graph Convolutional Networks (arXiv), p. 1609.02907.

Liang, C., Yu, S., and Luo, J. (2019). Adaptive multi-view multi-label learning for identifying disease-associated candidate miRNAs. PLoS Comput. Biol. *15*, e1006931.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. R. News *2*, 18–22.

Ling, H., Fabbri, M., and Calin, G.A. (2013). MicroRNAs and other non-coding RNAs as targets for anticancer drug development. Nat. Rev. Drug Discov. *12*, 847.

Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., and Ferrando, A.A. (2005). MicroRNA expression profiles classify human cancers. Nature *435*, 834–838.

Luo, J., Xiao, Q., Liang, C., and Ding, P. (2017). Predicting MicroRNA-disease associations using Kronecker regularized least squares based on heterogeneous omics data. IEEE Access *5*, 2503–2513.

Margineantu, D.D., and Dietterich, T.G. (1997). Pruning Adaptive Boosting (Citeseer), pp. 211–218.

Matsui, M., and Corey, D.R. (2017). Non-coding RNAs as drug targets. Nat. Rev. Drug Discov. *16*, 167–179.

Mishra, R., Bhattacharya, S., Rawat, B.S., Kumar, A., Kumar, A., Niraj, K., Chande, A., Gandhi, P., Khetan, D., and Aggarwal, A. (2020). MicroRNA-30e-5p has an integrated role in the regulation of the innate immune response during virus infection and systemic lupus erythematosus. Iscience *23*, 101322.

Rish, I. (2001). An empirical study of the naive Bayes classifier (In: IJCAI 2001 workshop on empirical methods in artificial intelligence), pp. 41–46.

Rupaimoole, R., and Slack, F.J. (2017). MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. Nat. Rev. Drug Discov. *16*, 203.

Siegel, R.L., Miller, K.D., and Jemal, A. (2016). Cancer statistics, 2016. CA Cancer J. Clin. *66*, 7–30.

Tao, Z., Shi, A., Lu, C., Song, T., Zhang, Z., and Zhao, J. (2015). Breast cancer: epidemiology and etiology. Cell Biochem. Biophys. *72*, 333–338.

Torre, L.A., Siegel, R.L., and Jemal, A. (2016). Lung cancer statistics. In Lung Cancer and Personalized Medicine (Springer), pp. 1–19.

Wang, L., You, Z.-H., Chen, X., Li, Y.-M., Dong, Y.-N., Li, L.-P., and Zheng, K. (2019). LMTRDA: using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities. PLoS Comput. Biol. *15*, e1006865.

Wong, L., You, Z.-H., Guo, Z.-H., Yi, H.-C., Chen, Z.-H., and Cao, M.-Y. (2020). MIPDH: a novel computational model for predicting microRNA–mRNA interactions by DeepWalk on a heterogeneous network. ACS Omega *5*, 17022–17032.

Xu, J., Li, C.-X., Lv, J.-Y., Li, Y.-S., Xiao, Y., Shao, T.-T., Huo, X., Li, X., Zou, Y., and Han, Q.-L. (2011). Prioritizing candidate disease miRNAs by topological features in the miRNA target–dysregulated network: case study of prostate cancer. Mol. Cancer Ther. *10*, 1857–1866.

Xu, P., Guo, M., and Hay, B.A. (2004). MicroRNAs and the regulation of cell death. Trends. Genet. *20*, 617–624.

Xuan, P., Han, K., Guo, M., Guo, Y., Li, J., Ding, J., Liu, Y., Dai, Q., Li, J., and Teng, Z. (2013). Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. PLoS one *8*, e70204.

Xuan, P., Sun, H., Wang, X., Zhang, T., and Pan, S. (2019). Inferring the disease-associated miRNAs based on network representation learning and convolutional neural networks. Int. J. Mol. Sci. *20*, 3648.

Yang, Z., Ren, F., Liu, C., He, S., Sun, G., Gao, Q., Yao, L., Zhang, Y., Miao, R., and Cao, Y. (2010). dbDEMC: A Database of Differentially Expressed miRNAs in Human Cancers, *4* (BioMed Central), p. S5.

