scientific reports

Check for updates

OPEN

# A structural deep network embedding model for predicting associations between miRNA and disease based on molecular association network

Hao-Yuan Li[1,5], Hai-Yan Chen[2,5], Lei Wang[3]✉, Shen-Jian Song[4]✉, Zhu-Hong You[3], Xin Yan[1] & Jin-Qian Yu[1]

**Previous studies indicated that miRNA plays an important role in human biological processes especially in the field of diseases. However, constrained by biotechnology, only a small part of the miRNA-disease associations has been verified by biological experiment. This impel that more and more researchers pay attention to develop efficient and high-precision computational methods for predicting the potential miRNA-disease associations. Based on the assumption that molecules are related to each other in human physiological processes, we developed a novel structural deep network embedding model (SDNE-MDA) for predicting miRNA-disease association using molecular associations network. Specifically, the SDNE-MDA model first integrating miRNA attribute information by Chao Game Representation (CGR) algorithm and disease attribute information by disease semantic similarity. Secondly, we extract feature by structural deep network embedding from the heterogeneous molecular associations network. Then, a comprehensive feature descriptor is constructed by combining attribute information and behavior information. Finally, Convolutional Neural Network (CNN) is adopted to train and classify these feature descriptors. In the five-fold cross validation experiment, SDNE-MDA achieved AUC of 0.9447 with the prediction accuracy of 87.38% on the HMDD v3.0 dataset. To further verify the performance of SDNE-MDA, we contrasted it with different feature extraction models and classifier models. Moreover, the case studies with three important human diseases, including Breast Neoplasms, Kidney Neoplasms, Lymphoma were implemented by the proposed model. As a result, 47, 46 and 46 out of top-50 predicted disease-related miRNAs have been confirmed by independent databases. These results anticipate that SDNE-MDA would be a reliable computational tool for predicting potential miRNA-disease associations.**

MicroRNAs (miRNAs) are one type of small non-coding RNA with length of 20–25 nucleotides[1]. They normally influence their target messenger RNAs (mRNAs) by base pairing binding to the 3′ untranslated region (UTR) sites of mRNAs[2]. These small molecules could function as negative regulator of target gene expression in post-transcriptional[3]. With the development of molecular biology, increasing miRNAs have been detected[4]. To date, the famous miRbase database have collected 48,860 mature miRNAs from 271 organisms containing more than 1000 human miRNAs[5]. In addition, researchers have found that miRNAs are related with multiple significant cell biological activities, involving diffusion, aging, development, death and so on[6–9].

In recent years, an increasing number of experiments have demonstrated that there are close relationships between miRNA with disease[10–13]. In particular, miRNAs have been new biomarkers for human cancer, which is important to cancer preventions and treatments[14]. Therefore, identifying the miRNA-disease associations has gradually become a hot topic in biology[15]. Early traditional biological experiments identified the disease-related

[1]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China. [2]Xinjiang Autonomous Region tax Service, State Taxation Administration, Urumqi 830011, China. [3]Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China. [4]Science & Technology Department of Xinjiang Uygur Autonomous Region, Urumqi 830011, China. [5]These authors contributed equally: Hao-Yuan Li and Hai-Yan Chen. ✉email: leiwang@ms.xjb.ac.cn; ArchibaldDaniel@163.com

miRNAs by detecting the expression level of miRNAs in biological disease process[16]. For example, Yohei et al. found that miR-200c could build a molecular link between breast cancer cells and normal cells[17]. Liu et al. point out that many miRNAs are disordered in cancer and this situation occurs because miRNAs participate in tumorigenesis and function as oncogenes[18]. Thum et al. reported that miR-21 adjust expression of the ERK-MAP kinase to effect on structure and function of heart[19]. Traditional experiments achieve high accuracy, while it has the limitations of long experimental time, high cost, and low success rate[20]. To resolve these issues, for effectively and accurately predict potential miRNA-disease associations, increasing researchers adopted computational model and select the most possible related miRNAs for further traditional biological experiments[21].

With the development of biotechnology, some databases were constructed by collecting these biological data. These datasets provide the possibility to classify associations of miRNA-disease through computational methods[20,22–25]. Over the years, these methods mostly are according to the assumption that these functionally similar miRNAs tend to be related with semantically similar diseases[2,26–28]. These models could be split into under similarity network models and machine learning models[29]. For example, Jiang et al.[22] presented a computational model to speculate the relationship between miRNA and disease based on a hypergeometric distribution model. This is an early calculation model by fusing multiple sources of information. However, this method built the miRNA-related network by functional similarity, which is limited by the relationship between miRNAs. Based on random walk method, Xuan et al.[30] presented MIDP and MIDPE, an extension method of MIDP. MIDP constructed the network by combining the information of each node including similarity, prior information and various ranges of topological structure. This model could effectively reduce noise from data by restarting the walk. Furthermore, You et al.[31] proposed PBMDA constructed a heterogeneous graph including three subgraphs. PBMDA is a depth-first algorithm based on path, which could fully use the topology information of heterogeneous network. In particularly, the priority of new associations between diseases and miRNAs could be identified by evaluating the score of the path. Chen et al.[32] proposed a computational method adopted the extreme gradient boosting named EGBMMDA. This is the first learning method based on decision tree for classifying miRNA-disease relationships. EGBMMDA built a comprehensive feature vector by various methods such as statistical, graph theory and matrix factorization. These studies have continually improved the performance of computational method and played an important guiding role in traditional biological experiments[33]. Therefore, accurately and effectively predict associations between miRNA-disease through computational method become urgently demanded[34].

In this study, based on the assumption of molecules are related to each other in human physiological processes, we developed a structural deep network embedding-based model (SDNE-MDA) for predicting miRNA-disease association using molecular association network. The flow chart of SDNE-MDA is shown as Fig. 1. Specifically, we first constructed the molecular association network (MAN)[35] by combining multiple different molecules with edges of them. This study extracted behavior information from the heterogeneous network by the structural deep network embedding (SDNE)[36], which could maintain the overall structure of large network to the greatest extent. Secondly, SDNE-MDA obtained the miRNA attribute information by the chaos game representation (CGR) algorithm and disease attribute information by disease semantic similarity. After then, we formed the feature descriptor by fusing the behavior information and attribute information of miRNAs and diseases. Finally, these feature descriptors are trained and classified by the CNN to predict miRNA-disease associations. Five-fold cross validation experiment was carried out for SDNE-MDA to verify the performance of prediction and achieved the AUC of 0.9447 with the prediction accuracy of 87.38%. To further evaluate SDNE-MDA, we contrasted the proposed model with two feature extraction models and classifier models. Besides, we carry out SDNE-MDA with three significant human diseases involving breast cancer, kidney cancer and lymphoma. And as a result, 47, 46 and 46 out of top-50 candidate related miRNAs are confirmed by known databases and recent literature, respectively. These experiment result demonstrated that SDNE-MDA is a precisely and effectively computational method for predicting potential associations between miRNA with disease.

## Materials and methods

**Benchmark database.** Human miRNA-disease associations benchmark database HMDD v3.0[37] was adopted as data support in this paper, which collected 32,281 confirmed miRNA-disease associations, involving 1102 miRNAs and 850 diseases. Here, after data processing, we chose 16,427 known miRNA-disease associations as positive samples including 1023 miRNAs and 850 diseases. What's more, we defined the adjacency matrix $AM$ to represent the miRNA-disease associations. When the miRNA $mi(a)$ have a verified association with the disease $di(b)$, we set $AM(mi(a), di(b)) = 1$, otherwise $AM(mi(a), di(b)) = 0$. In this paper, we introduce two other independent databases (dbDEMC[38] and miR2Ddisease[39]) to verified the result of case study.

