Predicting miRNA-Disease Associations From miRNA-Gene-Disease Heterogeneous Network With Multi-Relational Graph Convolutional Network Model

Wei Peng, Zicheng Che, Wei Dai, Shoulin Wei, and Wei Lan

Abstract—MiRNAs are reported to be linked to the pathogenesis of human complex diseases. Disease-related miRNAs may serve as novel bio-marks and drug targets. This work focuses on designing a multi-relational Graph Convolutional Network model to predict miRNA-disease associations (HGCNMDA) from a Heterogeneous network. HGCNMDA introduces a gene layer to construct a miRNA-gene-disease heterogeneous network. We refine the features of nodes into initial and inductive features so that the direct and indirect associations between diseases and miRNA can be considered simultaneously. Then HGCNMDA learns feature embeddings for miRNAs and disease through a multi-relational graph convolutional network model that can assign appropriate weights to different types of edges in the heterogeneous network. Finally, the miRNA-disease associations were decoded by the inner product between miRNA and disease feature embeddings. We apply our model to predict human miRNA-disease associations. The HGCNMDA is superior to the other state-of-the-art models in identifying missing miRNA-disease associations and also performs well on recommending related miRNAs/diseases to new diseases/ miRNAs. The codes are available at https://github.com/weiba/HGCNMDA.

Index Terms—Disease, heterogeneous network embedding, MiRNA, MiRNA-disease association prediction, multi-relational graph convolutional network

1 Introduction

MiRNAs(MicroRNA) are a class of small single-stranded non-coding RNA molecules about 21 nucleotides in length [1], [2]. They control the expression of genes and are involved in many critical biological processes, such as cell growth, cell differentiation, immune reactions. MiRNAs are also linked to the pathologies of human complex diseases, ranging from cancer to common diseases, such as cardiovascular diseases, Parkinson's disease, immune-related disease [1]. Many studies are interested in finding disease-related miRNAs because they may serve as novel bio-marks and drug targets. High throughput sequencing techniques have uncovered a massive number of miRNAs. However, very few parts of them are determined to be closely associated with diseases.

Recently, many computational methods have been designed to detect the miRNA-disease associations. These methods usually infer novel associations from similar miRNAs or similar diseases. Hence, they build a miRNA-disease heterogamous network whose nodes include miRNAs and diseases, and edges include known miRNA-disease associations or connect similar miRNAs or similar diseases. The common assumption behind these methods is that the miRNAs with similar functions tend to share associations with similar diseases and vice versa [1], [3]. The early-stage methods measure the correlation scores between miRNAs and diseases by counting the number of paths connecting them in the heterogamous network [4], [5]. However, the shortcoming of the path-based methods lies in high time complexity in path enumeration. A group of method conducts random walks or heat diffusion on the heterogeneous network to propagate the association information across similar neighbors [6], [7], [8], [9], [10]. The network propagation-based methods can quickly capture the associations between miRNA and diseases by simultaneously spreading messages along multiple paths of the networks.

- *Wei Peng is with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China, and also with the Computer Technology Application Key Lab of Yunnan Province, Kunming University of Science and Technology, Kunming, Yunnan 650500, China. E-mail: weipeng1980@gmail.com.*
- *Zicheng Che is with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China. E-mail: czc2019220@163.com.*
- *Wei Dai and Shoulin Wei are with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China, and also with the Computer Technology Application Key Lab of Yunnan Province, Kunming University of Science and Technology, Kunming, Yunnan 650500, China. E-mail: dw@cnlab.net, weishoulin@kust.edu.cn.*
- *Wei Lan is with the School of Computer, Electronic and Information, Guangxi University, Nanning, Guangxi 530004, China. E-mail: lanwei@gxu.edu.cn.*

Manuscript received 25 February 2022; revised 15 June 2022; accepted 19 June 2022. Date of publication 1 July 2022; date of current version 26 December 2023.
This work was supported in part by the National Natural Science Foundation of China under Grants 61972185 and 62072124, in part by the Natural Science Foundation of Yunnan Province of China under Grant 2019FA024, in part by Yunnan Key Research and Development Program under Grant 2018IA054, and in part by Yunnan Ten Thousand Talents Plan young. (Corresponding author: Wei Peng.)
Digital Object Identifier no. 10.1109/TCBB.2022.3187739

Nevertheless, these methods may introduce false-positive links between miRNAs and diseases due to the noise in the networks. Network embedding is an emerging technique that learns feature vectors for miRNAs and disease in a latent space while keeping their connections in the original network as much as possible. The network embedding-based methods usually define a cost function that requires minimal errors between the reconstructed miRNA-disease association matrix and the known ones while preserving the original miRNA and disease similarity in the latent space. Hence, they show robustness against the network noise. There are a lot of network embedding techniques, such as Regularized Least Squares (RLS) [11], matrix factorization [12], [13], [14], [15], [16], matrix completion [17], [18], [19], which have been applied in the field of disease-related miRNA prediction. The graph convolutional networks (GCNs) are a class of deep learning methods for processing data represented by graph data structures. They can capture the interaction between network structure and node attributes simultaneously in the network embedding. Recently, GCN models have obtained wide attention from bioinformatics researchers [20]. Li *et al.* [21] designed a novel graph convolutional network with neural inductive matrix completion (NIMCGCN) for miRNA-disease association prediction. They implement separate graph convolution operations on the miRNA and disease similarity network to learn miRNA and disease latent feature representations. Then, they complete the missing values in the miRNA-disease association matrix according to these features. Li and Su [22] predict miRNA-disease associations via a graph convolution auto-encoder. They regard the miRNA similarity and their known connections with diseases as miRNA features. Similarly, they combine the disease similarity and the verified miRNA-disease associations as disease features. They run graph convolutional auto-encoder on miRNA-based and disease-based sub-networks, respectively, and combine the embeddings of miRNAs and diseases to calculate association scores of miRNA-disease.

Although the previous works have achieved significant progress and impressive performances, most rely on the miRNA-disease heterogeneous network composed of miRNA, disease, similarity networks, and the known associations. Complex diseases are caused by a combination of genetic and environmental factors. Introducing more information reveals new insights into disease pathogenesis and helps determine disease-related miRNAs. MiRNAs regulate the expression of their target genes, and the dysfunctions of genes can trigger diseases. Genes act as essential agents bridging miRNAs and diseases. Hence, we want to leverage gene information to improve the prediction of miRNA-disease associations. Previous methods used genetic information to intuitively explore the relationship between miRNA and disease. Zeng *et al.* [4] calculate the similarity of miRNAs according to the significance of two microRNAs sharing target genes. Peng *et al.* [15] measure the disease similarity based on the functional similarity of disease-related genes. Some methods [23], [24] measure the degree of microRNA related to the disease by referring to the correlation of their related genes. Peng *et al.* [25] and Li *et al.* [26] use a deep learning framework to mine links between miRNA-disease pairs. They learn miRNA and disease feature embedding by calculating the association between miRNA and genes, between diseases and genes. The difference between the two models is that Peng *et al.* [25] employ a two-layer regression model while Li *et al.* [26] use a GCN model to learn features.

