Front. Comput. Sci., 2024, 18(6): 186912
https://doi.org/10.1007/s11704-023-3610-y

RESEARCH ARTICLE

DMFVAE: miRNA-disease associations prediction based on deep matrix factorization method with variational autoencoder

Pijing WEI[1], Qianqian WANG[2], Zhen GAO[2], Ruifen CAO[2], Chunhou ZHENG (✉)[3]

1 Information Materials and Intelligent Sensing Laboratory of Anhui Province, Institutes of Physical Science and Information Technology, Anhui University, Hefei 230601, China

2 Key Lab of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei 230601, China

3 Key Lab of Intelligent Computing and Signal Processing of Ministry of Education, School of Artificial Intelligence, Anhui University, Hefei 230601, China

Abstract MicroRNAs (miRNAs) are closely related to numerous complex human diseases, therefore, exploring miRNA-disease associations (MDAs) can help people gain a better understanding of complex disease mechanism. An increasing number of computational methods have been developed to predict MDAs. However, the sparsity of the MDAs may hinder the performance of many methods. In addition, many methods fail to capture the nonlinear relationships of miRNA-disease network and inadequately leverage the features of network and neighbor nodes. In this study, we propose a deep matrix factorization model with variational autoencoder (DMFVAE) to predict potential MDAs. DMFVAE first decomposes the original association matrix and the enhanced association matrix, in which the enhanced association matrix is enhanced by self-adjusting the nearest neighbor method, to obtain sparse vectors and dense vectors, respectively. Then, the variational encoder is employed to obtain the nonlinear latent vectors of miRNA and disease for the sparse vectors, and meanwhile, node2vec is used to obtain the network structure embedding vectors of miRNA and disease for the dense vectors. Finally, sample features are acquired by combining the latent vectors and network structure embedding vectors, and the final prediction is implemented by convolutional neural network with channel attention. To evaluate the performance of DMFVAE, we conduct five-fold cross validation on the HMDD v2.0 and HMDD v3.2 datasets and the results show that DMFVAE performs well. Furthermore, case studies on lung neoplasms, colon neoplasms, and esophageal neoplasms confirm the ability of DMFVAE in identifying potential miRNAs for human diseases.

Keywords miRNA-disease association, deep matrix factorization, self-adjusted nearest neighbor, variational encoder, network structure

Received July 30, 2023; accepted December 6, 2023

E-mail: zhengch99@126.com

Special Issue—Bioinformatics (CCF CBC2022 Award Papers)

1 Introduction

MicroRNAs (miRNAs) are endogenous noncoding RNAs with a length of about 22−24 nucleotides [1,2], which play an important role in regulating various biological processes and are essential for living organisms [3]. Additionally, substantial evidence has indicated that miRNAs are implicated in a wide range of diseases [4,5]. For instance, it has been demonstrated that miR-21 is involved in several vital biological processes, including neurogenesis, differentiation and development [6]. Therefore, it is essential to predict miRNAs associated with major human diseases for understanding the pathogenesis of these diseases, as well as for their preventing, diagnosing and treating.

Due to the high cost and time-consuming nature of traditional biological experimental methods, various state-of-the-art computational methods were proposed. Chen et al. penned a comprehensive review of miRNAs and complex diseases [7], which not only delves into the function of miRNAs, miRNA-target interactions, associations between miRNAs and diseases, but also introduce a plethora of computational models and sheds light on several significant public miRNA-related databases. Chen et al. scrutinized 20 state-of-the-art computational models from diverse perspectives, enumerated the crucial factors associated with prediction, and outlined the future progression of computational models. Subsequently, Huang et al. updated the review in three areas [8–10]. In terms of computational models, they introduced 29 advanced models and specifically categorized them into model-based integration and non-integration approaches. In addition, they analyzed the strengths and weaknesses of each model category and proposed future research directions from different perspectives to improve model performance [8]. In terms of databases, they provided a detailed introduction of different databases and the mainstream web servers since 2017, and explained the principles for integrating different data sources using different methods [9]. While model performance is often illustrated by

cross-validation and case studies, there is no universally accepted strategy for evaluating MDA computational models. In terms of performance evaluation of computational models, they analyzed the results of 29 models and recommended a feasible evaluation workflow to facilitate a fair and systematic evaluation of prediction performance [10]. Based on the analysis of previous computational methods, we categorize the methods into three categories.

The first category is based on graph theory to predict miRNA-disease associations (MDAs). Some typical models in this category such as RWBRMDA [6], LWBRW [11], GMDA [12]. However, the predictive performance of graph-theoretical-based methods is heavily dependent on the availability of known MDA information. These methods may not perform as well when applied to new diseases or diseases containing only a few known related miRNAs.The second category of methods is based on matrix factorization algorithms. For instance, the LRMCMDA method predicted MDAs by using bipartite graphs to infer negative samples and low-rank matrix completion to obtain the final score [13]. MDHGI, proposed by Chen et al., is different from previous heterogeneous networks-based computing models, in which matrix decomposition is used before constructing heterogeneous networks to efficiently improve the prediction accuracy of the model [14]. Chen et al. constructed a new MDA prediction model, NCMCMDA, by transforming neighborhood constraint problems into optimization problems solved using fast iterative algorithms [15]. Lu et al. presented the PMDA method, which uses a neighborhood learning method to reduce redundancy and uses a matrix factorization method to predict MDAs [16]. Additionally, there are similar methods used to predict circRNA-disease associations (CDAs). For example, Zhang et al. proposed PCD_MVMF, which builds reliable heterogeneous networks by combining multiple pieces of data. The integrated heterogeneous network is then learned using metapath2vec++ and the final prediction is completed using matrix factorization [17]. Although traditional matrix factorization methods achieve good results, it is difficult to learn the underlying representations and complex data structures.The third category is based on deep learning models. Deep learning methods are capable to combine low-level features with nonlinear functions to form abstract high-level features, which enables them to discover effective feature representations for data. For example, CNNMDA adopted convolutional neural network (CNN) to predict potential MDAs [18]. Li et al. inferred disease-associated miRNAs using graph neural network-based encoder to aggregate the neighbor information of nodes [19]. VGAE-MDA [20] calculated the correlation score of the two subnetworks through two variational graph autoencoders, and combined the scores of the trained network to obtain the final score of MDAs. DFELMDA obtained a low-dimensional feature representation through autoencoder for each association pair and used ensemble learning to predict MDAs [21].

In recent years, many scholars have achieved promising results in the field of association prediction by combining deep learning and matrix factorization methods. For example, Zeng et al. proposed DMFLDA, which was the first to use the deep matrix factorization method to predict the potential association of lncRNA-disease [22]. Similarly, Lu et al. presented DMFCDA to predict circRNA-disease associations using a similar approach [23]. However, they ignored the similarity information. Furthermore, Liu et al. adopted an unsupervised stack autoencoder method to obtain nonlinear latent features and combined similarity information to predict MDAs [24].

Although great efforts have been made to explore the MDA with good performance, there are still some limitations: (i) the MDAs data is sparse, which may negatively impact the prediction performance; (ii) traditional matrix factorization can only extract linear features of miRNA and disease, however cannot capture the nonlinear relationship of the miRNA-disease heterogeneous network; (iii) the field of deep matrix factorization (such as DMFLDA [22], DMFCDA [23], and SMALF [24]) does not fully reflect the characteristics of the network and node neighbors when constructing features; (iv) CNN cannot extract global information and does not consider the importance of features. In addition, most existing MDAs prediction methods are trained and tested on balanced data, such as [20,25,26]. The known MDAs are taken as positive samples and the unknown MDAs are taken as negative samples. To ensure a balanced dataset, an equal number of samples from unknown set are selected as negative samples, thus maintaining 1:1 ratio between positive and negative samples. It is worth noting that the distribution of these balanced data does not conform to the natural distribution of MDA, and the natural distribution of MDA is extremely unbalanced [27]. For unbalanced data, the number of negative samples is extremely unequal to the number of positive samples. While many methods achieve good performance on these balanced data, they do not perform as well on unbalanced data.