**Molecular associations network.** In this study, we combined multiple biological molecular information according the Molecular association network (MAN). The MAN is a heterogeneous information network proposed by Guo et al.[40]. Currently, this complex network consists of five types of molecular (miRNA, lncRNA, protein, disease, drug) and associations between them. The heterogeneous information network MAN provided a new comprehensive view to explore the complex physiological process and human disease. The structure diagram of molecular association network is as shown in Fig. 2. In this study, we download the information of molecular and associations between them from multiple databases. The number of different molecules is shown in Table 1, and the associations between them are shown in the following Table 2.

**Chaos game representation (CGR) algorithm.** MiRNA sequences contain a lot of complex information. However, most of the existing sequence feature information extraction algorithms only quantify one of position information and nonlinear information. In order to measure the similarity of these information con-

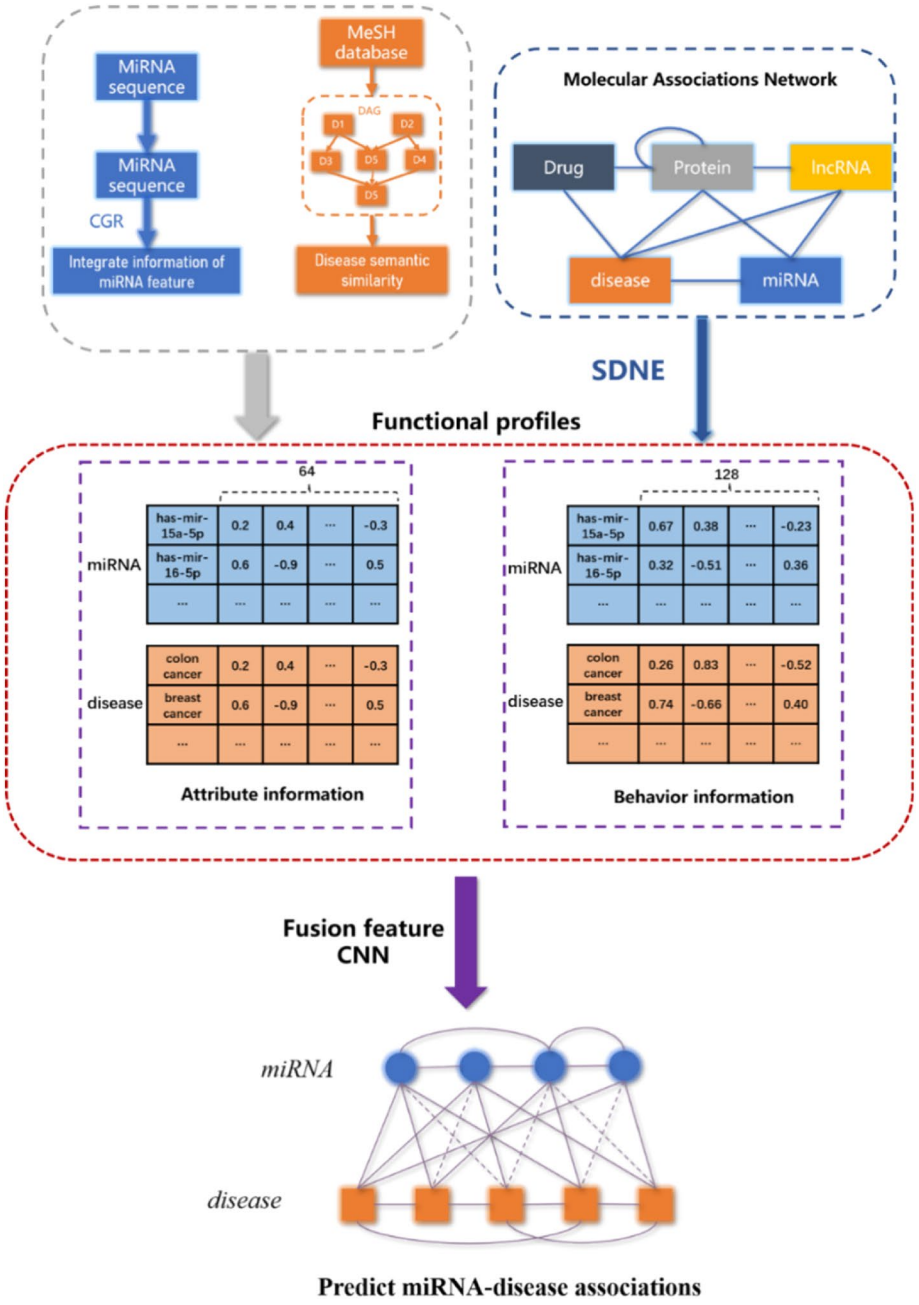


**Figure 1.** Flowchart of SDNE-MDA to predict potential miRNA-disease associations.

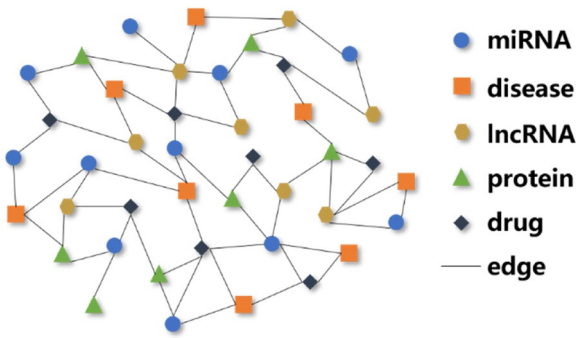


**Figure 2.** Structure diagram of molecular association network.

| Molecular | Number |
|---|---|
| MiRNA | 1023 |
| Disease | 2026 |
| Drug | 1025 |
| LncRNA | 769 |
| Protein | 1647 |
| Total | 6528 |

**Table 1.** The number of different types of nodes in MAN.

| Association | Database | Number |
|---|---|---|
| miRNA-disease | HMDD[41] | 16,427 |
| miRNA-protein | miRTarBase[42] | 4944 |
| Drug-protein | DrugBank[43] | 11,107 |
| lncRNA-disease | LncRNADisease[44], LncRNASNP2[45] | 1264 |
| Protein–protein | STRING[46] | 19,237 |
| miRNA-lncRNA | lncRNASNP2[45] | 8374 |
| lncRNA-protein | LncRNA2Target[47] | 690 |
| Drug-disease | CTD[48] | 18,416 |
| Protein-disease | DisGeNET[49] | 25,087 |
| Total | | 105,546 |

**Table 2.** The number and database of different types of associations in MAN.

tained in the miRNA sequences comprehensively. In this study, we chose chaos game representation (CGR)[50] to quantize position and nonlinear information to calculate miRNA sequence similarity by pearson coefficient. Firstly, the positions of four nucleotides of miRNA are mapped to Euclidean space by the following formula:

$$T_i = T_{i-1} + c * (T_{i-1} - G_i) \tag{1}$$

$$G_i = \begin{cases} (0,0), \ if \ type \ of \ nucleotide \ is \ A \\ (0,1), \ if \ type \ of \ nucleotide \ is \ C \\ (1,0), \ if \ type \ of \ nucleotide \ is \ U \\ (1,1), \ if \ type \ of \ nucleotide \ is \ G \end{cases} \tag{2}$$

where $T_i$ is the position of $i$th nucleotide, and it is related to the position of the previous nucleotide $T_{i-1}$ and the nucleotide coefficient $G_i$. In this paper, the contribution parameter $c$ is equal to 0.5 and $T_0$ is (0.5, 0.5).

Secondly, we divided the CGR space into 64 subspaces as shown in Fig. 3. The attribute information of each subspace $SS_i$ would be represented by integrating the position information $X_i$, $Y_i$ and nonlinear information $Z_i$ by the following formula:

$$X_i = \sum x, \quad if \ point \ in \ subspace \ SS_i \tag{3}$$

$$Y_i = \sum y, \quad if \ point \ in \ subspace \ SS_i \tag{4}$$

$$Z_i = \frac{num_i - \frac{\sum_{t=1}^{64} num_t}{64}}{\sqrt{\frac{1}{64}\sum_{r=1}^{64}(num_r - \frac{\sum_{t=1}^{64} num_t}{64})^2}} \tag{5}$$

$$SS_i = (X_i, Y_i, Z_i), i = 1, 2, \ldots, 64 \tag{6}$$

where $num_i$ is the number of points in subspace $SS_i$.