Considering the complexity of the relationship among miRNA, disease, and gene. we propose a multi-relational Graph Convolutional Network model to predict MiRNA-Disease Associations (HGCNMDA) from a Heterogeneous network. HGCNMDA introduces a gene layer to construct a miRNA-gene-disease heterogeneous network. GCN models depend on neighbor's attributes to learn feature embedding, and the neighbors' attributes may update with their neighbors' attributes. We prepare two attributes for every node in the network, i.e., initial and inductive attributes. An initial attribute is generated by passing similarity data to a nonlinear layer. The inductive attribute aggregates local neighbor attributes through a graph convolutional network. Hence, the nodes of the heterogeneous network, i.e., the miRNAs, diseases, genes, may have different types of attributes. We categorize the network nodes according to their attributes and redefine their relations based on known associations and their relevance to prediction tasks. After that, HGCNMDA learns feature embeddings for miRNAs, diseases, and genes through a multi-relational graph convolutional network model that can integrate features from fine-grained node types by assigning suitable weights to different types of edges. Finally, the miRNA-disease associations were decoded by the inner product between miRNA and disease feature embeddings. HGCNMDA learns parameters based on known miRNA-disease, miRNA-gene, and disease-gene associations in a supervised end-to-end manner.

We tested our model on the human miRNA disease association dataset HMDDv3.2 under different cross-validation settings. The Experimental results show that our HGCNMDA model outperforms the state-of-the-art models [6], [21], [25], [26], [27]. The ablation study shows that introducing gene information and considering fine-grained edge types can improve the performance of our model. Case studies on two diseases, i.e., Lymphoma and Stomach Neoplasms, further prove the effectiveness of our model.

Compared with previous work, our main contributions are as follows.

1) Our model introduces an additional gene layer and constructs a miRNA-gene-disease heterogeneous network. Introducing more information reveals new insights into disease pathogenesis and helps determine disease-related miRNAs.
2) We refined the features of nodes into initial and inductive features so that the direct and indirect associations between diseases and miRNA can be considered simultaneously.
3) We design a multi-relational graph convolutional network model for gathering neighbor features, which assigns appropriate weights to different types of edges in a supervised manner and integrates information from different sources naturally and systematically. Our model can learn embedding features that are more effective in uncovering potential miRNA disease associations.

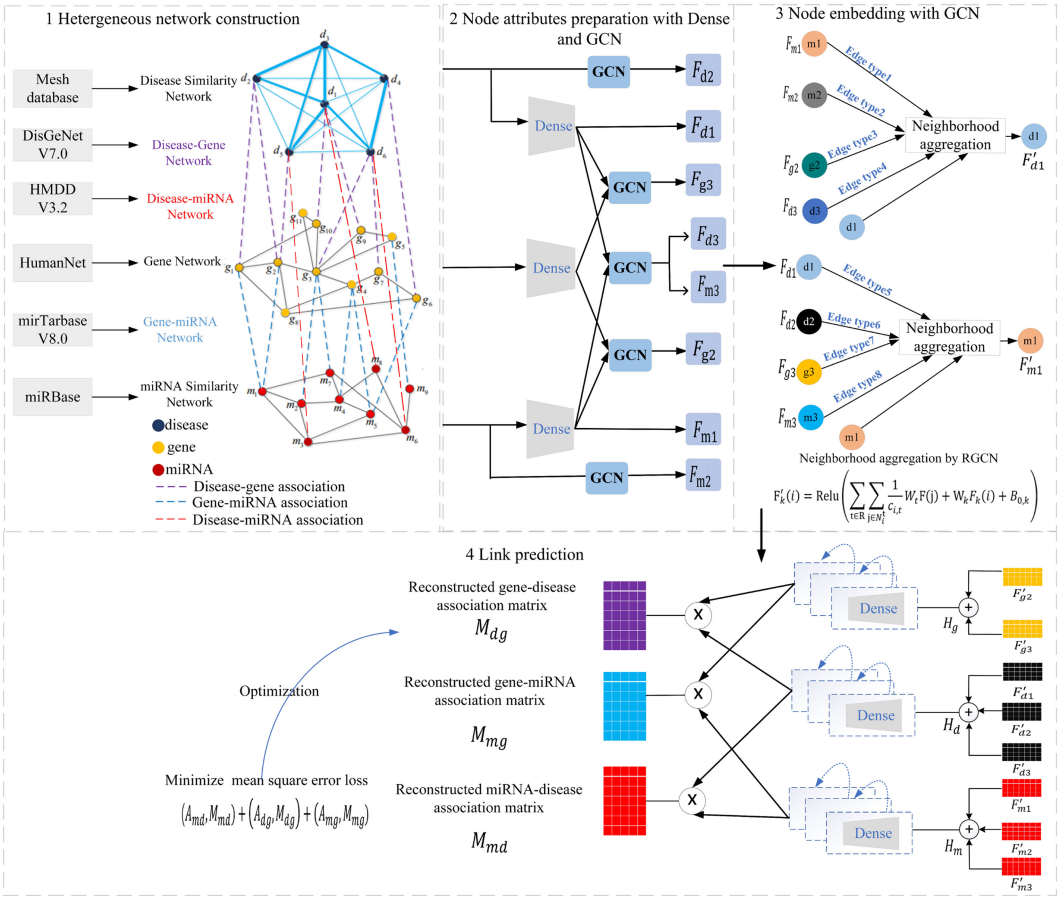


Fig. 1. The workflow HGCNMDA. HGCNMDA contains four main steps. (a) Hetergeneous network construction. (b) Node attributes preparation with Dense and GCN. (c) Node embedding with GCN. (d) Link prediction.

2 Material

The miRNA-gene-disease heterogeneous network comprises a miRNA layer, a disease layer, a gene layer and the associations linking the nodes of different layers. The known human miRNA-disease associations were from the HMDDv3.2 database [28]. The associations between miRNAs and genes were collected from mirTarbaseV8.0 [29]. We moved out the miRNAs with no relational genes. We use miRNA similarity data in [22] to generate the miRNA similarity network. The relationship between diseases and genes were obtained from the DisGeNET v7.0 database [30]. We selected the manually verified disease-gene associations for the experiment and moved out the diseases with no relational genes. The similarities between diseases were from paper [22], which were calculated based on disease semantic similarity in the Mesh database [31]. The gene network was retrieved from the HumanNet database. We only focused on the genes associated with miRNAs or diseases. Finally, our experiments involved 757 miRNAs, 435 diseases, 11216 genes, 7694 miRNA-disease associations, 148775 miRNA-gene associations and 154131 disease-gene associations. The links between miRNAs are 572103, between diseases are 34945, between genes are 1029638. We regarded the remaining unknown associations between the 757 miRNAs and 435 diseases as the negative set.

3 Methods

In our work, we cast the miRNA-disease association prediction as a link prediction problem. Our model, HGCNMDA, works in four steps to predict the links. First, we construct a miRNA-gene-disease heterogeneous network, where nodes are miRNAs, diseases, genes, and edges are generated from the miRNA and disease similarity, gene connections, miRNA-gene associations, disease-gene and the known miRNA-disease associations. The potential miRNA-disease associations can be considered as missing links in the network, and our goal is to predict these links. Our method employs a multi-relational graph convolution network (GCN) encoder to learn the nodes' latent representations (embeddings) from their attributes, considering the graph's topological structure and neighborhood. Hence, the second step of our model is to prepare the attributes for the nodes. We refined the attributes of nodes into initial and inductive attributes so that the direct and indirect association between diseases and miRNA can be considered simultaneously. Then we employ a multi-relational GCN encoder to learn the nodes' latent representations with finer-grained node type and edge type information. Finally, we adopt the learned feature embeddings to detect the links between miRNAs and diseases with an edge decoding model. All parameters of our model are learned in a supervised end-to-end way. Fig. 1 illustrates the framework of our model.