Yi, H.-C., You, Z.-H., Huang, D.-S., Guo, Z.-H., Chan, K.C., and Li, Y. (2020). Learning representations to predict intermolecular interactions on large-scale heterogeneous molecular association network. Iscience *23*, 101261.

You, Z.-H., Huang, Z.-A., Zhu, Z., Yan, G.-Y., Li, Z.-W., Wen, Z., and Chen, X. (2017). PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. PLoS Comput. Biol. *13*, e1005455.

Zheng, K., You, Z.-H., Wang, L., and Guo, Z.-H. (2020a). iMDA-BN: identification of miRNA-disease associations based on the biological network and graph embedding algorithm. Comput. Struct. Biotechnol. J. *18*, 2391–2400.

Zheng, K., You, Z.-H., Wang, L., Zhou, Y., Li, L.-P., and Li, Z.-W. (2020b). Dbmda: a unified embedding for sequence-based mirna similarity measure with applications to predict and validate mirna-disease associations. Mol. Ther. Nucleic Acids *19*, 602–611.

iScience, Volume *24*

Supplemental information

DANE-MDA: Predicting microRNA-disease associations via deep attributed network embedding

Bo-Ya Ji, Zhu-Hong You, Yi Wang, Zheng-Wei Li, and Leon Wong

Supplemental Figures

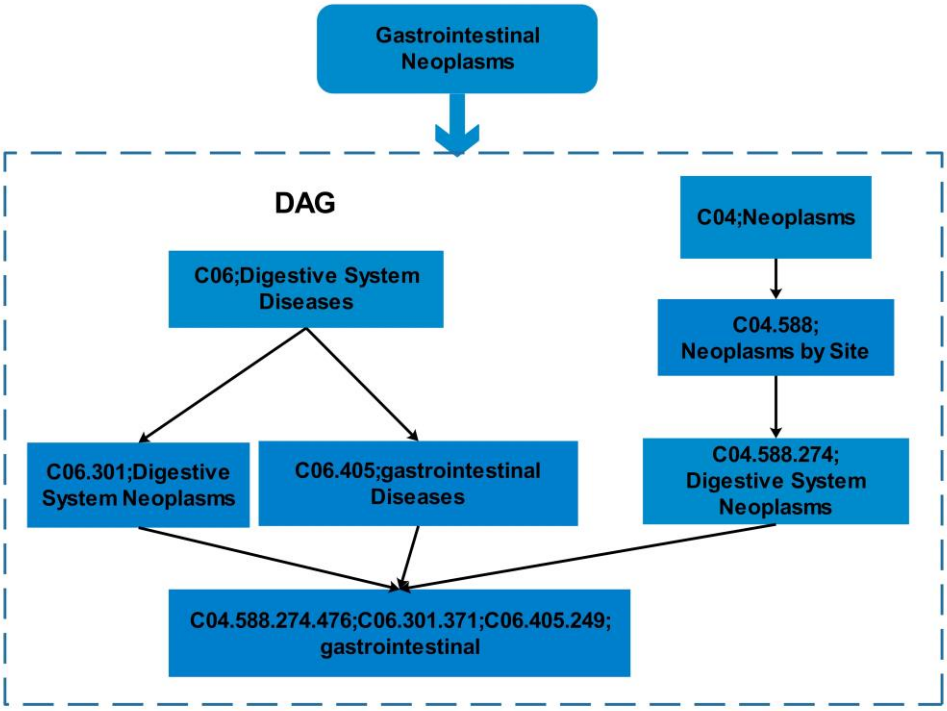

Gastrointestinal Neoplasms
DAG
C04;Neoplasms
C06;Digestive System Diseases
C04.588; Neoplasms by Site
C06.301;Digestive System Neoplasms
C06.405;gastrointestinal Diseases
C04.588.274; Digestive System Neoplasms
C04.588.274.476;C06.301.371;C06.405.249; gastrointestinal

Figure S1. The directed acyclic graph (DAG) of gastrointestinal neoplasms. Related to Figure 1.