Finally, each miRNA sequence information could be represented by the descriptor $m(i)$. And we calculate sequence similarity $M_{sim}(m(i), m(j))$ by Pearson correlation coefficient.

$$m(i) = (SS_i, SS_2, \ldots, SS_{64}) \tag{7}$$

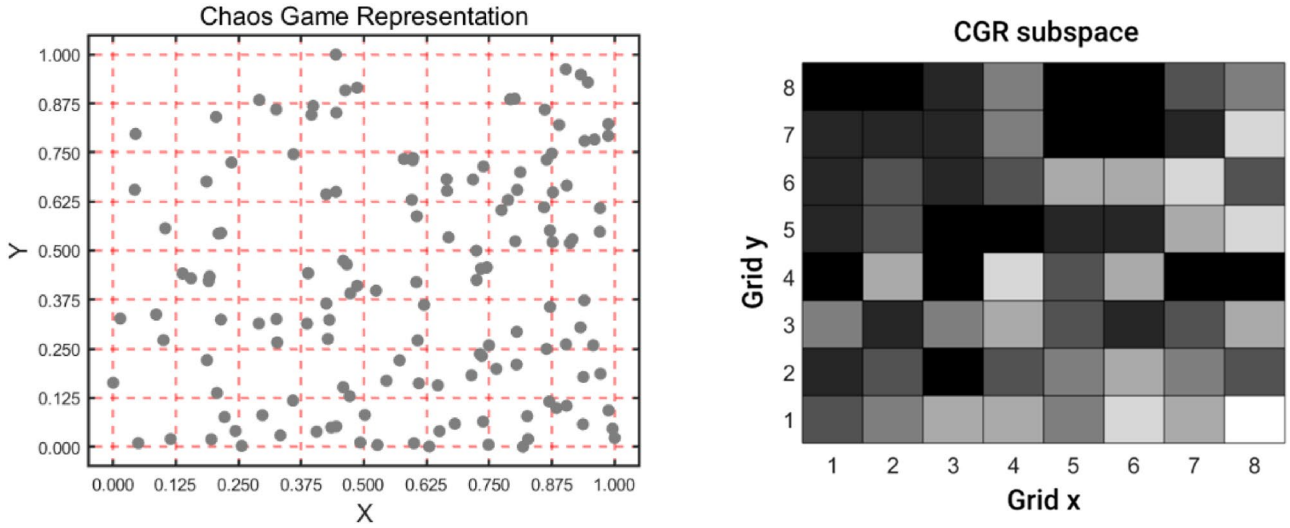


**Figure 3.** The CGR of has-mir-3976 plotted in 8 × 8 subspaces and the matrix of its nucleotides with probabilities for chaos game representation.

$$M_{sim}\big(m(i), m(j)\big) = \frac{Cov(m(i), m(j))}{m(i) \times m(j)} \tag{8}$$

**Disease semantic similarity.** In this study, the Directed Acyclic Graph (DAG)[51] of diseases could be obtained from the Medical Subject Headings (Mesh)[52]. In the system, a disease $d(a)$ could be defined by $DAG(d(a)) = (L(d(a)), E(d(a)))$, where $L(d(a))$ is a node set including $d(a)$ and ancestor nodes of $d(a)$, and $E(d(a))$ indicates directed edge set of all relationships from ancestor node to child node. The semantic value of $d(a)$ was contributed by term $T$ as the formula:

$$\begin{cases} D_{d(a)}(T) = 1 & if\ T = d(a) \\ D_{d(a)}(T) = max\{\vartheta * D_{d(a)}(T') | T' \in children\ of\ T\} & if\ T \neq d(a) \end{cases} \tag{9}$$

where $\vartheta$ is a parameter of semantic contribution, and $\vartheta$ is equal to 0.5 as previous study. Therefore, $DV(D)$ of $D$ could be calculated as follows:

$$DV(D) = \sum\nolimits_{T \in A_D} D_D(T) \tag{10}$$

According the assumption that two diseases should have higher similarity if they hold more same parts in DAG, the similarity of the diseases $d(a)$ with $d(b)$ could be obtained as follows:

$$S(d(a), d(b)) = \frac{\sum_{T \in A_{d(a)} \cap A_{d(b)}} (D_{d(a)}(T) + D_{d(b)}(T))}{DV(d(a)) + DV(d(b))} \tag{11}$$

**Structural deep network embedding.** Since existing network embedding algorithms could not keep the high-order proximity of large-scale networks, this paper adopted the structural deep network embedding (SDNE) to extract the behavior information of miRNAs and diseases. Many existing network embedding models are shallow model (e.g. Laplacian Eigenmaps[53], Graph Factorization[54]), which are unable to validly extract the highly non-linear structural information of network. SDNE is a semi-supervised model for network embedding. For the part of supervised, first-order similarity based on Laplacian matrix would be adopted to preserve local network information. And the part of unsupervised, SDNE used deep autoencoder modeling second-order similarity to save the global network information. Therefore, the loss function of SDNE is divided into two parts, i.e. Laplacian matrix model and Deep autoencoder model.

*First-order similarity.* To make adjacent nodes of graph closer in the latent space, the loss function of first-order similarity could be obtained as following formula:

$$L_{1st} = \sum\nolimits_{i,j=1}^{n} s_{i,j} \| y_i^{(k)} - y_j^{(k)} \|_2^2 = \sum\nolimits_{i,j=1}^{n} s_{i,j} \| y_i - y_j \|_2^2 \tag{12}$$

where $s_{i,j}$ is the adjacency matrix for heterogeneous information network and $y_i^{(k)}$ indicates the node $i$ of $k$-th layer.

*Second-order similarity.* For the capturing of global structure information, SDNE construct the deep autoencoder model. Any given $x_i$ could be convert into the latent representation of $k$th layer as:

$$y_i^{(1)} = \sigma\left(W^{(1)}x_i + b^{(1)}\right) \tag{13}$$

$$y_i^{(k)} = \sigma\left(W^{(k)}y_i^{(k-1)} + b^{(k)}\right), k = 2, \ldots, K \tag{14}$$

here $W^{(k)}$ is the $k$th layer weight matrix and $b^{(k)}$ as a parameter. According the optimization goal of the autoencoder is to reduce the reconstruction error in input and output, therefore, we could define the loss function as follows:

$$L = \sum\nolimits_{i=1}^{n} \|\widehat{x_i} - x_i\|_2^2 \tag{15}$$

The adjacency matrices are often very sparse, which means zero elements are far more than non-zero elements. Therefore, the loss function would be optimized as:

$$L_{2\text{nd}} = \sum\nolimits_{i=1}^{n} \|(\widehat{x_i} - x_i) \odot b_i\|_2^2 = \|(\widehat{X} - X) \odot B\|_{\text{F}}^2 \tag{16}$$

where $\odot$ is the Hadamard product (multiplying the corresponding elements).