3.1 Heterogeneous Network Construction

This work constructs a miRNA-gene-disease heterogeneous network comprised of a miRNA similarity network, disease similarity network, gene network, miRNA-disease association network, miRNA-gene and disease-gene association network. Suppose n_m, n_d and n_g denote the number of miRNAs, diseases and genes, respectively. The following section will introduce the network construction in detail.

3.1.1 MiRNA Similarity Network

The miRNA similarity network is built according to the sequence similarity of miRNAs in a similar way of [22]. It is based on the assumption that miRNAs with similar sequences tend to share similar functions. The miRNA sequences were collected from the miRBase database, and the Needleman-Wunsch algorithm was employed to calculate the sequence similarity of miRNAs. For the convenience of description, let $G_m \in R^{n_m \times n_m}$ represent the miRNA similarity network, which stores the sequence similarity scores between miRNAs. $A_m \in [0,1]^{n_m \times n_m}$ is the adjacent matrix for G_m, whose values are 1 if corresponding miRNAs are similar, otherwise are zeros. $\widetilde{L}_m = \widetilde{D}_m^{-1/2} \widetilde{A}_m \widetilde{D}_m^{-1/2}$ is the normalized Laplace matrix of $\widetilde{A}_m$, where $\widetilde{A}_m = A_m + I_m$, I_m is the identity matrix and $[\widetilde{D}_m]_{ii} = \sum_j [\widetilde{A}_m]_{ij}$ is the diagonal matrix.

3.1.2 Disease Similarity Network

The disease similarity between two diseases is calculated based on their semantic similarity in the Mesh database [31], where disease terms are organized as a hierarchical directed acyclic graph (DAG). A disease in the DAG is connected from a more general parent term to a more specific child term. A disease semantic is defined as the sum of semantic contribution of itself and its children. The disease similarity between two diseases, d1 and d2 can be defined as the semantic contribution of their shared parents over the sum of the semantic contribution of the two diseases. Given $G_d \in R^{n_d \times n_d}$ represent the disease similarity network, whose elements are the semantic similarities between diseases. $A_d \in [0,1]^{n_d \times n_d}$ denotes the adjacent matrix of G_d, whose value is 1 if corresponding diseases are connected. Otherwise the value is zero. $\widetilde{L}_d = \widetilde{D}_d^{-1/2} \widetilde{A}_d \widetilde{D}_d^{-1/2}$ is the normalized Laplace matrix of $\widetilde{A}_d$, where $\widetilde{A}_d = A_d + I_d$ and $[\widetilde{D}_d]_{ii} = \sum_j [\widetilde{A}_d]_{ij}$.

3.1.3 Gene Network

Gene network was constructed by referring to public database HumanNet, whose version was HumanNet-XC. Let $\mathrm{G_g} \in \mathrm{R^{n_g \times n_g}}$ denotes the gene network whose element values were connection probability between genes undergoing min-max normalization. Given two genes i and j, their value in the adjacent matrix is defined as $\mathrm{G}_g(\mathrm{i,j}) = \frac{G'_g(i,j) - \mathrm{Min}(G'_g)}{\mathrm{Max}(G'_g) - \mathrm{Min}(G'_g)}$, where the values of G'_g were originally retrieved from the database HumanNet. $\mathrm{A_g}$ denotes the adjacent matrix of $\mathrm{G_g}$, whose element value is one if corresponding genes are connected with probability larger than zero. Otherwise, the value is zero.

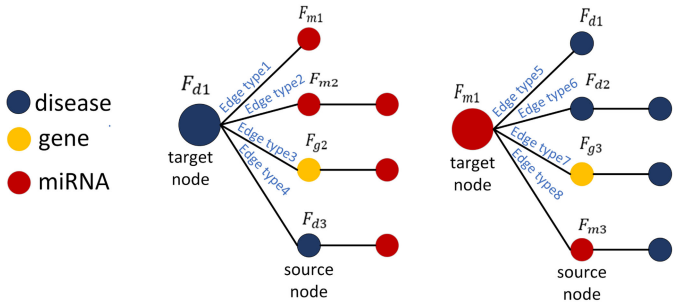


Fig. 2. Four types of connections linking to diseases or miRNAs.

3.1.4 Human miRNA-Disease Associations, Disease-Gene Associations and miRNA-Gene Associations

Here, we define $A_{md} \in R^{n_m \times n_d}$ to record the known miRNA-disease associations. If a miRNA is known to associated with a disease, $\mathrm{A_{md}(i,j)} = 1$, otherwise $\mathrm{A_{md}(i,j)} = 0$. Since nodes of the miRNA-disease association network are heterogeneous, we extended the matrix A_{md} to a matrix $\widetilde{A}_{md}$ of shape $(\mathrm{n_m + n_d}) \times (\mathrm{n_m + n_d})$. Specifically, $\widetilde{A}_{md}$ is defined as $\widetilde{A}_{md} = \begin{vmatrix} I_m & A_{md} \\ A_{md}^T & I_d \end{vmatrix}$. The $\widetilde{A}_{md}$ can be normalized as $\tilde{L}_{md} = \widetilde{D}_{md}^{-1/2} \widetilde{A}_{md} \widetilde{D}_{md}^{-1/2}$, where $\tilde{D}_{md} = \begin{vmatrix} \widetilde{D}_{md(m)} & 0 \\ 0 & \widetilde{D}_{md(d)} \end{vmatrix}$ with $[\widetilde{D}_{md(m)}]_{ii} = \sum_j [A_{md}]_{ij} + 1, [\widetilde{D}_{md(d)}]_{ii} = \sum_j [A_{md}^T]_{ij} + 1$. Similarly, we assume that $A_{mg} \in [0,1]^{n_m \times n_g}$ and $A_{dg} \in [0,1]^{n_d \times n_g}$ store the miRNA-gene and disease-gene associations. $\widetilde{L}_{mg}$ and $\widetilde{L}_{dg}$ are the normalized Laplace matrix of $\widetilde{A}_{mg}$ and $\widetilde{A}_{dg}$.

3.2 Node Attributes Preparation

Diseases or miRNAs have complex interactions with other nodes in the heterogeneous network. They aggregate neighbors' features and themselves through the GCN model to update their feature representations. We call the disease or miRNA nodes to learn feature representations as target nodes. Their neighbors are source nodes. We hope these target nodes can gather features from the source nodes contributing to miRNA-disease association identification. Hence, in this work, diseases only consider four types of connections with miRNAs (See Edge types 1-4 in Fig. 2). The four types of connections link diseases with miRNAs directly, link diseases with miRNAs indirectly through miRNAs similar to source nodes, indirectly through genes, and indirectly through diseases similar to target nodes. The four types of edges help diseases collect neighbors' features from two types of miRNAs. The one has initial input features, and the other updates their features by their neighbors. Similarly, miRNAs only consider four types of connections with diseases, including links from miRNAs to diseases directly and from miRNAs to diseases indirectly through diseases similar to source nodes or indirectly through genes or indirectly through miRNAs similar to target nodes (See Edge types 5-8 in Fig. 2). MiRNAs can learn features from diseases with two different types of features. Therefore, we prepare two kinds of features for every node. One is the initial feature,

and the other is the inductive attribute. We introduce initial attributes for every node passing the similarity data into a nonlinear layer. The inductive attributes of every node are encoded by implementing graph convolution operation on the heterogeneous network to collect attributes from their local neighbors.