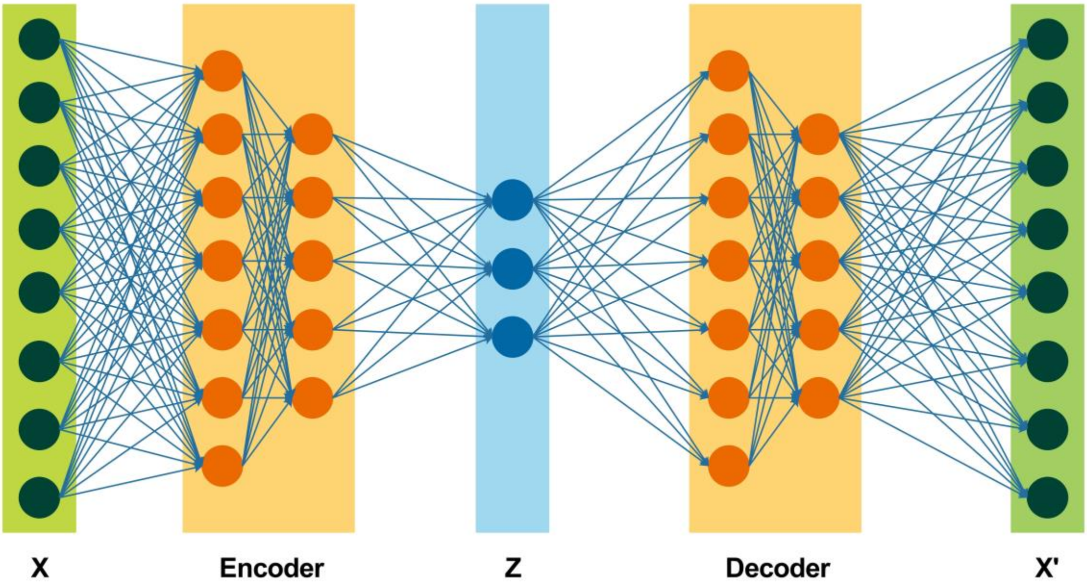


Figure S2. The simplified schematic diagram of deep stacked auto-encoder neural network. Related to Figure 1.