To address these issues, we propose a deep matrix factorization model with a variational autoencoder (DMFVAE) for predicting MDAs. DMFVAE extracts sparse features and dense features through two separate ways. The sparse vectors for each miRNA and disease are obtained by decomposing the original association matrix. The dense vectors are obtained by decomposing the enhanced association matrix, which is enhanced by self-adjusting the nearest neighbor (EASNN) method to mitigate the sparsity of the associated data. In order to solve the problem that traditional matrix factorization can only obtain linear features, variational autoencoders (VAE) is used to obtain complex nonlinear latent vectors for each miRNA and disease. Additionally, node2vec method is employed to obtain the embedding vectors of each miRNA and disease network structure based on miRNA and disease dense vector. To address the drawback of CNN in failing to extract local features, channel attention (CA) is added between convolution and pooling layers. The latent vector features pertaining to each miRNA and disease, along with the embedded features of the network structure, are fused to create combined sample features. These combined features are then utilized for the final prediction. Through these processes, DMFVAE can overcome the limitations of

existing methods and provide more accurate predictions of MDAs. Most models are currently trained and tested on the balanced dataset of HMDD v2.0, but the unbalanced data better reflects the natural distribution of MDAs. Therefore, we conduct extensive experiments not only on the balanced dataset but also on the unbalanced dataset. For unbalanced datasets, we take all unknown MDAs as negative samples. To demonstrate the generalization ability of our model, we also conduct partial experiments on the larger HMDD v3.2 dataset, which includes both balanced and unbalanced data. The experimental results on both HMDD v2.0 (balanced and unbalanced) and HMDD v3.2 (balanced and unbalanced) datasets show that, compared to other baseline methods, DMFVAE outperforms not only on the balanced dataset but also on the unbalanced dataset. Furthermore, we implement case studies on three common diseases, namely lung neoplasms (LN), colon neoplasms (CN), and esophageal neoplasms (EN), and the results demonstrated the superiority of DMFVAE in predicting potential MDAs.

2 Results

2.1 Hyperparameter analysis

In order to determine the optimal choice of model, we analyze some parameters. We conduct experiments to fine-tune the parameters of the feature dimensions, as shown in Fig. 1. In the projection layer, we select the latent vector dimension as the best value of 128 from the options {32, 64,128,256}. In the embedding layer, we conducted a comparative analysis of various dimensions {16, 32, 64, 128} for the embedding features and discovered that setting the dimension to 32 yields optimal performance. In the prediction layer part, there are five parameters, including the number of layers, the number of convolutions, the size of convolution, and pooled kernel, the value of dropout, and the number of the first fully connected layer. After experiments, it can be seen that the optimal selection of these five parameters is 1, 32, 1*2, 0.5, 32. The specific parameter selection changes are shown in Supplementary Fig. S1.

2.2 Prediction performance

To facilitate comparison with other models, we evaluate the performance of DMFVAE on the HMDD v2.0 (495 miRNAs, 383 diseases) balanced and unbalanced datasets using 10 times five-fold cross-validations (5CV). In addition, to verify the generalization ability of the model, we use the same model to experiment on a larger dataset, namely HMDD V3.2 (788 miRNAs, 374 diseases). The specific details on how to use 5CV can be found in the Supplementary instructions. In experiments, several metrics are used to measure the prediction performance of DMFVAE, including accuracy (ACC), precision, recall, F1-score (F1), area under the curve (AUC) and area under the precision-recall curve (AUPR). Correspondingly, the specific formulas can be found in Supplementary Formula 1.

In the experiment, we obtain the feature dimension of the final training sample as 2*160 through the parameter tuning process, where the potential vector feature dimension of each miRNA (disease) is 128 and the feature dimension of the embedding vector is 32. In addition, the average of 10 times 5CV is used as the final results. The results in Table 1 demonstrate that DMFVAE performs well not only on balanced datasets but also on unbalanced datasets in inferring unknown MDAs predictions. In addition, the corresponding ROC curves can be seen in Supplementary Fig. S2.

2.3 Validation of the effect of EASNN

In general, the performance on balanced datasets is usually better than it is on unbalanced datasets. However, the results

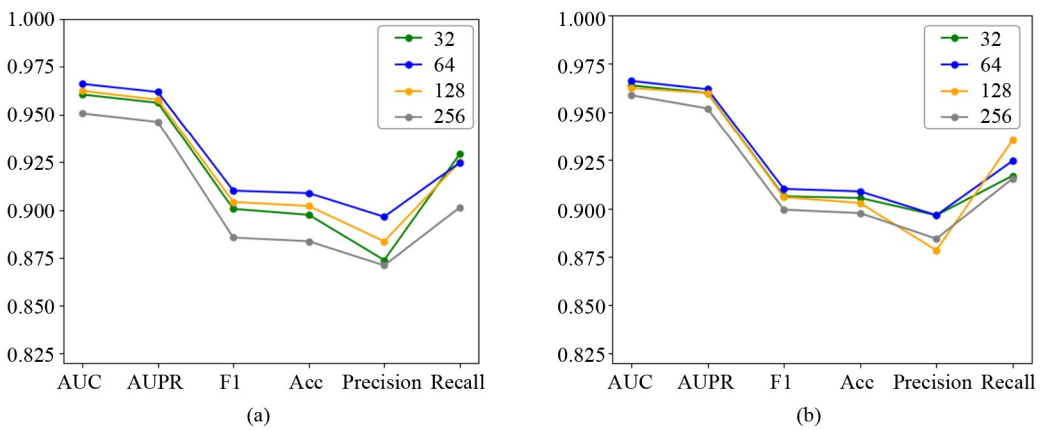


Fig. 1 The effects of different feature dimensions on HMDD v2.0 balanced dataset. (a) The dimensions of latent vector feature; (b) the dimensions of embedding feature

Table 1 The 5CV results on HMDD v2.0 balanced and unbalanced datasets, where the Std and Aver represent standard deviation and average value, respectively

	Fold	AUC	AUPR	F1	ACC	Precision	Recall
Balanced	Std	0.0012	0.0014	0.0031	0.0033	0.0057	0.0035
	Aver	0.9662	0.9619	0.9102	0.9089	0.8966	0.9247
Unbalanced	Std	0.0001	0.0031	0.0031	0.0001	0.0082	0.0086
	Aver	0.9678	0.6556	0.6055	0.9802	0.7053	0.5313

of our experiment have the opposite result. We hypothesize that EASNN can mitigate the sparsity of associated data by replacing zero values with numerical values. This, in turn, enables the model to learn more effectively, ultimately leading to a significant improvement in performance when dealing with unbalanced datasets. To validate the hypothesis, we compare the results of methods without EASNN (-noEASNN) and using EASNN (-withEASNN) module on both balanced and unbalanced datasets. For the balanced-noEASNN and unbalanced-noEASNN models, the EASNN method is removed, and the remaining steps are performed on balanced and unbalanced datasets, respectively. For the balanced-withEASNN and unbalanced-withEASNN models, the EASNN method is utilized.

Table 2 shows that EASNN can effectively improve the performance of the unablanced datasets. According to the experimental results, it can be found that the AUC on the balanced dataset decreases slightly after EASNN is removed, while the AUC on the unbalanced dataset decreases more. If EASNN is removed, the AUC value on the balanced dataset is higher than it on the unbalanced dataset. The above series of results confirm our hypothesis that EASNN can reduce the sparsity of matrices by adaptively selecting neighbors.