Let F_{d1}, F_{m1}, F_{g1} be the initial attributes of diseases, miRNAs and genes, respectively. We separately put the disease similarity network G_d, miRNA similarity network G_m, and gene network G_g through three fully connected networks as follows.

$$F_{d1} = \mathrm{relu}(W_d G_d + b_d) \tag{1}$$

$$F_{m1} = \mathrm{relu}(W_m G_m + b_m) \tag{2}$$

$$F_{g1} = \mathrm{relu}\big(W_g G_g + b_g\big) \tag{3}$$

where W_d, W_m, W_g, b_d, b_m, b_g are the parameters of the fully connected layers.

We calculate inductive attributes for every node using a graph convolutional network model because the nodes can also aggregate attributes from their local neighbors. For example, a miRNA node of the heterogeneous network may connect to three different neighbor types, including miRNAs, diseases, and genes. Hence, we run five separate graph convolutional network models on the miRNA similarity network, disease similarity network, known miRNA-disease association network, miRNA-gene association network, and disease-gene association network to produce inductive attributes for miRNAs, diseases and genes. We get three types of inductive attributes for miRNAs (F_{m2}, F_{m3}, F_{m4}), three types of inductive attributes for diseases (F_{d2}, F_{d3}, F_{d4}), and two types of inductive attributes for genes (F_{g2}, F_{g3}). Mathematically, the inductive attribute of miRNA (F_{m2}) generated from the miRNA similarity network is defined as follows.

$$F_{m2}^{(l)} = \mathrm{relu}\Big(\widetilde{L}_m F_{m2}^{(l-1)} \theta_{m2}^{(l-1)} + b_{m2}^{(l-1)}\Big) \tag{4}$$

$$F_{m2}^{(0)} = G_m$$

Where l is the layers of the GCN model. $\theta_{\mathrm{m2}}^{(1-1)} \in R^{n_m \times h}$ is the weight parameter matrix of the lth layer GCN model with n_m of miRNAs and h of hidden layer nodes, $\theta_{\mathrm{m2}}^{(1-1)}$ is the bias matrix of the lth layer GCN model. Similarly, the following equation can calculate the inductive attributes of disease (F_{d2}) on the disease similarity network(G_d).

$$F_{d2}^{(l)} = \mathrm{relu}\Big(\widetilde{L}_d F_{d2}^{(l-1)} \theta_{d2}^{(l-1)} + b_{m2}^{(l-1)}\Big) \tag{5}$$

$$F_{d2}^{(0)} = G_d$$

We ran graph convolution operation on the miRNA-disease associations to obtain another inductive attribute for miRNAs (F_{m3}) and diseases (F_{d3}) (see Eq. (6)).

$$\begin{vmatrix} F_{m3}^{(l)} \\ F_{d3}^{(l)} \end{vmatrix} = \mathrm{F}_{\mathrm{md}}^{(l)} = \mathrm{relu}\Big(\tilde{L}_{md}\ \mathrm{F}_{md}^{(l-1)} \theta_{\mathrm{m3}}^{(l-1)} + \mathrm{b}_{\mathrm{m3}}^{(l-1)}\Big) \tag{6}$$

$$\tilde{L}_{md} = \begin{vmatrix} \widetilde{D}_{md(m)}^{-1} & \widetilde{D}_{md(m)}^{-1/2} A_{md} \widetilde{D}_{md(d)}^{-1/2} \\ \widetilde{D}_{md(d)}^{-1/2} A_{md}^T \widetilde{D}_{md(m)}^{-1/2} & \widetilde{D}_{md(d)}^{-1} \end{vmatrix}$$

$$F_{md}^{(0)} = \begin{vmatrix} F_{m1} \\ F_{d1} \end{vmatrix}$$

Where l is the layers of the GCN model. $\theta_{m3}^{(l)}$ and $\mathrm{b}_{m3}^{(l)}$ are parameter matrices of the lth layer GCN model. A_{md} is the adjacent matrix of the known miRNA-disease associations. $\tilde{L}_{md}$ is the normalized Laplace matrix of the matrix $\widetilde{A}_{md}$. Similarly, we utilize the GCN model on the miRNA-gene associations to define the inductive attributes for miRNAs (F_{m4}) and gene(F_{g2}) by Eq. (7) and on the disease-gene associations to produce the inductive attributes for diseases (F_{d4}) and gene (F_{g3}) by Eq. (8).

$$\begin{vmatrix} F_{m4}^{(l)} \\ F_{g2}^{(l)} \end{vmatrix} = \mathrm{F}_{\mathrm{mg}}^{(l)} = \mathrm{relu}\Big(\tilde{L}_{mg}\ \mathrm{F}_{mg}^{(l-1)} \theta_{m4}^{(l-1)} + \mathrm{b}_{m4}^{(l-1)}\Big) \tag{7}$$

$$\tilde{L}_{mg} = \begin{vmatrix} \widetilde{D}_{mg(m)}^{-1} & \widetilde{D}_{mg(m)}^{-1/2} A_{mg} \widetilde{D}_{mg(g)}^{-1/2} \\ \widetilde{D}_{mg(g)}^{-1/2} A_{mg}^T \widetilde{D}_{mg(m)}^{-1/2} & \widetilde{D}_{mg(g)}^{-1} \end{vmatrix}$$

$$F_{mg}^{(0)} = \begin{vmatrix} F_{m1} \\ F_{g1} \end{vmatrix}$$

$$\begin{vmatrix} F_{d4}^{(l)} \\ F_{g3}^{(l)} \end{vmatrix} = \mathrm{F}_{\mathrm{dg}}^{(l)} = \mathrm{relu}\Big(\tilde{L}_{dg}\ \mathrm{F}_{dg}^{(l-1)} \theta_{d4}^{(l-1)} + \mathrm{b}_{d4}^{(l-1)}\Big) \tag{8}$$

$$\tilde{L}_{dg} = \begin{vmatrix} \widetilde{D}_{dg(d)}^{-1} & \widetilde{D}_{dg(d)}^{-1/2} A_{dg} \widetilde{D}_{dg(g)}^{-1/2} \\ \widetilde{D}_{dg(g)}^{-1/2} A_{dg}^T \widetilde{D}_{dg(d)}^{-1/2} & \widetilde{D}_{dg(g)}^{-1} \end{vmatrix}$$

$$F_{dg}^{(0)} = \begin{vmatrix} F_{d1} \\ F_{g1} \end{vmatrix}$$

Here A_{mg} and A_{dg} denote the miRNA-gene association matrix and disease-gene association matrix. $\tilde{L}_{mg}$ and $\tilde{L}_{dg}$ are the normalized Laplace matrices of the matrix $\widetilde{A}_{mg}$ and $\widetilde{A}_{dg}$, respectively. $\widetilde{A}_{mg}$ and $\widetilde{A}_{dg}$ can be calculated by plussing an identity matrix with A_{mg} and A_{dg}. l is a predefined parameter denoting the layers of the GCN model.