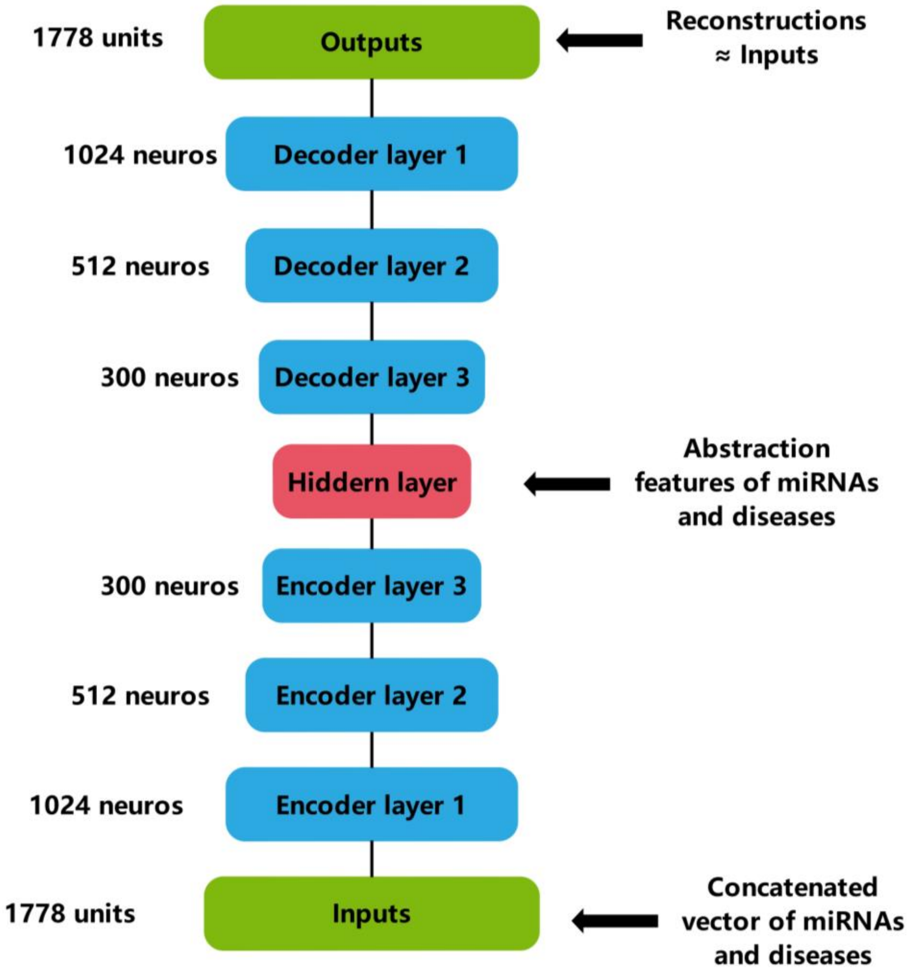


Figure S3. The architecture of our deep stacked auto-encoder neural network model. Related to Figure 1.

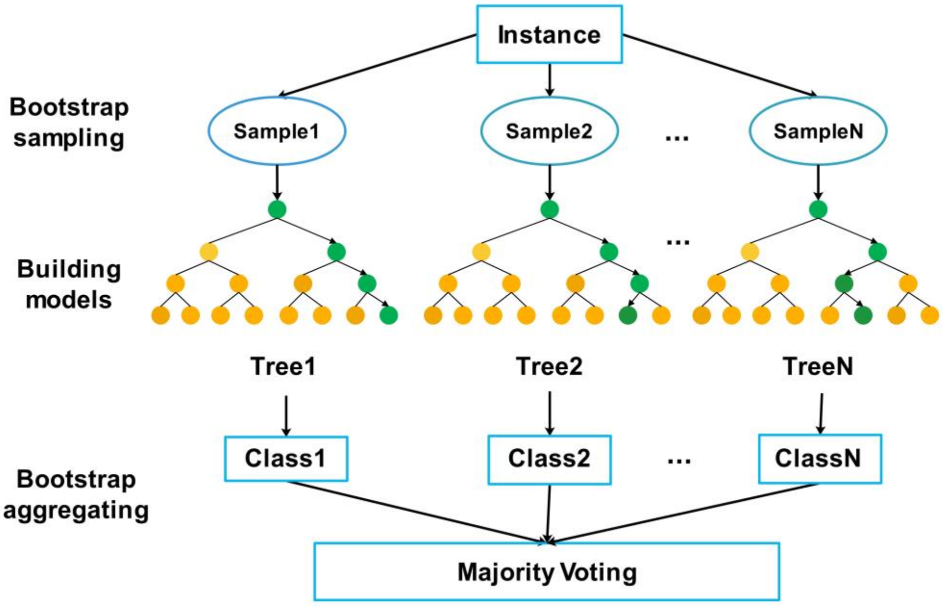


Figure S4. The simplified flowchart of the Random Forest classifier. Related to Figure 1.

Transparent Methods

miRNA sequence similarity information

In this study, the attribute feature of miRNAs was represented by the sequence similarity information. Generally, miRNA sequences are usually denoted by simplified letters of four nitrogenous bases: uracil (U), cytosine (C), guanine (G), and adenine (A). We downloaded the miRNA sequence information from the public miRBase database (Griffiths-Jones et al., 2006) and then utilized the 3-mer method to obtain the numerical statistical features of miRNA sequences. Specifically, we first set up a sliding window with a window size of 3 and a sliding distance of 1, to split the miRNA sequence into multiple 3-monomeric units (3-mers). Second, the occurrence number of each 3-mer is divided by the corresponding miRNA sequence length to obtain its occurrence frequency, and the occurrence frequency of non-occurring 3-mers is set to 0. Finally, each miRNA sequence is converted into a 64-dimensional vector based on the 64 kinds of 3-mer combinations. On this basis, we continue to use the most common similarity measurement method Euclidean distance to calculate the miRNA sequence similarity (MSS), defined as follows:

$$\mathrm{Sim}(\mathrm{M},\ \mathrm{M}') = \sqrt{\sum_{i=1}^{n}(M_i - M'_i)^2} \tag{1}$$

$$\mathrm{MSS(M)} = (\mathrm{Sim}(M, M'_1), \mathrm{Sim}(M, M'_2), \dots, \mathrm{Sim}(M, M'_m)) \tag{2}$$

where M and M' represent the numerical statistical feature vectors of two miRNA sequences, *n* represents the vector length, and *m* means the number of miRNAs.

Disease semantic similarity

In this study, the disease semantic similarity was used to represent the attribute feature of diseases. The Medical Subject Heading (MeSH) descriptors of diseases provide a strict disease classification system, which can be obtained from the U.S. National Library of Medicine (https://www.nlm.nih.gov/) (Lipscomb, 2000). MeSH descriptors are divided into 16 categories: category A is anatomical terms, category B is organisms, category C used in this study is disease terms, and so on. On this basis, the relationship among various diseases can be represented as a directed acyclic graph (DAG), where the nodes represent the MeSH descriptors of the diseases, and the directed edges point from more general items (parent nodes) to more specific ones (child nodes). Besides, there are one or more tree numbers of each MeSH descriptor to indicate its position in the DAGs. The child node's tree number is its parent node's tree number appended by its information. Figure S1 shows an example of the DAG for gastrointestinal neoplasms. For instance, disease A can be defined as DAG(A) = (D(A), E(A)), in which D(A) is meant as A and its ancestor nodes, and E(A) is meant as all the direct edges. On this basis, the semantic contribution of disease term *t* in DAG(A) to disease A is defined as follows:

$$\begin{cases} D_A(t) = 1 & if\ t = A \\ D_A(t) = max\{\Delta * D_A(t') | t' \in children\ of\ t\} & if\ t \neq A \end{cases} \tag{3}$$

where Δ is the semantic contribution attenuation factor, which means that its semantic contribution to disease A will decrease as the distance between item *t* and disease A increases. Disease A is at the bottom of the DAG, so we defined its contribution value as 1. According to the above formula, the contribution of items at different levels to the semantic value of disease A can be differentiated. Finally, the semantic value of disease A is achieved by summarizing all the contributions from itself and its ancestor diseases, as shown below:

$$\mathrm{DV(A)} = \sum_{t \in D(A)} D_A(t) \tag{4}$$

Hence, the disease semantic similarity (DSS) between diseases d_i and d_j is acquired based on the nodes shared by the two disease DAGs as follows:

$$\mathrm{DSS(d_i, d_j)} = \frac{\sum_{t \in D(d_i) \cap D(d_j)} (D_{d_i}(t) + D_{d_j}(t))}{DV(d_i) + DV(d_j)} \tag{5}$$

Network structure feature of miRNAs and diseases

In this study, the local network structure feature of miRNAs and diseases was represented by the probability of direct transitions between each miRNA-disease association pair. First, we generated an adjacency matrix *R* based on the constructed attributed miRNA-disease association network. The row and column number of *R* is 901 and 850, representing the number of miRNAs and diseases. The element R_{ij} in the matrix represents the relationship between miRNA m_i and disease d_j. If there is an association between m_i and d_j, the R_{ij} is equal to 1, otherwise, equal to 0. Second, we normalized the adjacency matrix *R* by row to generate the network structure feature matrix *S*, which shows the connection probability between miRNAs and diseases within one step, given by:

$$\mathrm{S}_{\mathrm{ij}} = \frac{R_{ij}}{\sum_{k\in N} R_{ik}} \tag{6}$$

where *N* is the column number of matrix *R*, and S_{ij} is the associated probability of miRNA m_i and disease d_j. Thus, the structural feature matrix *S* should satisfy the following constraints:

$$0 \leqslant S_{ij} \leqslant 1 \tag{7}$$

$$\forall\, \mathrm{i} \in [1, 2, \cdots, \mathrm{N}],\ \sum_{k=1}^{N} S_{ij} = 1 \tag{8}$$