Integrating the first-order similarity and second-order similarity, the finally loss function of SDNE is shown as follows:

$$L_{mix} = L_{2nd} + \alpha L_{1st} + \upsilon L_{reg} = \|(\widehat{X} - X) \odot B\|_{\text{F}}^2 + \alpha \sum\nolimits_{i,j=1}^{n} s_{i,j}\|y_i - y_j\|_2^2 + \upsilon L_{reg} \tag{17}$$

where $L_{reg}$ is a regularization term, and $\alpha$ is a parameter to control the loss of the first-order similarity. The regularization term is shown as:

$$L_{reg} = \frac{1}{2}\sum_{k=1}^{K}\left(W_F^{(k)2} + \hat{W}_F^{(k)2}\right) \tag{18}$$

**Integration of feature information.** In this study, we firstly obtained miRNA sequence similarity and disease semantic similarity and convert them into attribute feature information $M_{sim}(i)$, $D_{sim}(j)$ of same dimension by stacked autoencoder. The dimension of $M_{sim}(i)$ and $D_{sim}(j)$ is 64. After then, the behavior feature information of miRNAs $M_b(i)$ and diseases $D_b(j)$ were extracted by the structural deep network embedding based on the molecular association network. The dimension of $M_b(i)$ and $D_b(j)$ is 128. Finally, a complete sample feature descriptor is constructed by fusing above information based on the HMDD v3.0 database. The feature descriptor was a 384-dimensional vector as follows:

$$FD(i,j) = \left[M_b(i), M_{sim}(i), D_b(j), D_{sim}(j)\right] \tag{19}$$

**Convolutional neural network algorithm.** Convolutional neural network (CNN) is a deep-structured feedforward neural network with convolution calculations. CNN could shift-invariant classify the input information based on layer structure by representation learning capability. With the development of research, CNN has been successfully utilized in bioinformatics[55]. Therefore, in this paper, we adopted the CNN to train and predict potential miRNA-disease association. Specifically, CNN has a multi-layer structure including input, convolutional layer, pooling layer, fully-connected layer and output as shown in Fig. 4. The input layer is a matrix of all feature descriptor $FD(i,j)$ with size 26284 × 384. Two convolutional layers $C1$ and $C2$ are obtained by 32 filters with 3 × 1 convolution kernel and 64 filters with 3 × 1 convolution kernel. In this study, we adopted max-pooling 2 × 1 kernel to subsample the $C2$. After repeatedly convolution and pooling, CNN classifies the features from fully-connected layer and output the probability distribution.

## Results and discussion

**Performance evaluation.** In this experiment, we implemented the five-fold cross validation to evaluate the performance of proposed model under HMDD v3.0[37]. These known miRNA-disease pairs would be randomly split into five subsets with no intersection. Each cross validation, one of five subsets would be set as test set and remaining data sets as train set. To avoid the revelation of test data, we constructed the heterogeneous information network by only training data and extract the behavior information. In this study, a class of evaluation criteria were used to assess SDNE-MDA, including accuracy (Acc.), sensitivity (Sen.), specificity (Spec.), precision (Prec.), Matthews Correlation Coefficient (MCC) and area under curve (AUC). As a result, the average Acc, Sen, Spec, Prec, MCC and AUC achieved 87.38%, 87.28%, 87.47%, 87.45%, 74.76% and 0.9447 with standard deviations of 0.44%, 0.93%, 1.01%, 0.82%, 0.88% and 0.0027, respectively as shown in Table 3. In addition, the receiver operating characteristics (ROC) curve and area under precision-recall (PR) curve by SDNE-MDA based on HMDD are shown in Fig. 5.

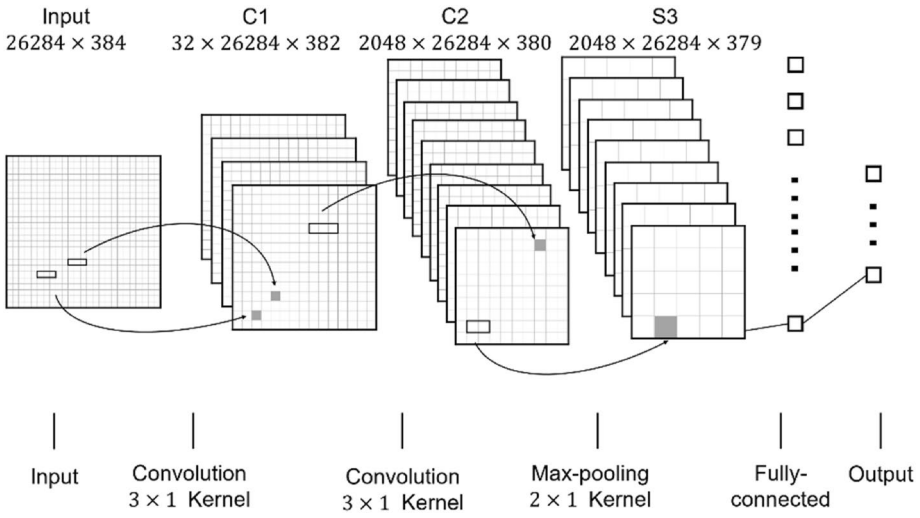


**Figure 4.** Structure of the CNN algorithm.

| Evaluation criteria | Result |
|---|---|
| Acc. (%) | 87.38 ± 0.44 |
| Sen. (%) | 87.28 ± 0.93 |
| Spec. (%) | 87.47 ± 1.01 |
| Prec. (%) | 87.45 ± 0.82 |
| MCC (%) | 74.76 ± 0.88 |
| AUC | 0.9447 ± 0.0027 |

**Table 3.** Five-fold cross validation results performed by SDNE-MDA on HMDD v3.0.

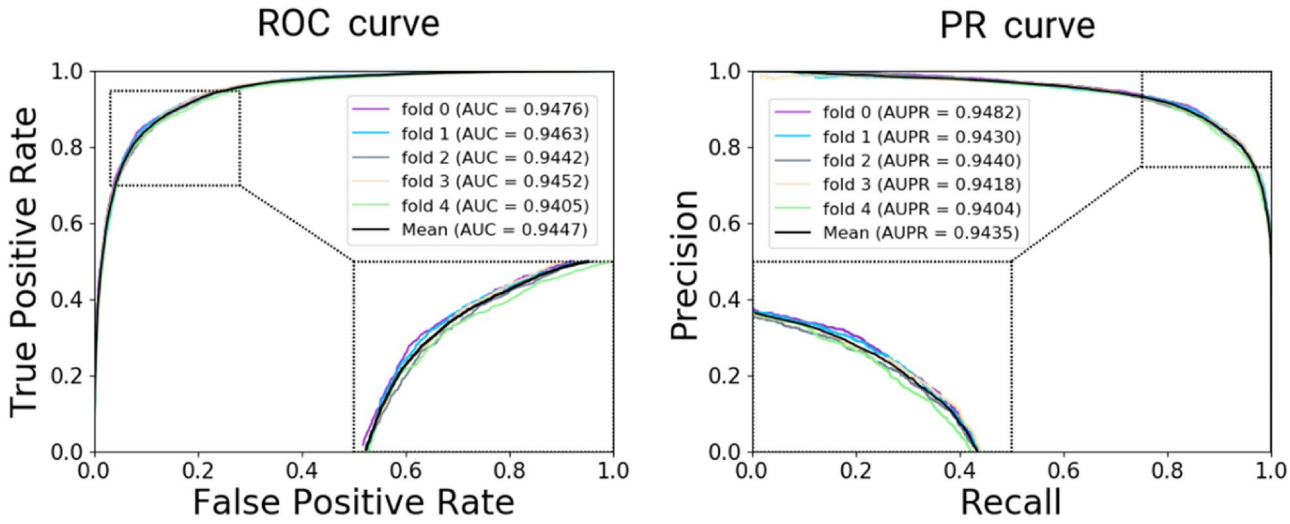


**Figure 5.** The ROC and PR curves performed in terms of five-fold cross validation by SDNE-MDA on HMDD v3.0.

**Comparison with different feature extraction methods.** In this study, these nodes in the network could be represented by the attribute and behavior information. Both types of information may influence the result of prediction, so we compared the different feature extraction methods including SDNE-MDA_AI composed of attribute information, SDNE-MDA_BI composed of behavior information and SDNE-MDA composed of both them. In addition, attribute information of other nodes has scarcely effect on prediction of potential miRNA-disease relationships. For reducing the redundancy of model, we only considered the attribute information of miRNAs and diseases. The detail result of comparison between proposed model with different feature

| Feature | Acc. (%) | Sen. (%) | Spec. (%) | Prec. (%) | MCC (%) | AUC |
|---|---|---|---|---|---|---|
| SDNE-MDA_AI | 79.60 ± 0.35 | 81.29 ± 1.87 | 77.92 ± 1.43 | 78.65 ± 0.73 | 59.26 ± 0.73 | 0.8636 ± 0.0037 |
| SDNE-MDA_BI | 83.95 ± 0.72 | 83.08 ± 6.30 | 84.83 ± 5.94 | 84.95 ± 4.07 | 68.32 ± 1.27 | 0.9187 ± 0.0048 |
| SDNE-MDA | 87.38 ± 0.44 | 87.28 ± 0.93 | 87.47 ± 1.01 | 87.45 ± 0.82 | 74.76 ± 0.88 | 0.9447 ± 0.0027 |

**Table 4.** The comparison results between SDNE-MDA_AI model, SDNE-MDA_BI model and SDNE-MDA model based on HMDD database.