3.3 Node Embedding With Graph Convolution Network

After preparing attributes for every node in the heterogeneous network, we encode the representations for a given miRNA, disease or gene by employing a multi-relational GCN model that incorporates the neighborhood representation according to different relations types. Every node with various attributes (i.e., initial and inductive attributes) affects its neighbors differently. Hence, disease or miRNA nodes only consider connecting edges relevant to the prediction task (See Fig. 2). The multi-relational GCN model tackles each edge type independently and can naturally encode a given node by passing neighbors' features conducive to exploring the

TABLE 1
Table 1 the Connection Status Between the Eight Types of Nodes in the Heterogeneous Network

	m1	m2	m3	d1	d2	d3	g2	g3
m1			1	1	1			1
m2				1				
m3	1							
d1	1	1				1	1	
d2	1							
d3				1				
g2				1				
g3	1							

potential miRNA-disease associations. As Fig. 2 shown, the disease node d1 with F_{d1} features updates its presentation by collecting features from four types of connecting neighbors. The four type neighbors are m1, m2, g2 and d3 with attributes F_{m1}, F_{m1}, F_{g2} and F_{d3}. Similarly, the miRNA node m1 with F_{m1} features updates its presentation by collecting features from four types of connecting neighbors. The four type neighbors are d1, d2, g3 and m3 with attributes F_{d1}, F_{d2}, F_{g3} and F_{m3}. These attributes are generated according to the explanation in Section 3.2. Note that the features propagation among these nodes is bidirectional. For example, disease node d1 gathers features from node m2, but also d1's features are passed to node m2. Hence, we only focus on eight types of nodes in the heterogeneous network, including m1, m2, m3, d1, d2, d3,g2, g3 with attributes F_{m1}, F_{m2}, F_{m3}, F_{d1}, F_{d2}, F_{d3}, F_{g2}, F_{g3}, respectively. We refer to m1, m2, m3 as three miRNA node types, refer to d1, d2, d3 as disease node types, and refer to g2, g3 as gene node types. According to Fig. 2, the connection status between the eight types of nodes is laid out in Table 1. Table values of 1 mean the two types of nodes connect according to their original similarity network or association matrix. Otherwise, the two types of nodes will not connect. For example, the value of m1 and d1 is 1. The node types of m1 and d1 will connect if the corresponding miRNA associates with the disease according to their association matrix. Hence, we establish a multi-relational graph and introduce Eq. (9) to encode a node feature by incorporating the neighborhood aggregation with finer-grained edge type information. The key idea is to collect features for a given node vi from the features vector of itself and the transformed feature vectors of neighboring nodes by assigning suitable weights to different types of edges.

$$\mathrm{F}'_k(i) = \mathrm{relu}\left(\sum_{\mathrm{t}\in\mathrm{R}}\sum_{\mathrm{j}\in N_i^{\mathrm{t}}}\frac{1}{c_{i,\mathrm{t}}}W_{\mathrm{t}}\mathrm{F(j)} + \mathrm{W}_k F_k(i) + B_{0,k}\right) \quad (9)$$

Suppose there are eight node types $\{$ m1, m2, m3, d1, d2, d3, g2, g3$\}$, let F=$\{F_{m1}, F_{m2}, F_{m3}, F_{d1}, F_{d2}, F_{d3}, F_{g2}, F_{g3}\}$ be the feature set of every node type. $\mathrm{F}_k(i)$ and $\mathrm{F}'_k(i)$ denote the input and the output features of the node vi of the kth node type, respectively. t indicates the edge types. Here, we define fourteen types of edges (see Table 1), i.e., the edges connecting node types m1 and m3, m3 and m1, m1 and d1, d1 and m1, m1 and d2, d2 and m1, m1 and g3, g3 and m1, d1 and m2, m2 and d1, d1 and d3, d3 and d1, d1 and g2, g2 and d1. $\mathrm{N}_{\mathrm{i}}^t$ is the set of neighbors related to the node vi by the edge type t. F(j) refers to the input feature of the node $v_j \in \mathrm{N}_{\mathrm{i}}^t$. W_t refers to the weight parameter of the edge type t. $c_{i,t} = |N_i^t|$ depicts the number of neighbors related to node vi through edge type t. W_k is weight parameter of the node itself in the node type k. We use relu as the activation function.

After aggregating neighbors' information through the multi-relational GCN model, we obtain the final feature representation of every node type. They are defined as $F' = \{F'_{m1}, F'_{m2}, F'_{m3}, F'_{d1}, F'_{d2}, F'_{d3}, F'_{g2}, F'_{g3}\}$. $F'_{m1}, F'_{m2}, F'_{m3}$ are all feature vectors of miRNAs. We add them bit by bit to generate the final miRNAs feature representation (called H_m). Similarly, we add F'_{d1}, F'_{d2}, F'_{d3} bit by bit to obtain the final feature representation for diseases(called H_d) and add F'_{g2}, F'_{g3} bit by bit to get the final feature representations for genes (called H_g).

$$H_m = F'_{m1} + F'_{m2} + F'_{m3} \quad (10)$$
$$H_d = F'_{\mathrm{d1}} + F'_{\mathrm{d2}} + F'_{\mathrm{d3}} \quad (11)$$
$$H_g = F'_{\mathrm{g2}} + F'_{\mathrm{g3}} \quad (12)$$

3.4 Link Prediction

Finally, by separately passing the representation of miRNAs (H_m), diseases(H_d) and genes(H_g) into three different three-layer fully connected layers, we employ an edge decoder to reconstruct the miRNA-disease, miRNA-genes, and disease-genes associations.

$$Y_m^3(H_m) = \mathrm{relu}\big(W_m^3\big(\mathrm{relu}\big(W_m^2\big(\mathrm{relu}\big(W_m^1 H_m + b_m^1\big)\big) + b_m^2\big)\big) + b_m^3\big) \quad (13)$$
$$Y_d^3(H_d) = \mathrm{relu}\big(W_d^3\big(\mathrm{relu}\big(W_d^2\big(\mathrm{relu}\big(W_d^1 H_d + b_d^1\big)\big) + b_d^2\big)\big) + b_d^3\big) \quad (14)$$
$$Y_g^3(H_g) = \mathrm{relu}\Big(W_g^3\Big(\mathrm{relu}\Big(W_g^2\Big(\mathrm{relu}\Big(W_g^1 H_g + b_g^1\Big)\Big) + b_g^2\Big)\Big) + b_g^3\Big) \quad (15)$$

$W_m^{(l)}$(here,$l = \{1, 2, 3\}$) denotes the weight matrix of the lth layer in the fully connected layer. $Y_m^3(H_m) \in R^{\mathrm{n_m}\times\mathrm{t}}$ is the output feature matrix of miRNA of the 3th fully connected layer, where n_m and t are the number of miRNAs and the output dimensions. The predicted association probability between miRNA and disease is calculated by the inner product of their representations (See Eq. (16)).

$$M_{md} = Y_m^3(H_m) \times Y_{\mathrm{d}}^3(H_d)^T \quad (16)$$

We also leverage the representations of miRNA, genes and diseases to calculate the relation between miRNAs and genes (see Eq. (17)), between diseases and genes (see Eq. (18)). To ensure the representations can preserve the structure of the heterogeneous network, we design a loss function to combine the mean square error loss between the three reconstructed matrices and the original ones (Eq. (19).

$$M_{mg} = Y_m^3(H_m) \times Y_{\mathrm{g}}^3\left(H_g\right)^T \tag{17}$$

$$M_{dg} = Y_{\mathrm{d}}^3(H_d) \times Y_{\mathrm{g}}^3\left(H_g\right)^T \tag{18}$$

$$\begin{aligned} Loss = {} & (1-\alpha)||P_\delta(A_{md} - M_{md})||_F^2 \, | \\ & + \alpha||P_{\bar{\delta}}(A_{md} - M_{md})||_F^2 \\ & + ||A_{dg} - M_{dg}||_F^2 \\ & + ||A_{mg} - M_{mg}||_F^2 + \|W\|^2 \end{aligned} \tag{19}$$