2.4 Ablation experiments

In order to validate the significance of integrating nonlinear latent vector features with embedding features of network structure, we perform ablation experiments on the features. Specifically, we set up three sets of experiments. The first set is DMF which only uses nonlinear latent vectors. The second set is MDF-node2vec which only uses the network structure embedding vectors obtained by node2vec. The third set is DMFVAE which adopts sample features, where sample features refer to the results of concating latent vector features and network structure embedding features. Due to the different foci of balanced and unbalanced data evaluation performance indicators, we measure them with different average values (Avg) [27]. For balanced datasets, the average of the performance evaluation metrics is calculated from AUC, AUPR, F1 and Acc. However, in unbalanced datasets, the average performance evaluation metric is calculated from AUPR and F1. Moreover, in order to see the degree of dispersion of the 5CV results, we add error bars to some experiments to show the margin of error for each evaluation metric.

The results are shown in Fig. 2. From the results of the single evaluation metric and Avg, it can be seen that the nonlinear potential vectors, generated by DMF, have more significant impact on predicting potential MDA compared to the embedding vectors generated by DMF-node2vec. In addition, the combined sample features generated by DMFVAE are better than using only one feature.

2.5 Comparison with other methods

To further demonstrate the effectiveness of the DMFVAE method, we compare DMFVAE with ten state-of-the-art methods, including VAEMDA [28], SMALF [24], and ERMDA [29], GRPAMDA [30] for balanced dataset and GBDT-LR [31], MDA-GCNFTG [27], GAEMDA [19], NIMGSA [32] for unbalanced datasets. It must be noted here that since some models do not provide code (or incomplete code) and use similar methods, we choose different comparison models in the corresponding dataset. For example, VAEMDA, SMALF (using deep matrix factorization), ERMDA and GRPAMDA were only experimented on balanced datasets, and GBDT-LR, MDA-GCNFTG,

Table 2 The results of experiments comparing on HMDD v2.0 balanced and unbalanced datasets with EASNN removed and retained, respectively

	AUC	AUPR	F1	ACC	Precision	Recall
Balanced-noEASNN	0.9624	0.9585	0.9045	0.9028	0.8890	0.9210
Balanced-withEASNN	0.9662	0.9619	0.9102	0.9089	0.8966	0.9247
Unbalanced-noEASNN	0.9587	0.5836	0.5032	0.9779	0.7081	0.3930
Unbalanced-withEASNN	0.9678	0.6556	0.6055	0.9802	0.7053	0.5313

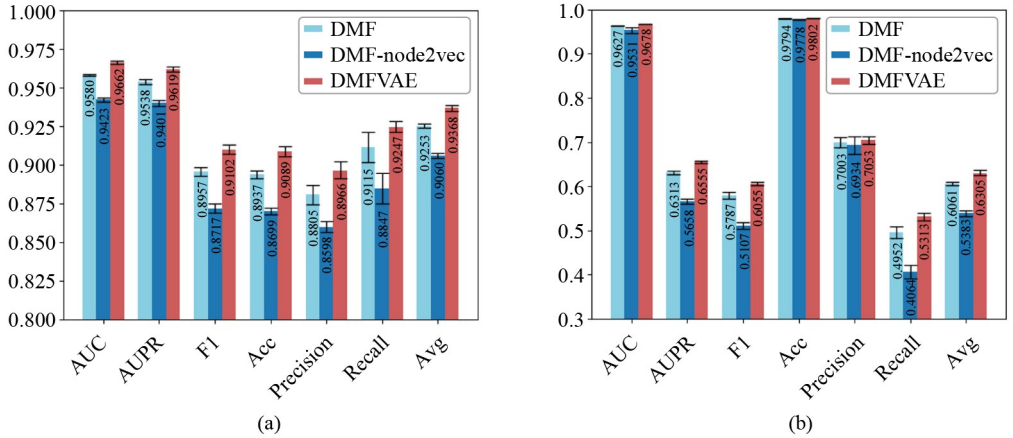


Fig. 2 The comparison by using different features on HMDD v2.0 balanced and unbalanced datasets. For balanced dataset, the average value is calculated from AUC, AUPR, F1 and Acc. For unbalanced dataset, average value is calculated from AUPR and F1. (a) The balanced dataset; (b) the unbalanced dataset

GAEMDA, and NIMGSA were experimented on unbalanced datasets. For a fair comparison, the above models are all evaluated on 5CV and trained using the same training and testing dataset based on HMDD v2.0. Due to the small gap between the results of some models, we add the same Avg as the ablation experiment to measure the experimental results of the balanced and unbalanced datasets.

Figure 3 shows the experimental results comparing DMFVAE with the other five methods on both balanced and unbalanced datasets. As can be seen from the results, DMFVAE has the highest evaluation indicators. For the unbalanced dataset, although DMFVAE has a slightly lower F1 score than GAEMDA and a slightly lower recall than MDA-GCNFTG, the Avg is higher than other models. Therefore, the overall performance of DMFVAE is significantly better than these methods. In addition, the individual values of the evaluation indicators can be found in the Supplementary Tables S1−S2.

2.6 Performance of DMFVAE on HMDD v3.2

To study the generalization ability of DMFVAE, we further apply it to the HMDD v3.2 dataset. Similarly, we use 5CV to evaluate all methods on both balanced and unbalanced datasets. First, we plot the corresponding ROC curves. Second, we compare it with other models. For the balanced dataset, we select several recent studies that use the same data for comparison, such as ABMDA [33], VGAMF [34], and MLRDFM [35]. However, since the previous researchers did not conduct experiments on the HMDD v3.2 unbalanced dataset, we reproduce VGAMF and MLRDFM for comparison, which are only performed on the balanced dataset.

The results of Table 3 show that the AUC of DMFVAE in balanced datasets and unbalanced datasets reach 0.9682 and 0.9705, respectively, achieving better results on larger datasets. The corresponding ROC curves can be seen in the Supplementary Fig. S3. The results of Fig. 4 show that, for the balanced dataset, our model outperforms these models based on various evaluation indicators. For the unbalanced dataset, despite the precision value is lower than VGAMF, other metrics and Avg are higher than other models. Overall, our model is superior to other models. The specific experimental results for balanced and unbalanced can be seen in the Supplementary Tables S3−S4.

2.7 Case studies

To investigate the performance of DMFVAE in inferring unknown miRNA-disease interactions in practical applications, we select three common diseases LN, CN and EN for case studies by training a DMFVAE model to identify unknown associations based on all known MDA data. Then, we validate the predicted associations by referring to HMDD v3.2, dbDEMC [36] and miR2Disease [37] databases.