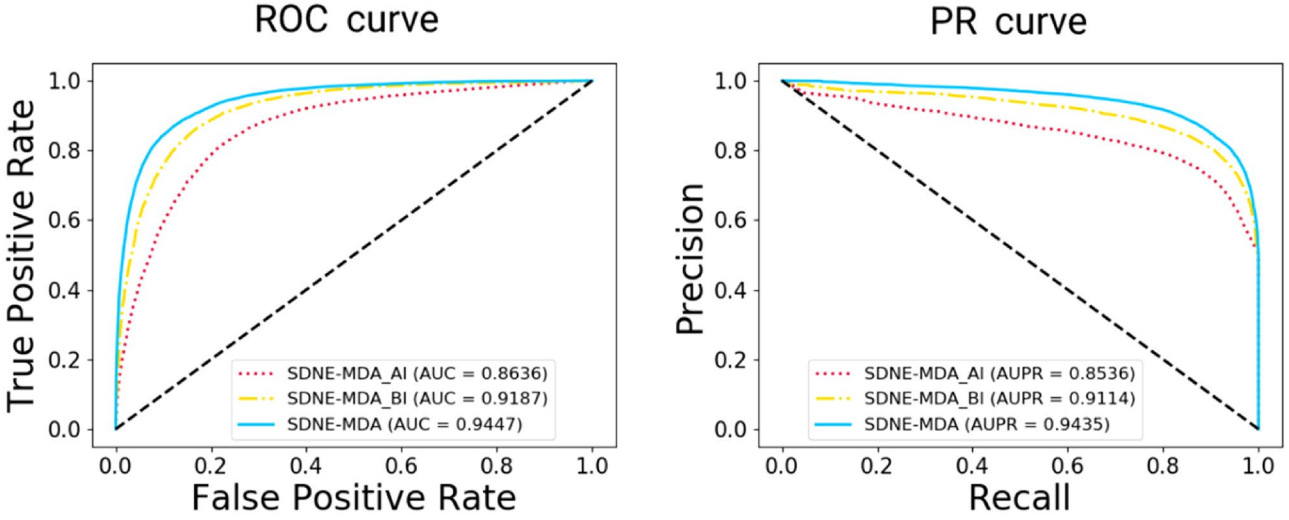


**Figure 6.** ROC and PR curves performed by SDNE-MDA_AI, SDNE-MDA_BI and SNDE-MDA model in terms of five-fold cross validation based on HMDD database.

| Model | Acc. (%) | Sen. (%) | Spec. (%) | MCC (%) | AUC |
|---|---|---|---|---|---|
| SDNE-MDA | 87.38 ± 0.44 | 87.28 ± 0.93 | 87.47 ± 1.01 | 74.76 ± 0.88 | 0.9447 ± 0.0027 |
| Bagging | 84.52 ± 0.62 | 84.77 ±0.80 | 84.27 ± 1.34 | 69.05 ± 1.23 | 0.8985 ± 0.0042 |
| LogisticRegression | 85.13 ± 0.86 | 84.42 ± 0.92 | 85.84 ± 1.19 | 70.27 ± 1.71 | 0.9272 ± 0.0080 |
| NaiveBayes | 75.90 ± 1.27 | 60.04 ± 3.94 | 91.76 ± 1.64 | 54.68 ± 1.77 | 0.8881 ± 0.0059 |
| Adaboost | 85.69 ± 0.51 | 84.74 ± 1.72 | 86.63 ± 2.07 | 71.43 ± 1.06 | 0.9226 ± 0.0036 |
| MLP | 85.78 ± 1.06 | 84.75 ± 0.87 | 86.82 ± 2.78 | 71.72 ± 1.83 | 0.9320 ± 0.0051 |

**Table 5.** The comparison results between SDNE-MDA with other four different classifier models in terms of five-fold cross validation based on HMDD v3.0 database.

extraction models are shown in Table 4. The accuracy of SDNE-MDA is 7.78% and 3.43% higher than that of SDNE-MDA_AI and SDNE-MDA_BI, respectively. In addition, the AUC of proposed model is 0.0811 and 0.0260 higher than SDNE-MDA_AI and SDNE-MDA_BI. The ROC curves and PR curves of three experiments are shown in Fig. 6. These results indicated that integrating the two kind of information to represent the node achieved more distinguished performance.

**Comparison with different classifier models.** In this study, the CNN was adopted to train and identify potential relationships between miRNA and disease. To further evaluate SDNE-MDA, we compare proposed model with Bagging, Logistic Regression, Naive Bayes and Adaboost classifier model. In this experiment, we implemented the five-fold cross validation in these different classifier models based on the HMDD v3.0. Finally, the proposed model yielded average AUC of 0.9447 based on five-fold cross validation and outperformed Bagging (0.8998), LogisticRegression (0.9270), Naive Bayes (0.8881), Adaboost (0.9226) and MLP (0.9320). The AUC of CNN is 0.0259 higher than the mean AUC of all five model, and the accuracy is 1.60% higher than that of the second highest methods. The detail results of the comparison between SDNE-MDA and other four classifier models are shown in Table 5, and we drew the ROC curves as shown in Fig. 7. Therefore, CNN algorithm is the optimal selection for the proposed model to predicting potential miRNA-disease associations.

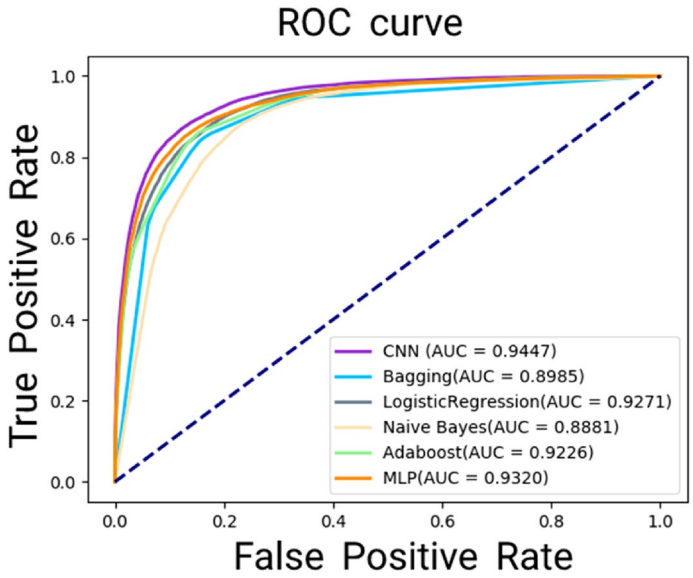


**Figure 7.** Performance comparison between SDNE-MDA with other four different classifier models based on HMDD v3.0 database.

| Method | AUC |
|---|---|
| RWRMDA | 0.8617 |
| MTDN | 0.8872 |
| EGBMMDA | 0.9048 |
| LMTRDA | 0.9054 |
| DBMDA | 0.9129 |
| PBMDA | 0.9172 |
| SDNE-MDA | 0.9447 |

**Table 6.** The comparison results between SDNE-MDA with other related works.

**Comparison with related work.** An increasing number of researchers have focused on the prediction of miRNA-disease associations, and a mass of model have been proposed. To further evaluate the predictive performance of our method, the SDNE-MDA was compared with six state-of-the-art classical methods under five-fold cross validation, including RWRMDA[56], MTDN[57], EGBMMDA[32], LMTRDA[58], DBMDA[59] and PBMDA[31]. Since these algorithms have not calculated multiple evaluation criteria, we only compare the AUC on the terms of five-fold cross validation based HMDD database. The detail results of the comparison between SDNE-MDA and other six related works are shown in Table 6. The proposed method is 0.0399 higher than the average AUC of all algorithms, and 0.0275 higher than that of the second highest methods. This is mainly due to SDNE-MDA integrated two types of information of miRNAs and diseases, and extract the feature more comprehensively. Therefore, the proposed model is an effective and reliable computational tool for predicting potential miRNA-disease associations.