Where δ and $\bar{\delta}$ denote the positive and negative sample of the training set. $P_\delta(.)$ is the projection of the matrix onto the set δ. A_{md}, A_{dg} and A_{mg} store the known miRNA-disease, gene-disease, and gene-miRNA association matrix. W denotes the parameters of the model. Parameter α balances the weight of positive sample loss and the negative sample loss. Our model is implemented with python3.6 and Pytorch1.6.0. Adam algorithm is used to minimize. The learning rate is set to 0.0001. For the GCN model, the dimension of the latent spaces is set 256. The single fully connected layer for generating initial attributes contains 256 hidden nodes. Before decoding, the fully connected layers for nonlinear transformation have three hidden layers with 256, 128, and 64 units, respectively. We tune the parameters α in Eq. (19) and epoch for different cross-validation task. For the randomly zeroing cross-validation, α and epoch are set to 0.4 and 100. For the multi-column zeroing, α and epoch are set to 0.5 and 100. For the multi-row zeroing cross-validation, α and epoch are set to 0.3 and 50.

4 Results

To evaluate the effectiveness of our model HGCNMDA, we compared it with five baselines, including MDACNN [25], NIMCGCN [21], GCSENet [26], ThrRWMDE [6], CCA-based [27]. The NIMCGCN uses graph convolutional networks (GCNs) to learn miRNA and disease potential feature representations from miRNA and disease similarity networks. The MDACNN, GCSENet, CCA-based, and ThrRWMDE introduce gene information for disease-related miRNA identification. The MDACNN and GCSENet adopt a supervised deep-learning framework for miRNA-disease association determination. They design the miRNA features and disease features by calculating the associations between miRNAs and genes, between diseases and genes. The difference between the two models lies in the MDACNN employs a two-layer regression model while GCSENet uses a GCN model to learn features. The CCA-based method predicts miRNA-disease association by analyzing the canonical correlation between miRNAs and their target genes. ThrRWMDE runs random walks on a three-layer heterogeneous network to predict microRNA-disease associations based on network propagation. For a fair comparison, all methods take the same input similarity data and miRNA-disease association data. Their parameters are set by the recommendation in their paper or adjusted appropriately to perform best.

4.1 Experimental Setting

We tested our method and the baselines on the HMDDv3.2. The performance of every method was measured by the receiver operating characteristic (ROC) curve and the area under the curve (AUC). We also compared their best F1 scores and corresponding precision, recall values. In the experiments, we conduct cross-validation under three different Settings:

1) Randomly zeroing cross-validation: All known miRNA-disease associations as positive samples are randomly divided into five non-overlapping parts. One part of positive samples and an equal amount of negative samples were randomly selected as test data. The remaining positive samples and the remaining negative samples are selected for training. Randomly zeroing cross-validation aims to test the ability of every model to find missing miRNA-disease associations.
2) Multi-column zeroing cross-validation: The columns of the miRNA-disease association matrix correspond to diseases. The values of 1/5 columns of the miRNA-disease association matrix were cleared for testing. All the remaining columns were the training set. Multi-column zeroing cross-validation tests the performance of every model on detecting miRNAs associated with new diseases.
3) Multi-row zeroing cross-validation: The rows of the miRNA-disease association matrix refer to miRNAs. In multi-row zeroing cross-validation, 1/5 miRNAs were selected as testing data, and all their associations were removed. The remaining miRNAs were the training set. We repeat every cross-validation ten times and report the average results.

4.2 Parameter Setting

We leverage a multi-layer graph convolutional network (GCN) model to generate the inductive attributes for every node in the heterogeneous network. The parameter l controls the number of layers of the GCN model (see Eqs. (4), (5), (6), (7), and (8)). The larger the value of l is, the deeper order neighbors the GCN model involves. Since, in this work, diseases/miRNA only consider four types of connections with miRNAs/diseases (See Fig. 2). To investigate how the parameter l affects the prediction performance, we set different values of l ranging from 1 to 3 for an inductive attribute while only collecting features for a given node from the neighbors with initial features and the considered inductive attributes. For example, we set parameter l in Eq. (4) from 1 to 3 to obtain three different inductive miRNA attributes (F_{m2}) and encode miRNAs and diseases for the prediction task by collecting neighborhood features through edge types 1, 2 and 5. The result is illustrated in Fig. 3a. Fig. 3b shows the prediction performance of our model that sets different values for the parameter l in Eq. (5) and encode miRNAs and diseases by aggregating neighborhood features through edge types 1, 6 and 5. As shown in Fig. 3, we observed that our model performs best when setting GCN layer l to 1 for all inductive attributes except for the inductive attributes F_{m2}. Introducing deep neighbors may bring noisy features for the miRNA-disease association prediction task. In contrast, high order neighbors in the miRNA similarity network may contribute to identifying potential miRNA-disease associations. Hence, we set the GCN layer parameter l to 1 for all inductive attributes at following experiments.

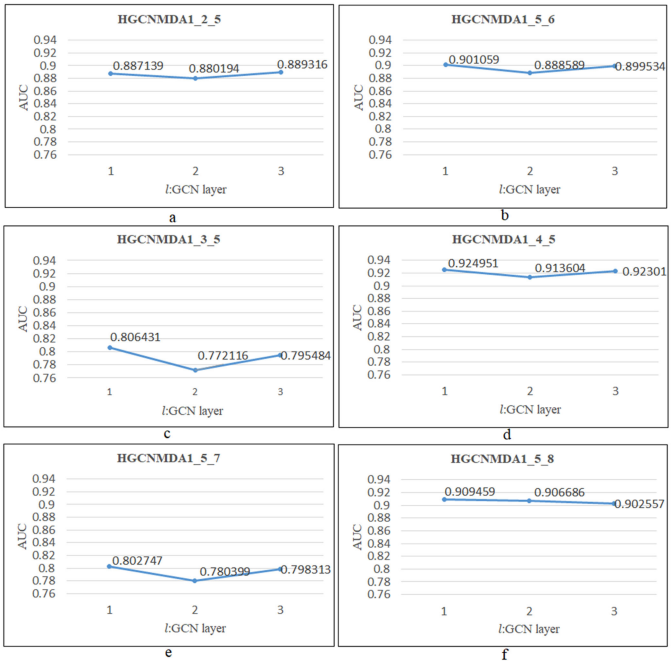


Fig. 3. The prediction performance on different GCN layer.

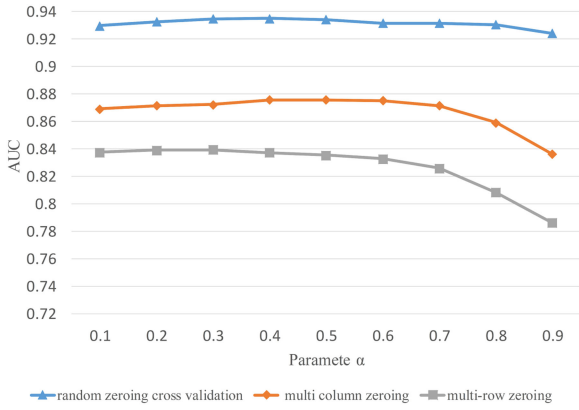


Fig. 4. The prediction performance on different α.