Studies have shown that early-stage LN have a high cure rate and can be screened for early-stage LN through regular physical examinations. Many studies have shown that some miRNAs can be used as biomarkers for LN [38]. Table 4 gives the top 20 LN-related miRNAs that we identified. We can all find evidence to verify them. For instance, the 6th miRNA, ‘mir-196b’ is a novel distinguishing marker of adenocarcinoma (AC) and squamous cell carcinoma (SCC) [39]. It can improve the treatment for different categories of

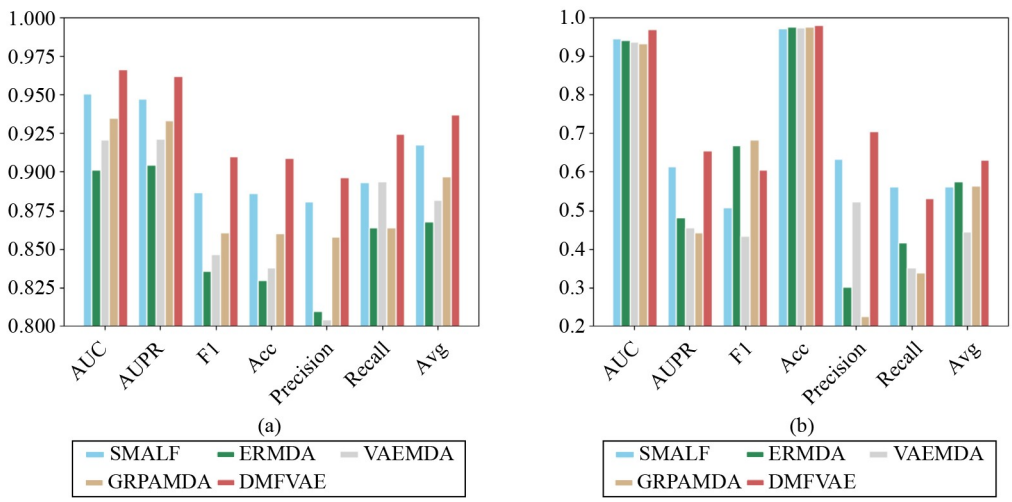


Fig. 3 DMFVAE compared with other models on HMDD v2.0 balanced and unbalanced datasets. For balanced dataset, the average value is calculated from AUC, AUPR, F1 and Acc. For unbalanced dataset, the average value is calculated from AUPR and F1. (a) The balanced dataset; (b) the unbalanced dataset

Table 3 The 5CV results on HMDD v3.2 balanced and unbalanced datasets, where the Std and Aver represent standard deviation and average value, respectively

	Fold	AUC	AUPR	F1	ACC	Precision	Recall
Balanced	Std	0.0004	0.0010	0.0019	0.0022	0.0047	0.0041
	Aver	0.9682	0.9639	0.9140	0.9123	0.8967	0.9322
Unbalanced	Std	0.0005	0.0033	0.0038	0.0002	0.0081	0.0081
	Aver	0.9705	0.6853	0.6191	0.9795	0.7142	0.5514

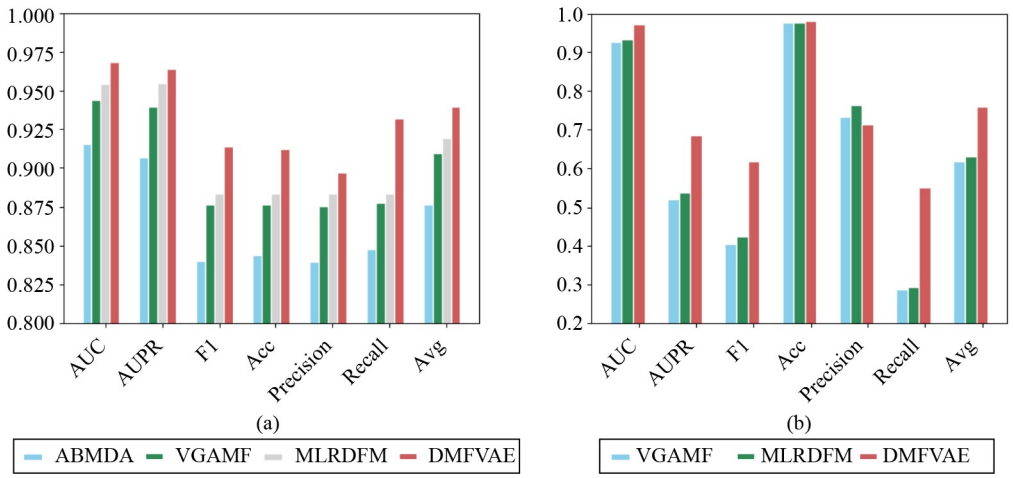


Fig. 4 DMFVAE compared with other models on HMDD v3.2 balanced and unbalanced datasets. For balanced dataset, the average value is calculated from AUC, AUPR, F1 and Acc. For unbalanced dataset, the average value is calculated from AUPR and F1. (a) The balanced dataset; (b) the unbalanced dataset

Table 4 Top 20 candidate miRNAs associated with LN, where H3, DEMC and miR represent HMDD v3.2, dbDEMC and miR2Disease, respectively

Rank	miRNA	Evidence	Rank	miRNA	Evidence
1	hsa-mir-211	H3, DEMC, miR	11	hsa-mir-20b	DEMC
2	hsa-mir-130a	H3, DEMC, miR	12	hsa-mir-152	H3, db
3	hsa-mir-129	H3, DEMC	13	hsa-mir-99a	H3, DEMC, miR
4	hsa-mir-151a	DEMC	14	hsa-mir-23b	DEMC
5	hsa-mir-208a	H3	15	hsa-mir-449a	H3
6	hsa-mir-196b	H3, DEMC	16	hsa-mir-16	H3, DEMC, miR
7	hsa-mir-378a	DEMC	17	hsa-mir-106b	H3, DEMC
8	hsa-mir-302c	DEMC	18	hsa-mir-10a	H3, DEMC
9	hsa-mir-370	DEMC	19	hsa-mir-195	H3, DEMC, miR
10	hsa-mir-296	DEMC	20	hsa-mir-15a	H3, DEMC

non-small-cell lung cancer because overall survival with maintenance pemetrexed for patients with non-squamous histology is significant [40].

CN is a major cause of mortality and morbidity worldwide even though the incidence rate is higher in people aged 40-50 years [41], which is asymptomatic in its early stages, making it easy to miss a diagnosis. Therefore, there is a growing need for novel sensitive biomarkers that can help improve CN detection [42]. As shown in the Supplementary Table S5, all the top 20 candidate miRNAs for CN are validated in HMDD v3.2, dbDEMC, or miR2Disease. For instance, miR-21 is one of the most frequently upregulated miRNAs in various human tumors, and is highly up-regulated miRNA in colorectal tumors [43]. Through this marked upregulation, circulating miR-21 has been shown to be an effective biomarker for early detection of colorectal carcinoma (CRC), along with a prognostic marker for aggressive disease, as it is associated with poor patient survival [44].

Besides, in recent years, the incidence of EN in Asia has gradually increased [45]. Identification of EN biomarkers for early diagnosis has important implications for the diagnosis and treatment prospects of EN [46]. The results of Supplementary Table S6 indicated that nineteen of all the top twenty candidate miRNAs for EN are validated in HMDD v3.2, dbDEMC or miR2Disease. Only one miRNA, hsa-mir-122, has not been verified yet in these databases. However, miR-122 has been reported to have suppressive roles by regulating cyclin G1 negatively, reducing proliferation and inducing apoptosis in vitro [47]. In other words, all the predicted novel MDAs for EN are confirmed.

3 Methods

3.1 Human miRNA-disease associations

In this study, we mainly use HMDD v2.0 [48] as the benchmark dataset and the known MDAs matrices are downloaded directly from the HMDD database. HMDD v2.0 includes 5430 confirmed associations among 495 miRNAs and 383 diseases. To further validate the generalization ability of DMFVAE, we also apply our proposed method to the HMDD v3.2 dataset [49]. Influenced by previous researches [34,35], we use the same data HMDD v3.2, which has 8968 confirmed associations between 788 miRNAs and 374 diseases. The datasets are shown in Table 5.