**Case studies.** For further evaluating the prediction ability of SDNE-MDA, we implemented case studies based on three significant human diseases (Breast Neoplasms, Kidney Neoplasms, Lymphoma). In this study, these known miRNA-disease associations based on HMDD v3.0 database would be the training set. To avoid the overlap in the train data and prediction list, the test set is the unknown relationship pairs between three diseases and all possible miRNAs. As a result, 47, 46 and 46 of top-50 candidate related miRNAs were confirmed by independent databases. Therefore, SDNE-MDA is a feasible and reliable model for predicting potential relationships between miRNA and disease.

Breast Neoplasms is the most universal neoplasms in female and the risk of breast cancer is up to 13% in the United States. Although men may also develop breast cancer, 99% of patients are women. There are approximately 276,480 novel cases in women and 42,170 were die from breast cancer in 2020[60]. In previous few years, studies had indicated the expression level of miRNA have strong impact to growth and division of breast tumor cell[61]. Therefore, we implemented a case study of Breast Neoplasms-miRNA associations by SDNE-MDA. In the

| Rank | miRNA | Evidence | Rank | miRNA | Evidence |
|---|---|---|---|---|---|
| 1 | hsa-miR-124-3p | dbdemc | 26 | hsa-miR-200b-3p | dbdemc |
| 2 | hsa-miR-483-5p | dbdemc | 27 | hsa-miR-181d-5p | dbdemc |
| 3 | hsa-miR-200c-3p | dbdemc | 28 | hsa-miR-23b-3p | dbdemc |
| 4 | hsa-miR-101-3p | dbdemc | 29 | hsa-miR-532-5p | dbdemc |
| 5 | hsa-miR-27a-3p | dbdemc | 30 | hsa-miR-193b-3p | dbdemc |
| 6 | hsa-miR-28-5p | dbdemc | 31 | hsa-miR-126-3p | dbdemc |
| 7 | hsa-miR-455-5p | dbdemc | 32 | hsa-miR-92b-3p | dbdemc |
| 8 | hsa-miR-186-5p | dbdemc | 33 | hsa-miR-539-5p | dbdemc |
| 9 | hsa-miR-99b-5p | dbdemc | 34 | hsa-mir-138-2-3p | Unconfirmed |
| 10 | hsa-miR-141-3p | dbdemc | 35 | hsa-miR-506-3p | dbdemc |
| 11 | hsa-miR-330-5p | dbdemc | 36 | hsa-miR-223-3p | dbdemc |
| 12 | hsa-miR-19b-2-5p | dbdemc | 37 | hsa-miR-19a-3p | dbdemc |
| 13 | hsa-miR-154-5p | dbdemc | 38 | hsa-miR-29c-3p | dbdemc |
| 14 | hsa-miR-744-5p | dbdemc | 39 | hsa-miR-188-5p | dbdemc |
| 15 | hsa-miR-1271-5p | dbdemc | 40 | hsa-miR-25-3p | dbdemc |
| 16 | hsa-miR-377-3p | dbdemc | 41 | hsa-miR-300 | dbdemc |
| 17 | hsa-miR-200a-3p | dbdemc | 42 | hsa-miR-376b-3p | dbdemc |
| 18 | hsa-miR-211-5p | dbdemc | 43 | hsa-mir-208b-5p | Unconfirmed |
| 19 | hsa-miR-216a-5p | dbdemc | 44 | hsa-miR-376a-3p | dbdemc |
| 20 | hsa-miR-449b-5p | dbdemc | 45 | hsa-miR-543 | dbdemc |
| 21 | hsa-miR-346 | dbdemc | 46 | hsa-miR-130a-3p | dbdemc |
| 22 | hsa-miR-328-3p | dbdemc | 47 | hsa-miR-302a-3p | dbdemc |
| 23 | hsa-miR-494-3p | dbdemc | 48 | hsa-miR-29a-3p | dbdemc |
| 24 | hsa-mir-885-5p | Unconfirmed | 49 | hsa-miR-302e | dbdemc |
| 25 | hsa-miR-202-3p | dbdemc | 50 | hsa-miR-363-3p | dbdemc |

**Table 7.** Prediction of top 50 miRNAs related to Breast Neoplasms based on known miRNA-disease associations in HMDD V3.0 database.

prediction list shown as Table 7, 47 of top 50 predicted Breast Neoplasms related miRNAs were verified based on independent databases.

Kidney Neoplasms is a novel cancer with higher adult incidence[60]. In the past few years, however, morbidity and mortality of kidney neoplasms have been increasing. There are about 73,750 novel cases in kidney neoplasms with about 45,520 in male and about 28,230 in female in United States and about 14,830 deaths for this cancer (9860 men and 4970 women) in 2020. Recently, increasing researchers have indicated miRNAs are related with kidney neoplasms[62]. Thus, we take Kidney Neoplasms as a case study for SDNE-MDA and prioritize the candidate miRNAs. In the prediction list shown as Table 8, 46 of top-50 potential kidney neoplasms-related miRNAs were confirmed by independent databases.

Lymphoma is one of the most common malignant cancers (~4% of all new cancer) especially in teenagers in United States[60]. Lymphoma mainly contains two types of Hodgkin Lymphoma (HL) and non-Hodgkin Lymphoma (NHL). In 2020, it is estimated that about 85,720 new cases of Lymphoma (47,070 of men and 38,650 of women) and 20,910 deaths for HL and NHL (12,030 of men and 8,880 of women). Therefore, we implemented SDNE-MDA to prioritize possible miRNAs for Lymphoma based on HMDD v3.0. As shown in Table 9, 46 out of top 50 predicted Lymphoma candidate miRNAs were verified by independent databases.

## Conclusion

In previous few years, accumulating number of researches demonstrated that miRNAs have closely link with diseases. Various of biological experiments and computational methods are committed to classify the association of them. In this paper, we proposed a structural deep network embedding-based model SDNE-MDA to predict miRNA-disease associations. This model constructed a complex network MAN by fusing miRNAs, diseases and three related molecular (lncRNA, drug and protein) with their relationships. Through the comprehensive heterogeneous information network, potential miRNA-disease associations could be predicted more accurate and efficient. And CNN is utilized to train and classify the potential miRNA-disease associations. Compared with other classifiers and feature extraction models, SDNE-MDA showed outstanding performance. In addition, case studies were implemented on three significant human disease for further validate performance of SDNE-MDA. As a result, 47, 46 and 46 of top-50 predicted miRNAs have been confirmed by independent databases. These results demonstrated that SDNE-MDA is a reliable computational tool for predicting miRNA-disease associations.