Parameter α balances the weight of positive and negative sample losses. We tune the parameter α in Eq. (19) ranging from 0.1 to 0.9 for randomly zeroing, multi-column zeroing and multi-row zeroing cross-validations.From the line in Fig. 4, we observe that for random zeroing cross validation, when the parameter α When set to 0.4, our model has a relatively high AUC value. Under the cross verification of multi column zeroing and multi row zeroing, α setting 0.5 and 0.3 respectively will result in the highest AUC value. Therefore, we adopt α the value of 0.4 is used for random zeroing, 0.5 for multi column and 0.3 for multi row zeroing cross validation.

4.3 Comparison With Other Methods on HMDDv3.2 Dataset

We compared HGCNMDA with baselines on the HMDDv3.2 dataset under three cross-validation settings. We first assessed the performances of HGCNMDA under randomly zeroing cross-validation to test whether it can recover the known miRNA-disease associations. Table 2 shows the average AUC values, the best F1 scores, and corresponding Precision and Recall values of every model under randomly zeroing cross-validation. We observed that HGCNMDA leads to the highest average AUC, AUPR and F1 Score values under this situation. Its AUC, AUPR and F1 score values reach 93.5%, 93.5% and 86.4%, which are 3.1%, 2.9% and 2.8% better than the second-best method ThrRWMDE. We then evaluate the performance of HGCNMDA under randomly zeroing out multi-columns or multi-rows cross-validation to test its capacity of detecting associations for new diseases or new miRNAs. Table 3 reports that HGCNMDA still controls the highest average AUC, AUPR and F1 Score values when recommending miRNAs for new diseases. Its AUC, AUPR and F1 Scores are 87.59%, 19.92% and 26.69%, which achieves 4.7%, 4.8% and 4.3% improvement compared with the second-best method ThrRWMDE. As for multi-rows cross-validation, we observe from Table 4 that our HGCNMDA model still stands at the method list with top performance. It has the highest AUC value, 83.9%. Its AUPR and F1 score values are inferior to the CCA-based method and comparable with the ThrRWMDE and MDACNN.

The CCA-based method predicts potential diseases for a new miRNA by referring to the Canonical correlations between its target gene and disease profiles. It only focuses on a limited number of pairs of extracted canonical components. Hence the CCA-based method produces higher AUPR and F1 score values under multi-rows cross-validation. However, its AUC value is 79.14%, which is relatively lower than other methods. Moreover, it cannot predict potential miRNAs for new diseases. GCSENet runs GNN models on miRNA-gene and disease-gene association networks separately to learn miRNA features and disease features. However, the lengths of its miRNA and disease feature vectors are too long, relying on the number of genes. GCSENet cannot learn efficient miRNA and disease features for miRNA-disease association prediction and results in lower performance than all other methods under the three cross-validations. Our model is superior to all baselines in AUC values under the three cross-validation settings. The observed improvement in the performance of HGCNMDA can be partially attributed to its successful introduction of gene information in miRNA-disease association determination. Moreover, HGCNMDA tackles different-type neighbor nodes, assigns suitable weights to different-type edges in a supervised manner, and makes the embedding more effective in revealing the potential miRNA-disease associations.

4.4 Ablation Study

Our HGCNMDA aggregates different types of neighbors' features through different edges to learn feature embedding for miRNAs and diseases. We hope miRNAs and diseases can gather features from their neighbors that help predict miRNA-disease associations. Hence, HGCNMDA only considers four types of connections for miRNAs and four types of connections for diseases (See Fig. 2). We set up the following eight model variations to investigate which connections contribute to its excellent performance.

TABLE 2
Performance Comparison of Every Method Under Randomly Zeroing Cross Validation

Methods	AUC	AUPR	PRECISION	RECALL	F1 SCORE
ThrRWMDE	0.9049±0.000024	0.9067±0.000031	0.7867±0.000273	**0.8911±0.000344**	0.8353±0.000032
CCA-based	0.7898±0.000103	0.8446±0.000068	0.8254±0.000167	0.6726±0.000390	0.7410±0.000164
MDACNN	0.8613±0.000080	0.8590±0.000082	0.7471±0.000513	0.8472±0.000507	0.7935±0.000076
GCSENet	0.6974±0.000115	0.6869±0.000585	0.5826±0.000472	0.8406±0.001706	0.6870±0.000017
NIMCGCN	0.8087±0.000834	0.8132±0.000686	0.6896±0.001653	0.8284±0.000940	0.7510±0.000327
HGCNMDA	**0.9351±0.000017**	**0.9351±0.000018**	**0.8406±0.0004291**	0.8896±0.000516	**0.8639±0.000032**

TABLE 3
Performance Comparison of Every Method Under Randomly Zeroing Out Multi-Column Cross Validation

Methods	AUC	AUPR	PRECISION	RECALL	F1 SCORE
ThrRWMDE	0.8206±0.000298	0.1869±0.001453	0.2141±0.001661	0.3328±0.001329	0.2582±0.001075
CCA-based	-	-	-	-	-
MDACNN	0.8282±0.000293	0.1514±0.002273	0.1866±0.002838	0.2976±0.002848	0.2238±0.002029
GCSENet	0.6857±0.001803	0.1060±0.010546	0.0580±0.000262	**0.4025±0.022184**	0.0981±0.000475
NIMCGCN	0.5991±0.001730	0.0351±0.000094	0.0654±0.002040	0.3685±0.053311	0.0866±0.000948
HGCNMDA	**0.8759±0.000275**	**0.1992±0.004645**	**0.2151±0.003598**	0.3725±0.003543	**0.2669±0.002322**

TABLE 4
Performance Comparison of Every Method Under Randomly Zeroing Out Multi-Rows Cross Validation

Methods	AUC	AUPR	PRECISION	RECALL	F1 SCORE
ThrRWMDE	0.8259±0.000146	0.1891±0.000205	0.2614±0.000336	0.2926±0.000583	0.2752±0.000180
CCA-based	0.7914±0.000168	**0.2399±0.001069**	**0.2885±0.001282**	0.3538±0.000843	**0.3159±0.000533**
MDACNN	0.8373±0.000227	0.1866±0.000863	0.2264±0.001033	0.2852±0.002183	0.2496±0.000741
GCSENet	0.6763±0.002859	0.1138±0.018039	0.0561±0.000185	**0.4192±0.043769**	0.0949±0.000363
NIMCGCN	0.8083±0.000486	0.1524±0.000379	0.2344±0.000970	0.2604±0.002076	0.2425±0.000437
HGCNMDA	**0.8395±0.000200**	0.1706±0.000212	0.2618±0.000431	0.2882±0.000526	0.2734±0.000219

HGCNMDA1_5: HGCNMDA gathers features from neighbors linking directly to diseases or miRNAs. It only considers edge types 1 and 5 (See Fig. 2).

HGCNMDA1_2_5_6: HGCNMDA gathers features from neighbors linking directly to diseases or miRNAs. It also gathers features from neighbors linking indirectly to miRNAs or disease through nodes similar to source nodes. It only considers edge types 1, 2, 5 and 6(See Fig. 2).

HGCNMDA1_4_5_8: HGCNMDA gathers features from neighbors linking directly to diseases or miRNAs. It also gathers features from neighbors linking indirectly to miRNAs or disease through nodes similar to target nodes. In other words, HGCNMDA considers edge types 1, 4, 5 and 8 (See Fig. 2).

HGCNMDA1_2_4_5_6_8: It performs prediction by combining edge types of HGCNMDA1_2_5_6 and HGCNMDA 1_4_5_8. In other words, HGCNMDA considers edge types 1, 2, 4, 5, 6 and 8(See Fig. 2).