3.2 Disease semantic similarity

According to a previous study [50], disease semantic similarity can be calculated from the medical subject headings descriptors, which are available at the website of https://www.ncbi.nlm.nih.gov. Here, we use a directed acyclic graph (DAG) to represent the relationships among different diseases. For a disease p, we define $DAG_p = \left(T_p, E_p\right)$, where T_p represents p itself and its ancestor nodes, and E_p represents the corresponding edge set. In addition, in order to obtain the semantic similarity $SD\left(d_i, d_j\right)$ of diseases d_i and d_j, two

Table 5 Corresponding miRNA-disease association information summarized in the test data, where H2, H3, #P and #N represent HMDD v2.0, HMDD v3.2, the number of positive samples and the number of negative samples, respectively

data	miRNAs	diseases	#P	#N
H2 balanced	495	383	5430	5430
H2 unbalanced	495	383	5430	184155
H3 balanced	788	374	8968	8968
H3 unbalanced	788	374	8968	285744

different ways are employed, resulting in $SD1$ and $SD2$ respectively, and then the average is taken to obtain the final semantic similarity value. Specifically, $SD1$ is calculated based on the DAG of the disease. And the larger the DAG shared by two diseases, the higher the semantic similarity between the two diseases [51]. $SD2$ is calculated based on the number of DAGs in which the disease is located, and if the disease is present in fewer DAGs, it indicates that the disease is more specific, leading to a higher semantic contribution value [52]. The specific calculation formula can be found in Supplementary Formula 2.

3.3 DMFVAE

In this study, we propose a model named DMFVAE to predict potential MDAs. DMFVAE can be described in four steps, as shown in Fig. 5: (1) feature extraction: obtain sparse and dense vector of miRNA and disease using matrix decomposition, respectively, (2) projection layer: apply a variational autoencoder to generate the latent vectors representation of miRNA and disease, (3) embedding layer: adopt node2vec to obtain the network structure embeddings vector of miRNA and disease nodes, (4) prediction layer: conduct convolutional neural networks with channel attention to perform the final prediction. By integrating these different steps, DMFVAE can effectively address the limitations of existing methods and achieve accurate predictions of potential MDAs.

3.3.1 Feature extraction

There are two steps in this part: Firstly, the original association matrix is decomposed to obtain its row (or column) vector as the disease (or miRNA) sparse vector. Secondly, EASNN method is used to obtain the enhanced correlation matrix, then the dense vector of each miRNA (or

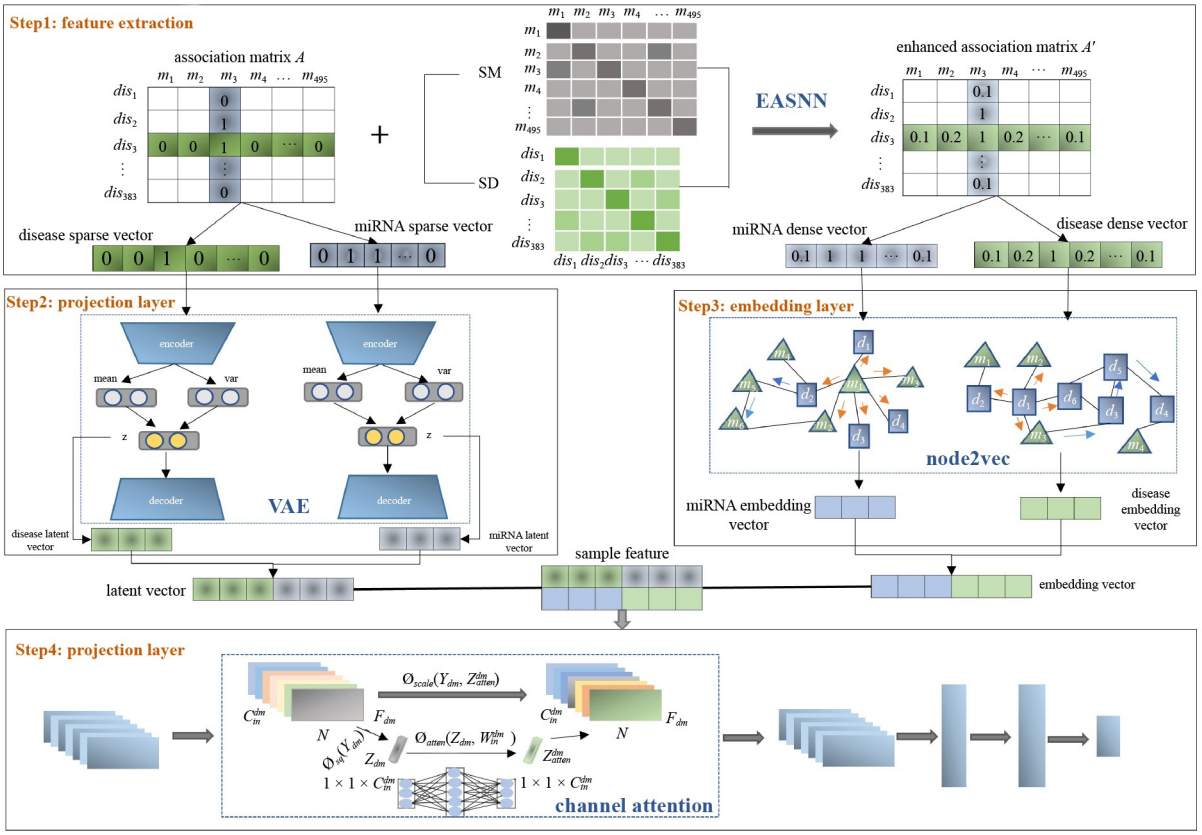


Fig. 5 Overview of the DMFVAE method architecture. (1) Feature extraction. The sparse and dense vector are constructed for each miRNA and disease by matrix factorization of the original association matrix A and the enhanced association matrix A' with the EASNN method. (2) Projection layer. A nonlinear latent vector of each miRNA and disease is obtained by using a variational autoencoder. (3) Embedding layer. The embedded features of miRNAs and disease network structures are acquired by extracting dense vectors using node2vec. (4) Prediction layer. Convolution is performed on the features of the samples (where the features represent the concatenated features of the nonlinear latent vectors and network structure embeddings for each miRNA and disease), and channel attention is used to assign different weights to each channel for final prediction

disease) is obtained.

Universally acknowledged, the MDAs matrix is very sparse. For example, there are only 5430 know associations in matrix $A \in R^{383\times495}$ and know associations only account for about 2.86% of the total number of 189585 data samples on HMDD v2.0. However, the lack of known information hinders the effective enhancement of model prediction performance. Therefore, we adopt an effective way to solve the problem called EASNN [53]. EASNN processes the similarity matrix *SM* (or *SD*) to identify the value that correspond to the most similar position of each miRNA (or disease). These values are then stored, while the remaining position are assigned to zeroes to obtain a new miRNA matrix M (or disease matrix D). Then the association matrix A is used to deduce a new interaction profile, A_{md}. The enhanced association matrix A' is obtained by taking the maximum value between A_{md} and A. The specific formula can be found in the Supplementary Formula 3.

3.3.2 Projection layer

In this section, we employ a deep matrix factorization approach, which uses a VAE to generate a potential vector representation of miRNA (or disease). Unlike traditional matrix factorization, it sends each row of vectors to a deep learning model for training to obtain the corresponding potential vectors. Therefore, it effectively overcomes the limitations of traditional matrix factorization methods which cannot capture complex nonlinear relationships.