Construct the attribute and structure matrix representation

The attribute matrix representation *A* for the attributed network is formed by combining the miRNA sequence similarity matrix *RM* and disease semantic similarity matrix *RD*. Moreover, since there is no attribute relationship between miRNAs and diseases, we set this part as the 0 matrices. The final attribute matrix representation is defined as follows:

$$\mathrm{A} = \begin{vmatrix} RM & 0 \\ 0 & RD \end{vmatrix} \tag{9}$$

The network structure matrix *S* is composed of the probability of direct transition between each miRNA-disease association pair. Similarly, since there is no structural relationship between miRNAs and diseases themselves, we also set this part as the 0 matrices. The final structure matrix representation is defined as follows:

$$\mathrm{S} = \begin{vmatrix} 0 & S \\ S^T & 0 \end{vmatrix} \tag{10}$$

where S^T is represented the transposed matrix of network structure matrix S. The number of rows and columns of the network structure matrix are both the sum of the number of miRNAs and diseases.

Step-based proximity calculation

For the purpose of catching the interactions between the attribute and network structure feature from diverse degrees of proximity, the graph-based random walk idea was borrowed to construct a step-based proximity matrix P^t at each step *t*. The first-degree proximity matrix P^1 is meant as the linear combination of the attribute feature matrix *A* and the network structure feature matrix *S*, in which only the first-order proximity between miRNA m_i and disease d_j in the network structure is considered, given by:

$$P^1_{(i,j)} = \mathrm{\alpha} S(\mathrm{i,j}) + (1-\mathrm{\alpha})\mathrm{A}(\mathrm{i,j}) \tag{11}$$

where $\alpha \in (0, 1)$ is the weight coefficient, which means the random walk preference ratio between attribute and structure feature matrix.

Furthermore, in order to catch the higher-degree structure proximity, we defined the (*t+1*)-th step-based proximity P^{t+1} as:

$$P^{t+1} = \alpha P^t \mathrm{S} + (1-\alpha)A \tag{12}$$

Specifically, the (*t+1*)-th step structure proximity was obtained by multiplying the *t*-th step proximity matrix P^t by the structure matrix S, and since the attribute features are static in network structure changes, the attribute proximity is always A. In this way, we obtained both the attribute features and the local and global structure features of the network from different degrees of proximity with the proximity matrix sequences: $P^1, P^2, \ldots, P^t$.

Diverse degrees of proximity fusion

In this part, to preserve both the attribute features, as well as the local and global network structure feature of miRNAs and diseases, an enhanced matrix Q is constructed by fusing the diverse degrees of proximity: $P^1, P^2, \dots, P^t$. Generally, it is a common fusion strategy to average the sum of all matrices. But intuitively, the closer (the smaller the degree) the connections between miRNAs and diseases, the closer the relationship between them. In other words, the low-order proximity nodes have a greater influence than high-order proximity ones. Hence, a weight function that decreases monotonously with the increase of step *t* is defined as:

$$\mathrm{Q} = \sum_{t=1}^{T} f(t) * P^t \tag{13}$$

where *f(t)* represents a decreasing function, and in this study, an exponential function modified by the parameter $\beta \in (0, 1)$ is used as the weighting strategy as shown below:

$$f(t) = \beta^t \tag{14}$$

Deep stacked auto-encoder neural network

In order to improve feature quality and reduce noise, we further learned the nonlinear and complex low-dimensional features in the fusion matrix Q. The deep stacked auto-encoder neural network (SAE) (Rumelhart et al., 1986) is utilized to obtain the embedding features of miRNAs and diseases. Specifically, SAE is a category of unsupervised learning for data compression, and the simplified SAE is a three-layer neural network model, including a data input layer, a hidden layer, and an output reconstruction layer. The encoding process is used to map the input data from the input layer to the hidden layer, and the decoding process is used to map the hidden data from the hidden layer to the output layer to reconstruct the input data. The schematic diagram of the simplified deep stacked auto-encoder is shown in Figure S2. Given the input data:

$$x = [x_1, x_2, \dots, x_{d(x)}]^T \in R^{d(x)} \qquad (15)$$

where *d(x)* means the dimension of the input data, and then the *x* is projected by the encoder from the input layer to the hidden layer data *z* with the mapping function *f* :

$$z = [z_1, z_2, \dots, z_{d(z)}]^T \in R^{d(z)} \qquad (16)$$

where *d(z)* means the dimension of the hidden layer data, and *f(x)* function is expressed as:

$$z = f(x) = s_f(Wx + b) \qquad (17)$$

where $W \in R^{d(x)*d(z)}$ is the weight matrix, $b \in R^{d(x)}$ is the deviation vector. The activation function s_f of the encoder can be a sigmoid function, a tanh function, or a rectified linear unit function (ReLu function).

In the decoder, the hidden layer representation *z* is mapped to the output layer $x' \in R^{d(x')}$ through the mapping function f', where the function is as follows:

$$x' = f'(z) = s_{f'}(W'z + b') \qquad (18)$$

where $W' \in R^{d(x^{'})*d(z)}$ is the weight matrix, $b' \in R^{d(x^{'})}$ is the deviation vector. Similarly, the activation function $s_{f'}$ of the decoder can also be a sigmoid, tanh, or ReLu function. Thus, the parameter set of SAE is:

$$\theta = \{W, W', b, b'\} \qquad (19)$$

To obtain the optimal model parameters, the loss function is reconstructed by computing the mean square reconstruction error to minimize:

$$J(W, W', b, b') = \frac{\sum_{i=1}^{N} \left\| x_i' - x_i \right\|^2}{2\mathrm{N}} \qquad (20)$$

where *N* is the total number of training samples. Figure S3 shows the architecture of our stacked auto-encoder model. Specifically, the input layer of the model is a concatenated vector of diseases and miRNAs. The encoder part contains a total of 3 layers, each containing 1024, 512, and 300 neurons. The decoder part has the reverse architecture of the encoder, each containing 300, 512, and 1024 neurons. Moreover, we set nonlinear activation functions to ReLU, the loss function to the mean squared error (MSE), which is minimized using Adam, the epochs to 100, and the batch size to 128.

Random Forest classifier

Generally, determining whether there is an association between miRNAs and diseases is regarded as a binary classification problem. In this study, the Random Forest (RF) classifier (Liaw and Wiener, 2002) is chosen for training the deep attributed network embedding features of miRNAs and diseases and predicting potential associations between them. In particular, Random Forest is a significant bagging-based ensemble learning method and has a lot of advantages, such as high accuracy rate, not easy to overfit, and good anti-noise ability, which could be utilized for regression, classification, and other problems. Its construction process is roughly as follows: (1) Generate *N* samples from the instance by utilizing the bootstrap sampling method. (2) Establish *N* decision tree models based on *N* training samples. For a single decision tree model, the best feature is used for each split according to the Gini index/information gain ratio/ information gain. (3) Use the majority voting mechanism to determine the final prediction result. Figure S4 shows the simplified flowchart of the Random Forest classifier.

Supplemental References

Griffiths-Jones, S., Grocock, R.J., Van Dongen, S., Bateman, A., and Enright, A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. Nucleic acids research *34*, D140-D144.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. R news *2*, 18-22.

Lipscomb, C.E. (2000). Medical subject headings (MeSH). Bulletin of the Medical Library Association *88*, 265.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning representations by back-propagating errors. nature *323*, 533-536.