| Rank | miRNA | Evidence | Rank | miRNA | Evidence |
|---|---|---|---|---|---|
| 1 | hsa-mir-146a-5p | dbdemc | 26 | hsa-mir-19a-5p | dbdemc |
| 2 | hsa-mir-223-5p | dbdemc | 27 | hsa-mir-133a-5p | Unconfirmed |
| 3 | hsa-mir-125b-5p | dbdemc | 28 | hsa-mir-29b-3p | dbdemc |
| 4 | hsa-mir-145-5p | dbdemc | 29 | hsa-mir-222-5p | dbdemc |
| 5 | hsa-mir-150-5p | dbdemc | 30 | hsa-mir-29c-5p | dbdemc |
| 6 | hsa-mir-181a-5p | dbdemc | 31 | hsa-mir-18a-5p | dbdemc |
| 7 | hsa-mir-182-5p | dbdemc | 32 | hsa-mir-1-3p | dbdemc |
| 8 | hsa-mir-26a-5p | dbdemc | 33 | hsa-mir-181b-5p | dbdemc |
| 9 | hsa-mir-9-5p | dbdemc | 34 | hsa-mir-206 | dbdemc |
| 10 | hsa-mir-31-5p | dbdemc | 35 | hsa-mir-124-5p | Unconfirmed |
| 11 | hsa-mir-16-5p | dbdemc | 36 | hsa-mir-205-5p | Unconfirmed |
| 12 | hsa-mir-143-5p | dbdemc | 37 | hsa-mir-23a-5p | dbdemc |
| 13 | hsa-mir-221-5p | dbdemc | 38 | hsa-let-7c-5p | dbdemc |
| 14 | hsa-mir-20a-5p | dbdemc | 39 | hsa-mir-22-5p | dbdemc |
| 15 | hsa-mir-26b-5p | dbdemc | 40 | hsa-mir-34b-5p | dbdemc |
| 16 | hsa-let-7b-5p | dbdemc | 41 | hsa-mir-19b-3p | dbdemc |
| 17 | hsa-mir-92a-3p | dbdemc | 42 | hsa-mir-132-5p | dbdemc |
| 18 | hsa-mir-29a-5p | dbdemc | 43 | hsa-mir-106b-5p | dbdemc |
| 19 | hsa-mir-375-5p | Unconfirmed | 44 | hsa-mir-34c-5p | dbdemc |
| 20 | hsa-mir-142-5p | dbdemc | 45 | hsa-mir-100-5p | dbdemc |
| 21 | hsa-let-7a-5p | dbdemc | 46 | hsa-mir-124-3p | dbdemc |
| 22 | hsa-mir-122-5p | dbdemc | 47 | hsa-mir-125a-5p | dbdemc |
| 23 | hsa-mir-146b-5p | dbdemc | 48 | hsa-mir-148a-5p | dbdemc |
| 24 | hsa-mir-30a-5p | dbdemc | 49 | hsa-mir-200b-5p | dbdemc |
| 25 | hsa-mir-24-3p | dbdemc | 50 | hsa-mir-486-5p | dbdemc |

**Table 8.** Prediction of top 50 miRNAs related to Kidney Neoplasms based on known miRNA-disease associations in HMDD V3.0 database.

| Rank | miRNA | Evidence | Rank | miRNA | Evidence |
|---|---|---|---|---|---|
| 1 | hsa-mir-34a-5p | dbdemc | 26 | hsa-mir-138-5p | dbdemc |
| 2 | hsa-mir-223-5p | dbdemc | 27 | hsa-mir-106a-5p | dbdemc |
| 3 | hsa-mir-125b-5p | dbdemc | 28 | hsa-mir-34b-5p | dbdemc |
| 4 | hsa-mir-145-5p | dbdemc | 29 | hsa-mir-140-5p | dbdemc |
| 5 | hsa-mir-182-5p | dbdemc | 30 | hsa-mir-132-5p | dbdemc |
| 6 | hsa-mir-27a-5p | Unconfirmed | 31 | hsa-mir-106b-5p | dbdemc |
| 7 | hsa-mir-9-5p | dbdemc | 32 | hsa-mir-100-5p | dbdemc |
| 8 | hsa-mir-26b-5p | dbdemc | 33 | hsa-mir-34c-5p | dbdemc |
| 9 | hsa-let-7b-5p | dbdemc | 34 | hsa-mir-148a-5p | dbdemc |
| 10 | hsa-mir-29a-5p | dbdemc | 35 | hsa-mir-124-3p | dbdemc |
| 11 | hsa-let-7a-5p | dbdemc | 36 | hsa-mir-25-5p | dbdemc |
| 12 | hsa-mir-192-5p | dbdemc | 37 | hsa-let-7i-5p | dbdemc |
| 13 | hsa-mir-146b-5p | dbdemc | 38 | hsa-mir-335-5p | dbdemc |
| 14 | hsa-mir-30a-5p | dbdemc | 39 | hsa-mir-141-5p | Unconfirmed |
| 15 | hsa-mir-24-3p | dbdemc | 40 | hsa-mir-99a-5p | dbdemc |
| 16 | hsa-mir-214-5p | dbdemc | 41 | hsa-mir-107 | dbdemc |
| 17 | hsa-mir-96-5p | dbdemc | 42 | hsa-mir-15b-5p | dbdemc |
| 18 | hsa-mir-183-5p | dbdemc | 43 | hsa-mir-144-5p | dbdemc |
| 19 | hsa-mir-206 | dbdemc | 44 | hsa-let-7e-5p | dbdemc |
| 20 | hsa-mir-181b-5p | dbdemc | 45 | hsa-mir-30d-5p | dbdemc |
| 21 | hsa-mir-1-3p | dbdemc | 46 | hsa-mir-218-5p | dbdemc |
| 22 | hsa-let-7c-5p | dbdemc | 47 | hsa-mir-130a-5p | Unconfirmed |
| 23 | hsa-mir-205-5p | dbdemc | 48 | hsa-mir-429 | Unconfirmed |
| 24 | hsa-mir-124-5p | dbdemc | 49 | hsa-mir-101-5p | dbdemc |
| 25 | hsa-mir-23a-5p | dbdemc | 50 | hsa-mir-195-5p | dbdemc |

**Table 9.** Prediction of top 50 miRNAs related to Lymphoma based on known miRNA-disease associations in HMDD V3.0 database.

Received: 29 November 2020; Accepted: 30 April 2021
Published online: 16 June 2021