In our deep matrix factorization model, the rows (or columns) of the association matrix are considered as sparse vectors for diseases (or miRNAs), and then the two sparse vectors are fed into the VAE to obtain the nonlinear relationship among miRNAs and diseases. The encoder of VAE produces approximate posterior probabilities $p_\phi(z|x)$ rather than a specific latent vector, whereas the decoder takes samples from this distribution and generates reconstruction probabilities $p_\theta(x|z)$, where ϕ,θ represent the parameters of the encoder and decoder, respectively. We suppose X is the known miRNA-disease samples, X' represents the output of the VAE. The marginal likelihood is expressed as follows:

$$L(X,X') = \sum_{i=1}^{N} \log p_\theta(x_i), \tag{1}$$

where N is the number of known $MDAs$ and the marginal log-likelihood of each sample can be described as below:

$$\log p_\theta(x_i) = D_{KL}(q_\theta(z \mid x_i) \| p_\theta(z \mid x_i)) + L(\theta,\phi;x_i), \tag{2}$$

where the first term of the equation is the KL divergence of the approximate $q_\phi(z)$ from the true posterior. In order to randomly generate a sample, we introduce a reparameterization trick. By using a deterministic-transformation $\varepsilon \in [0,1]$, mean μ and standard deviation σ', we obtain the reparameterization z of the VAE as follows:

$$z = \mu + \sigma' \times \varepsilon, \tag{3}$$

The lower bound of marginal log-likelihood can be represented as follows:

$$\begin{aligned} L(\phi,\theta;x_i) &\simeq -D_{KL}\left(q_\phi(z \mid x_i) \| p_\theta(z \mid x_i)\right) + \log p_\theta(x_i) \\ &= E_{q_\phi(z|x_i)}\left(-\log q_\phi(z \mid x_i) + \log p_\theta(x_i,z)\right) \\ &= -D_{KL}\left(q_\phi(z \mid x_i) \| p_\theta(z)\right) + E_{q_\phi(z|x_i)} \log p_\theta(x_i \mid z) \\ &= -D_{KL}\left(q_\phi(z \mid x_i) \| p_\theta(z)\right) + \frac{1}{S}\sum_{j=1}^{S} \log p_\theta\left(x_i \mid z^j\right), \end{aligned} \tag{4}$$

where $p_\theta(z)$ is the prior distribution and S is the number of latent vector samples z. The VAE updates the parameters of each node of the network iteratively to minimize the loss so that it obtains the low-dimensional and high-density latent vector features D_i and M_j^T.

3.3.3 Embedding layer

Although many scholars have applied deep matrix factorization methods to predictive association, few people combine network nodes and neighborhood information when constructing features in this field. Inspired by previous work, we initially use node2vec [54] to process the dense vector to obtain the embedding vector feature of each miRNA or disease network structure. Node2vec improves the generation mode of random walks, allowing the generated random walks to reflect the two sampling characteristics of depth-first sampling and breadth-first sampling strategies, thus improving the effect of network embedding. The embedding features of miRNAs and diseases network structure is obtained from the miRNA-disease association network and its transpose network, which are denoted as $M^a \in R^{nm\times e}$ and $D^a \in R^{nd\times e}$ respectively, where nm, nd are the number of miRNAs and disease respectively, and e is the dimension of the embedding feature.

3.3.4 Prediction layer

With the wide application of CNN [55] in images, more and more researchers have applied CNN to the field of biological information. However, CNN has limitations in extracting global information. This is because CNN uses local features to obtain global features without considering the importance of features. To address this problem, we add CA [56] between the convolutional and pooling layers to effectively extract global information. In this part, after obtaining the features of all samples, CNN is combined with the CA to achieve the final prediction. It is worth noting that the sample features referred here are the features obtained after integrating the nonlinear latent vectors and network structure embedding vectors of miRNAs and diseases.

In this section, we use attention on the channel to assign weights to features. To obtain the significance of different channels, the channel-based statistics are derived based on global and average pooling. For C_{in}^{dm} is the number of channels of features, $Z \in R^{1\times1\times C_{in}^{dm}}$ is generated by squeezing new features $X_{dm} = \left[x_1, x_2, \ldots, x_{in}^{dm}\right]$, where $X_{dm} \in R^{F_{dm}\times M\times c_{in}^{dm}}$, M represents the number of sample features. Specifically, for the c th feature matrix of new feature x_c, z_c is calculated as follows:

$$z_c = \phi_{sq}(x_c) = \frac{1}{F_{dm} \times M} \sum_{i=1}^{F_{dm}} \sum_{j=1}^{M} x_c(x, j). \quad (5)$$

To accurately capture the significance of each channel, the attention mechanism is employed to compute the attention weights of channels as follows:

$$Z_{att} = \phi_{atten}\left(Z_{dm}, W_{in}^{dm}\right) = \delta(W_2 \zeta(W_1 Z)), \quad (6)$$

where $\delta(\cdot)$ is Sigmoid activation, $\zeta(\cdot)$ is Relu activation, $W_{\text{in}}^{dm} = \{W_1, W_2\}$ is the training parameter. Finally, channel attention can be defined as $Z_{att} = \left[z_1^{att}, z_2^{att}, \ldots, z_{C_{in}^{dm}}^{att}\right]$. After obtaining the attention weights for different channels, we incorporate these attention with channel features for normalization as follows:

$$\tilde{x}_c = \phi_{scale}\left(X_c, Z_c^{att}\right) = Z_c^{\text{att}} \cdot x_c. \quad (7)$$

After the above steps, the normalized feature channel information $\tilde{x}_{dm} = \left[\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{C_{in}^{dm}}\right]$ are obtained and then the remaining pooling and fully connected layers are used for final miRNA-disease association prediction.

4 Discussion

In this study, we propose an effective and new method called DMFVAE for predicting disease-related miRNAs. The effectiveness of our model can be attributed to some strategies we employed which are focused on the characteristics of the data. Firstly, the disease-miRNA association data we employed exhibits a high level of sparsity which is characterized by numerous zero values. In some cases, entire rows or columns may consist of only zero values, indicating no association. To address this issue, we use the EASNN method to enhance the sparse association matrix, which can transform some zero values in the association matrix into values between 0 and 1 based on the similarity values of miRNA and diseases. It can effectively alleviate the sparsity issue of association data. Consequently, even for some new diseases or miRNAs that lack existing associations, EASNN can provide corresponding values to effectively enhance them.

Furthermore, we construct four datasets based on HMDD v2.0 and HMDD v3.2 databases, where the association information is represented by a 0−1 matrix. However, the miRNA-disease relationship is inherently complex and nonlinear, making it challenging to represent by simple matrix factorization. In this study, we employ the VAE method to solve this problem. It outputs the mean and standard deviation of the latent vector from the encoder and generates data by random sampling from the latent space. It not only mitigates the issue of garbage output that may be caused by random latent vectors, but also enhances the model's generalization ability. The nonlinear latent vectors are obtained by re-sampling, which allows the model to make improved predictions, even for specific diseases or miRNAs with limited associations in the original datasets.

Last but not least, in the association matrix we used, the miRNAs associated with different diseases are different, and different miRNAs have different contributions to a particular disease. When employing the traditional DeepWalk random walk method, all neighbors will be assigned equal contribution weights, which is not desirable in this case. Therefore, this study adopts the Node2vec method, which uses second-order random walks and controls the probability of visiting different vertices based on different parameters p and q to control whether it is biased towards depth-first or breadth-first search. For a specific disease, if it exhibits a greater number of associated miRNAs, the data is denser, and it may be biased towards a depth-first search. Conversely, if there are fewer miRNA associated with the disease and the data is sparse, breadth-first search is preferred to obtain more feature information. The same principle applies to miRNAs as well.

5 Conclusion

Abnormal expression of miRNA has been widely observed in the development of many complex human diseases. Identifying potential MDAs can help us better understand the pathogenesis of major human diseases and improve their prevention, diagnosis, and treatment. In this study, we propose a novel approach, DMFVAE, to predict MDAs. To effectively reduce the sparsity of the associated data, the EASNN method is used to enhance the association matrix. The VAE method is applied to obtain the nonlinear latent vectors of miRNA and disease. And the network structure embedding vector of miRNA and disease are obtained by node2vec. Furthermore, CA is employed to extract global features to address the limitation of CNNs which are primarily capable of capturing local features. The results of the 5CV show that our model not only achieved good results on balanced datasets, but also outperform other methods on unbalanced datasets. Additionally, case studies on three cancers further confirmed the ability of our model to identify potential candidates disease-associated miRNAs.

Although the proposed method, DMFVAE, achieves promising results for predicting MDAs, it has some limitations that require further investigation. Since DMFVAE only uses miRNA functional similarity and disease semantic similarity, we plan to incorporate more biological information, such as miRNA sequence or target information. Besides, the similarity information is not fully characterized, so we will extract features from the multi-view graph representation learning method.

Acknowledgments This work was supported by the National Natural Science Foundation of China (Grant Nos. 62202004, and 62322301), the Natural Science Foundation of Anhui Province (No. 2108085QF267), the University Synergy Innovation Program of Anhui Province (No. GXXT-2021-039), and the Anhui University Outstanding Youth Research Project (No. 2022AH020010).

Supplementary document Our supplementary document consists of three parts. The first part is the Supplementary Figures related to the experiment, the second part is the Supplementary Tables related to the experiment, and the third part is the Supplementary Formulas related to the experimental results and some methods. The supporting information is available online at journal.hep.com.cn and link.springer.com.

Competing interests The authors declare that they have no competing interests or financial conflicts to disclose.

References

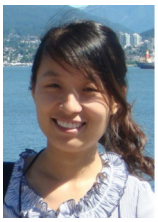
1. Gebert L F R, MacRae I J. Regulation of microRNA function in animals. Nature Reviews Molecular Cell Biology, 2019, 20(1): 21–37
2. Van Meter E N, Onyango J A, Teske K A. A review of currently identified small molecule modulators of microRNA function. European Journal of Medicinal Chemistry, 2020, 188: 112008
3. Hammond S M. An overview of microRNAs. Advanced Drug Delivery Reviews, 2015, 87: 3–14
4. Patanè S. The complex miRNAs-p53 signaling network in cardiovascular disease. Journal of the American College of Cardiology, 2017, 69(16): 2099–2100
5. Wang X, He Y, Mackowiak B, Gao B. MicroRNAs as regulators, biomarkers and therapeutic targets in liver diseases. Gut, 2021, 70(4): 784–795
6. Niu Y W, Wang G H, Yan G Y, Chen X. Integrating random walk and binary regression to identify novel miRNA-disease association. BMC Bioinformatics, 2019, 20(1): 59
7. Chen X, Xie D, Zhao Q, You Z H. MicroRNAs and complex diseases: from experimental results to computational models. Briefings in Bioinformatics, 2019, 20(2): 515–539
8. Huang L, Zhang L, Chen X. Updated review of advances in microRNAs and complex diseases: taxonomy, trends and challenges of computational models. Briefings in Bioinformatics, 2022, 23(5): bbac358
9. Huang L, Zhang L, Chen X. Updated review of advances in microRNAs and complex diseases: experimental results, databases, webservers and data fusion. Briefings in Bioinformatics, 2022, 23(6): bbac397
10. Huang L, Zhang L, Chen X. Updated review of advances in microRNAs and complex diseases: towards systematic evaluation of computational models. Briefings in Bioinformatics, 2022, 23(6): bbac407
11. Dai L Y, Liu J X, Zhu R, Wang J, Yuan S S. Logistic weighted profile-based bi-random walk for exploring miRNA-disease associations. Journal of Computer Science and Technology, 2021, 36(2): 276–287
12. Xuan P, Wang D, Cui H, Zhang T, Nakaguchi T. Integration of pairwise neighbor topologies and miRNA family and cluster attributes for miRNA−disease association prediction. Briefings in Bioinformatics, 2022, 23(1): bbab428
13. Xu J, Zhu W, Cai L, Liao B, Meng Y, Xiang J, Yuan D, Tian G, Yang J. LRMCMDA: predicting miRNA-disease association by integrating low-rank matrix completion with miRNA and disease similarity information. IEEE Access, 2020, 8: 80728–80738
14. Chen X, Yin J, Qu J, Huang L. MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. PLoS Computational Biology, 2018, 14(8): e1006418
15. Chen X, Sun L G, Zhao Y. NCMCMDA: miRNA–disease association prediction through neighborhood constraint matrix completion. Briefings in Bioinformatics, 2021, 22(1): 485–496
16. Lu X, Li J, Zhu Z, Yuan Y, Chen G, He K. Predicting miRNA-disease associations via combining probability matrix feature decomposition with neighbor learning. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022, 19(6): 3160–3170
17. Zhang Y, Lei X, Fang Z, Pan Y. CircRNA-disease associations prediction based on metapath2vec++ and matrix factorization. Big Data Mining and Analytics, 2020, 3(4): 280–291
18. Xuan P, Sun H, Wang X, Zhang T, Pan S. Inferring the disease-associated miRNAs based on network representation learning and convolutional neural networks. International Journal of Molecular Sciences, 2019, 20(15): 3648
19. Li Z, Li J, Nie R, You Z H, Bao W. A graph auto-encoder model for miRNA-disease associations prediction. Briefings in Bioinformatics, 2021, 22(4): bbaa240
20. Ding Y, Tian L P, Lei X, Liao B, Wu F X. Variational graph auto-encoders for miRNA-disease association prediction. Methods, 2021, 192: 25–34
21. Liu W, Lin H, Huang L, Peng L, Tang T, Zhao Q, Yang L. Identification of miRNA–disease associations via deep forest ensemble learning based on autoencoder. Briefings in Bioinformatics, 2022, 23(3): bbac104
22. Zeng M, Lu C, Fei Z, Wu F X, Li Y, Wang J, Li M. DMFLDA: a deep learning framework for predicting lncRNA−disease associations. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 18(6): 2353–2363
23. Lu C, Zeng M, Zhang F, Wu F X, Li M, Wang J. Deep matrix factorization improves prediction of human circRNA-disease associations. IEEE Journal of Biomedical and Health Informatics, 2021, 25(3): 891–899
24. Liu D, Huang Y, Nie W, Zhang J, Deng L. SMALF: miRNA-disease associations prediction based on stacked autoencoder and XGBoost. BMC Bioinformatics, 2021, 22(1): 219
25. Li J, Chen X, Huang Q, Wang Y, Xie Y, Dai Z, Zou X, Li Z. Seq-SymRF: a random forest model predicts potential miRNA-disease associations based on information of sequences and clinical symptoms. Scientific Reports, 2020, 10(1): 17901
26. Li J, Li Z, Nie R, You Z, Bao W. FCGCNMDA: predicting miRNA-disease associations by applying fully connected graph convolutional networks. Molecular Genetics and Genomics, 2020, 295(5): 1197–1209
27. Chu Y, Wang X, Dai Q, Wang Y, Wang Q, Peng S, Wei X, Qiu J, Salahub D R, Xiong Y, Wei D Q. MDA-GCNFTG: identifying miRNA-disease associations based on graph convolutional networks via graph sampling through the feature and topology graph. Briefings in Bioinformatics, 2021, 22(6): bbab165
28. Zhang L, Chen X, Yin J. Prediction of potential miRNA−disease associations through a novel unsupervised deep learning framework with variational autoencoder. Cells, 2019, 8(9): 1040
29. Dai Q, Wang Z, Liu Z, Duan X, Song J, Guo M. Predicting miRNA-disease associations using an ensemble learning framework with resampling method. Briefings in Bioinformatics, 2022, 23(1): bbab543
30. Zhong T, Li Z, You Z H, Nie R, Zhao H. Predicting miRNA−disease associations based on graph random propagation network and attention network. Briefings in Bioinformatics, 2022, 23(2): bbab589
31. Zhou S, Wang S, Wu Q, Azim R, Li W. Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression. Computational Biology and Chemistry, 2020, 85: 107200
32. Jin C, Shi Z, Lin K, Zhang H. Predicting miRNA-disease association based on neural inductive matrix completion with graph autoencoders and self-attention mechanism. Biomolecules, 2022, 12(1): 64