## References


1. Kloosterman, W. P. & Plasterk, R. H. A. The diverse functions of microRNAs in animal development and disease. *Dev. Cell* **11**, 441–450 (2006).
2. Ji, B.-Y. *et al.* Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model. *Sci. Rep.* **10**, 6658 (2020).
3. Ines, A. G. & Miska, E. A. MicroRNA functions in animal development and human disease. *Development* **132**, 4653–4662 (2005).
4. Guo, Z.-H. *et al.* A learning based framework for diverse biomolecule relationship prediction in molecular association network. *Commun. Biol.* **3**, 1–9 (2020).
5. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: From microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162 (2018).
6. Cheng, A. M., Byrom, M. W., Jeffrey, S. & Ford, L. P. Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. *Nucleic Acids Res.* **33**, 1290–1297 (2005).
7. Xantha, K. & Victor, A. Developmental biology. Encountering microRNAs in cell fate signaling. *Science* **310**, 1288–1289 (2005).
8. Miska, E. A. How microRNAs control cell division, differentiation and death. *Curr. Opin. Genet. Dev.* **15**, 563–568 (2005).
9. Xu, P., Guo, M. & Hay, B. A. MicroRNAs and the regulation of cell death. *Trends Genet.* **20**, 617–624 (2004).
10. Ramiro, G., Guido, M. & Croce, C. M. Targeting microRNAs in cancer: Rationale, strategies and challenges. *Nat. Rev. Drug Discov.* **9**, 775–789 (2010).
11. Farazi, T. A., Spitzer, J. I., Pavel, M. & Thomas, T. miRNAs in human cancer. *J. Pathol.* **223**, 102–115 (2015).
12. You, Z.-H. *et al.* PRMDA: Personalized recommendation-based miRNA-disease association prediction. *Oncotarget* **8**, 85568 (2017).
13. Wang, L. *et al.* Using two-dimensional principal component analysis and rotation forest for prediction of protein–protein interactions. *Sci. Rep.* **8**, 12874 (2018).
14. Bartels, C. L. & Tsongalis, G. J. MicroRNAs: Novel biomarkers for human cancer. *Clin. Chem.* **55**, 623–631 (2009).
15. Zheng, K. *et al.* MLMDA: A machine learning approach to predict and validate microRNA-disease associations by integrating of heterogenous information sources. *J. Transl. Med.* **17**, 1–14 (2019).
16. Chen, X., Xie, D., Zhao, Q. & You, Z.-H. MicroRNAs and complex diseases: From experimental results to computational models. *Brief. Bioinform.* **20**, 515–539 (2019).
17. Yohei, S. *et al.* Downregulation of miRNA-200c links breast cancer stem cells with normal stem cells. *Cell* **138**, 592–603 (2009).
18. Liu, B. *et al.* MiR-26a enhances metastasis potential of lung cancer cells via AKT pathway by targeting PTEN. *BBA Mol. Basis Disease* **1822**, 1692–1704 (2012).
19. Thum, T. *et al.* MicroRNA-21 contributes to myocardial disease by stimulating MAP kinase signalling in fibroblasts. *Nature* **456**, 980–984 (2008).
20. Chen, X. *et al.* WBSMDA: Within and between score for miRNA-disease association prediction. *Sci. Rep.* **6**, 21106 (2016).
21. Weidhaas, J. Using microRNAs to understand cancer biology. *Lancet Oncol.* **11**, 136–146 (2010).
22. Jiang, Q. *et al.* Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst. Biol.* **4**, S2 (2010).
23. Xuan, P. *et al.* Correction: Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PLoS ONE* **8**, e70204 (2013).
24. Chen, X. *et al.* HGIMDA: Heterogeneous graph inference for miRNA-disease association prediction. *Oncotarget* **7**, 65257 (2016).
25. Wang, L., Wang, H.-F., Liu, S.-R., Yan, X. & Song, K.-J. Predicting protein–protein interactions from matrix-based protein sequence using convolution neural network and feature-selective rotation forest. *Sci. Rep.* **9**, 9848 (2019).
26. Huang, Z.-A. *et al.* PBHMDA: Path-based human microbe-disease association prediction. *Front. Microbiol.* **8**, 233 (2017).
27. Chen, X. Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. *Sci. Rep.* **5**, 13186 (2015).
28. Pasquier, C. & Gardès, J. Prediction of miRNA-disease associations with a vector space model. *Sci. Rep.* **6**, 27036 (2016).
29. Li, J.-Q., Rong, Z.-H., Chen, X., Yan, G.-Y. & You, Z.-H. MCMDA: Matrix completion for MiRNA-disease association prediction. *Oncotarget* **8**, 21187 (2017).
30. Ping, X. *et al.* Prediction of potential disease-associated microRNAs based on random walk. *Bioinformatics* **31**, 1805–1815 (2015).
31. You, Z. H. *et al.* PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction. *Plos Computat. Biol.* **13**, e1005455 (2017).
32. Chen, X., Huang, L., Xie, D. & Zhao, Q. EGBMMDA: Extreme gradient boosting machine for MiRNA-disease association prediction. *Cell Death Dis.* **9**, 3 (2018).
33. Huang, Y.-A. *et al.* EPMDA: An expression-profile based computational model for microRNA-disease association prediction. *Oncotarget* **8**, 87033 (2017).
34. Chen, X., Cheng, J.-Y. & Yin, J. Predicting microRNA-disease associations using bipartite local models and hubness-aware regression. *RNA Biol.* **15**, 1192–1205 (2018).
35. Guo, Z.-H., Yi, H.-C. & You, Z.-H. Construction and comprehensive analysis of a molecular association network via lncRNA–miRNA–disease–drug–protein graph. *Cells*, **8**(8), 866 (2019).
36. Wang, D., Peng, C. & Zhu, W. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 1225–1234 (2016).
37. Huang, Z. *et al.* HMDD v3.0: A database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res.* **47**, D1013–D1017 (2018).
38. Yang, Z. *et al.* dbDEMC 2.0: Updated database of differentially expressed miRNAs in human cancers. *Nucleic Acids Res.* **45**, D812–D818 (2017).
39. Jiang, Q. *et al.* miR2Disease: A manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.* **37**, D98-104 (2009).
40. Guo, Z.-H., *et al.* Integrative construction and analysis of molecular association network in human cells by fusing node attribute and behavior information. *Mol. Therapy-Nucleic Acids* **19**, 498–506 (2020).
41. Zhou, H. *et al.* HMDD v3. 0: a database for experimentally supported human microRNA–disease associations. *Nucleic Acids Res.* **47**(D1), D1013–D1017 (2018).
42. Chou, C.-H., *et al.* miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* **46**(D1), D296–D302 (2017).
43. Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nuclc Acids Res.* **46**, D1074 (2018).
44. Chen, G. *et al.* LncRNADisease: A database for long-non-coding RNA-associated diseases. *Nuclc Acids Res.* **41**, D983–D986 (2013).
45. Miao, Y., Liu, W., Zhang, Q. & Guo, A. lncRNASNP2: An updated database of functional SNPs and mutations in human and mouse lncRNAs. *Nucleic Acids Res.* **46**, D276–D280 (2018).
46. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, gkw937 (2017).

47. Cheng, L. *et al.* LncRNA2Target v2.0: A comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* **47**, D140–D144 (2019).
48. Davis, A. P. *et al.* The Comparative Toxicogenomics Database: Update 2019. *Nucleic Acids Res.* **47**, D948–D954 (2019).
49. Janet, P. *et al.* DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* D833–D839 (2017).
50. Jeffrey, H. J. Chaos game representation of gene structure. *Nucleic Acids Res.* **18**, 2163–2170 (1990).
51. Kalisch, M. & Buehlmann, P. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8**, 613–636 (2012).
52. Lipscomb, C. E. Medical subject headings (MeSH). *Bull. Med. Libr. Assoc.* **88**, 265 (2000).
53. Belkin, M. & Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**(6), 1373–1396 (2003).
54. Ahmed, A., Shervashidze, N., Narayanamurthy, S., Josifovski, V. & Smola, A. J. Distributed large-scale natural graph factorization. In *Proceedings of the 22nd international conference on World Wide Web*, 37–48 (2013).
55. Wang, L., You, Z.-H., Huang, Y.-A., Huang, D.-S. & Chan, K. C. An efficient approach based on multi-sources information to predict circRNA-disease associations using deep convolutional neural network. *Bioinformatics* **36**, 4038–4046 (2020).
56. Chen, X., Liu, M. X. & Yan, G. Y. RWRMDA: Predicting novel human microRNA-disease associations. *Mol. BioSyst.* **8**, 2792–2798 (2012).
57. Xu, J. *et al.* Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: Case study of prostate cancer. *Mol. Cancer Ther.* **10**, 1857–1866 (2011).
58. Wang, L. *et al.* LMTRDA: Using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities. *PLoS Computat. Biol.* **15**, e1006865 (2019).
59. Zheng, K. *et al.* Dbmda: A unified embedding for sequence-based miRNA similarity measure with applications to predict and validate miRNA-disease associations. *Mol. Therapy-Nucleic Acids* **19**, 602–611 (2020).
60. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA Cancer J. Clin.* **69**, 7–34 (2019).
61. Iorio, M. V. *et al.* MicroRNA gene expression deregulation in human breast cancer. *Can. Res.* **65**, 7065–7070 (2005).
62. Muhamed Ali, A. *et al.* A machine learning approach for the classification of kidney cancer subtypes using miRNA genome data. *Mol. Therapy-Nucleic Acids* **8**, 2422 (2018).


## Acknowledgements

The authors would like to thank all anonymous reviewers for their constructive advice.

## Author contributions

H.L., H.C., Z.Y. and L.W. conceived the algorithm, carried out analyses, prepared the data sets, carried out experiments, and wrote the manuscript. S.S., X.Y. and J.Y. analyzed experiments. All authors reviewed the manuscript.

## Funding

This work is supported in part by the National Natural Science Foundation of China, under Grant 61702444, in part by the West Light Foundation of The Chinese Academy of Sciences, under Grant 2018-XBQNXZ-B-008, in part by the Chinese Postdoctoral Science Foundation, under Grant 2019M653804, in part by the Tianshan Youth—Excellent Youth, under Grant 2019Q029, in part by the Qingtan scholar talent project of Zaozhuang University.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to L.W. or S.-J.S.