33. Zhao Y, Chen X, Yin J. Adaptive boosting-based computational model for predicting potential miRNA-disease associations. Bioinformatics, 2019, 35(22): 4730–4738
34. Ding Y, Lei X, Liao B, Wu F X. Predicting miRNA-disease associations based on multi-view variational graph auto-encoder with matrix factorization. IEEE Journal of Biomedical and Health Informatics, 2022, 26(1): 446–457
35. Ding Y, Lei X, Liao B, Wu F X. MLRDFM: a multi-view Laplacian regularized DeepFM model for predicting miRNA-disease associations. Briefings in Bioinformatics, 2022, 23(3): bbac079
36. Yang Z, Ren F, Liu C, He S, Sun G, Gao Q, Yao L, Zhang Y, Miao R, Cao Y, Zhao Y, Zhong Y, Zhao H. dbDEMC: a database of differentially expressed miRNAs in human cancers. BMC Genomics, 2010, 11(S4): S5
37. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Research, 2009, 37: D98–D104
38. Seijo L M, Zulueta J J. Understanding the links between lung cancer, COPD, and emphysema: a key to more effective treatment and screening. Oncology, 2017, 31(2): 93–102
39. Hamamoto J, Soejima K, Yoda S, Naoki K, Nakayama S, Satomi R, Terai H, Ikemura S, Sato T, Yasuda H, Hayashi Y, Sakamoto M, Takebayashi T, Betsuyaku T. Identification of microRNAs differentially expressed between lung squamous cell carcinoma and lung adenocarcinoma. Molecular Medicine Reports, 2013, 8(2): 456–462
40. Ciuleanu T, Brodowicz T, Zielinski C, Kim J H, Krzakowski M, Laack E, Wu Y L, Bover I, Begbie S, Tzekova V, Cucevic B, Pereira J R, Yang S H, Madhavan J, Sugarman K P, Peterson P, John W J, Krejcy K, Belani C P. Maintenance pemetrexed plus best supportive care versus placebo plus best supportive care for non-small-cell lung cancer: a randomised, double-blind, phase 3 study. The Lancet, 2009, 374(9699): 1432–1440
41. Schabath M B, Cote M L. Cancer progress and priorities: lung cancer. Cancer Epidemiology, Biomarkers & Prevention, 2019, 28(10): 1563–1579
42. Cappell M S. Pathophysiology, clinical presentation, and management of colon cancer. Gastroenterology Clinics of North America, 2008, 37(1): 1–24
43. Aslam M I, Taylor K, Pringle J H, Jameson J S. MicroRNAs are novel biomarkers of colorectal cancer. British Journal of Surgery, 2009, 96(7): 702–710
44. Yamada A, Horimatsu T, Okugawa Y, Nishida N, Honjo H, Ida H, Kou T, Kusaka T, Sasaki Y, Yagi M, Higurashi T, Yukawa N, Amanuma Y, Kikuchi O, Muto M, Ueno Y, Nakajima A, Chiba T, Boland C R, Goel A. Serum miR-21, miR-29a, and miR-125b are promising biomarkers for the early detection of colorectal neoplasia. Clinical Cancer Research, 2015, 21(18): 4234–4242
45. El-Serag H B, Sweet S, Winchester C C, Dent J. Update on the epidemiology of gastro-oesophageal reflux disease: a systematic review. Gut, 2014, 63(6): 871–880
46. Sohda M, Kuwano H. Current status and future prospects for esophageal cancer treatment. Annals of Thoracic and Cardiovascular Surgery, 2017, 23(1): 1–11
47. Gramantieri L, Ferracin M, Fornari F, Veronese A, Sabbioni S, Liu C G, Calin G A, Giovannini C, Ferrazzi E, Grazi G L, Croce C M, Bolondi L, Negrini M. Cyclin G1 is a target of miR-122a, a microRNA frequently down-regulated in human hepatocellular carcinoma. Cancer Research, 2007, 67(13): 6092–6099
48. Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, Cui Q. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Research, 2014, 42(D1): D1070–D1074
49. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q. HMDD v3.0: a database for experimentally supported human microRNA−disease associations. Nucleic Acids Research, 2019, 47(D1): D1013–D1017
50. Xuan P, Han K, Guo M, Guo Y, Li J, Ding J, Liu Y, Dai Q, Li J, Teng Z, Huang Y. Prediction of microRNAs associated with human diseases based on weighted *k* most similar neighbors. PLoS One, 2013, 8(8): e70204
51. Wang D, Wang J, Lu M, Song F, Cui Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics, 2010, 26(13): 1644–1650
52. Pasquier C, Gardès J. Prediction of miRNA-disease associations with a vector space model. Scientific Reports, 2016, 6: 27036
53. Zhang Z W, Gao Z, Zheng C H, Wang Y T, Qi S M. MELPMDA: a new method based on matrix enhancement and label propagation for predicting miRNA-disease association. In: Proceedings of the 17th International Conference on Intelligent Computing Theories and Application. 2021, 536−548
54. Xie F, Yang Z, Song J, Dai Q, Duan X. DHNLDA: a novel deep hierarchical network based method for predicting lncRNA-disease associations. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022, 19(6): 3395–3403
55. Dhillon A, Verma G K. Convolutional neural network: a review of models, methodologies and applications to object detection. Progress in Artificial Intelligence, 2020, 9(2): 85–112
56. Tang X, Luo J, Shen C, Lai Z. Multi-view multichannel attention graph convolutional network for miRNA−disease association prediction. Briefings in Bioinformatics, 2021, 22(6): bbab174

Pijing Wei received the PhD degree in computer science and technology from Anhui University, China in 2020. She is currently a lecturer in the Institute of Physical Science and Information Technology, Anhui University, China. Her main research interests include bioinformatics, synthetic biology, cancer data mining, and machine learning.

Qianqian Wang received the BS degree in science from Anhui University of Science and Technology, China in 2021. She is currently pursuing the MS degree in the School of Computer Science and Technology, Anhui University, China. Her research interests include research of bioinformatics and deep learning.

Zhen Gao received the MS degree in computer science from Qufu Normal University, China in 2021. She is currently working toward the PhD degree in the School of Computer Science and Technology, Anhui University, China. Her research interests include research of bioinformatics, deep learning and gene regulatory networks.



Ruifen Cao received the PhD degree from Hefei Institute of Physical Sciences, Chinese Academy of Sciences, China in 2009. She is currently an associate professor at the School of Computer Science and Technology, Anhui University, China. Her research interests include artificial intelligence, medical image processing, and multimodal data fusion.

Chunhou Zheng received the the PhD degree in pattern recognition and intelligent system in 2006, from University of Science and Technology of China. From February 2007 to June 2009, he worked as a Postdoctoral Fellow in the Hefei Institutes of Physical Sceience, Chinese Academy of Sciences, China. From July 2009 to July 2010, he worked as a Postdoctoral Fellow in the Department of Computing, the Hong Kong Polytechnic University, China. He is currently a Professor in the School of Artificial Intelligence, Anhui University, China. His research interests include pattern recognition, synthetic biology and bioinformatics.