HGCNMDA3_7: HGCNMDA gathers features from neighbors linking indirectly to diseases or miRNAs through genes. It only considers edge types 3 and 7(See Fig. 2).

HGCNMDA1_3_5_7: Besides the edge types of HGCNMDA1_5, HGCNMDA additionally gathers features from neighbors linking indirectly to diseases or miRNAs through genes. In other words, HGCNMDA considers edge types 1, 5, 3 and 7(See Fig. 2).

HGCNMDA1_3_2_5_7_6: Besides the edge types of HGCNMDA1_2_5_6, HGCNMDA additionally gathers features from neighbors linking indirectly to diseases or miRNAs through genes. That is, HGCNMDA considers edge types 1, 5, 3, 7, 2 and 6(See Fig. 2).

HGCNMDA1_3_4_5_7_8: Besides the edge types of HGCNMDA1_4_5_8, HGCNMDA additionally gathers features from neighbors linking indirectly to diseases or miRNAs through genes. In other words, HGCNMDA considers edge types 1, 5, 3, 7, 4 and 8(See Fig. 2).

Table 5 presents the performance comparison between HGCNMDA and its variants on HMDDv3.2 dataset under randomly zero cross-validation. We observe that considering direct neighbors of miRNA and disease leads to good performance (see HGCNMDA1_5). HGCNMDA 1_4_5_8 performs better than HGCNMDA1_5. In contrast, HGCNMDA 1_3_5_7 performs worse than HGCNMDA 1_5. It suggests that HGCNMDA considering features from neighbors linking indirectly through nodes similar to target nodes improves the prediction performance. However, gathering neighbors' features similar to source nodes decreases the performance of HGCNMDA. Only gathering features

TABLE 5
Ablation Studying

Methods	AUC	AUPR	PRECISION	RECALL	F1 SCORE
$HGCNMDA1\_5$	0.9024±0.000038	0.9024±0.000052	0.7916±0.000418	0.8708±0.000410	0.8289±0.000039
$HGCNMDA1\_2\_5\_6$	0.8936±0.000037	0.8915±0.000052	0.7813±0.000315	0.8657±0.000281	0.8211±0.000039
$HGCNMDA1\_4\_5\_8$	0.9299±0.000025	0.9333±0.000024	**0.8466±0.000215**	0.8828±0.000284	0.8641±0.000036
$HGCNMDA1\_2\_5\_6\_4\_8$	0.9299±0.000028	0.9333±0.000030	0.8354±0.000425	0.8880±0.000596	0.8604±0.000035
$HGCNMDA3\_7$	0.7670±0.000082	0.7715±0.000111	0.6433±0.000375	0.8085±0.001033	0.7157±0.000042
$HGCNMDA1\_3\_5\_7$	0.7817±0.000239	0.7822±0.000209	0.6405±0.000631	0.8450±0.001094	0.7278±0.000152
$HGCNMDA1\_3\_2\_5\_7\_6$	0.8043±0.000181	0.7834±0.000324	0.6801±0.000404	0.8683±0.000595	0.7622±0.000086
$HGCNMDA1\_3\_4\_5\_7\_8$	0.9332±0.000013	0.9338±0.000018	0.8409±0.000253	0.8926±0.000217	**0.8657±0.000020**
$HGCNMDA_{mdmg}$	0.9310±0.000021	0.9321±0.000021	0.8287±0.000312	**0.9019±0.000253**	0.8634±0.000036
$HGCNMDA_{mddg}$	0.9296±0.000012	0.9325±0.000011	0.8383±0.000246	0.8858±0.000253	0.8611±0.000019
$HGCNMDA_{md}$	0.9299±0.000019	0.9337±0.000020	0.8350±0.000331	0.8952±0.000356	0.8637±0.000030
$HGCNMDA$	**0.9351±0.000017**	**0.9351±0.000018**	0.8406±0.0004291	0.8896±0.000516	0.8639±0.000032

from neighbors linking indirectly through genes cannot help find miRNA-disease associations. HGCNMDA $1\_3\_4\_5\_7\_8$ improves HGCNMDA $1\_4\_5\_8$ by additional consideration of gene information, suggesting the gene information can help determine disease-related miRNAs. Finally, the HGCNMDA model achieves higher AUC and F1 score values than all its variants, suggesting that HGCNMDA successfully improves the miRNA-disease association detection by integrating multiple neighbors features related to the association prediction task.

The loss function (Eq. (19)) combines errors in three matrix reconstructions. To test which part contributes to the excellent performance of HGCNMDA, we did an ablation study by setting the following variants. $HGCNMDA_{mdmg}$ adopts the loss function only considers the errors in miRNA-disease association matrix reconstruction and the errors in miRNA-gene association matrix reconstruction. $HGCNMDA_{mddg}$ adopts the loss function only considers the errors in miRNA-disease association matrix reconstruction and the errors in disease-gene association matrix reconstruction. $HGCNMDA_{md}$ employs the loss function, considering the miRNA-disease association matrix reconstruction errors. The results show that HGCNMDA model performs better under the constraint of minimizing the errors in reconstructing the associations among miRNAs, diseases and genes.

5 Case Study

To further illustrate the performance of HGCNMDA, we apply it to predict the miRNAs related to Lymphoma and Stomach Neoplasms. The dbDEMC and miRCancer database are used as the benchmark. dbDEMC [32] (database of

TABLE 6
Top 50 Related miRNAs of Lymphoma Predicted by HGCNMDA on HMDDv3.2 Dataset

Top1-25 miRNA	Evidence	Top26-50 miRNA	Evidence
hsa-mir-146a	dbDEMC,HMDDv3.2	hsa-mir-93	dbDEMC,HMDDv3.2
hsa-mir-155	dbDEMC,miRCancer,HMDDv3.2	hsa-let-7b	dbDEMC
hsa-mir-21	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-146b	dbDEMC
hsa-mir-17	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-214	dbDEMC
hsa-mir-34a	dbDEMC	hsa-mir-192	dbDEMC
hsa-mir-221	dbDEMC,HMDDv3.2	hsa-mir-195	dbDEMC
hsa-mir-126	dbDEMC,HMDDv3.2	hsa-mir-132	dbDEMC
hsa-mir-223	dbDEMC,miRCancer	hsa-mir-27b	dbDEMC,HMDDv3.2
hsa-mir-222	dbDEMC,HMDDv3.2	hsa-mir-30b	dbDEMC
hsa-mir-142	PMID:28031239,HMDDv3.2	hsa-mir-106a	dbDEMC
hsa-mir-150	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-125a	dbDEMC,HMDDv3.2
hsa-mir-145	dbDEMC	hsa-mir-19a	dbDEMC,miRCancer,HMDDv3.2
hsa-mir-15a	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-29c	dbDEMC,HMDDv3.2
hsa-mir-122	dbDEMC,HMDDv3.2	hsa-mir-26b	dbDEMC
hsa-mir-20a	dbDEMC,HMDDv3.2	hsa-mir-23a	dbDEMC
hsa-mir-29a	dbDEMC	hsa-mir-27a	dbDEMC
hsa-mir-210	dbDEMC,HMDDv3.2	hsa-mir-34b	dbDEMC
hsa-mir-31	dbDEMC,HMDDv3.2	hsa-mir-183	dbDEMC
hsa-mir-143	dbDEMC,HMDDv3.2	hsa-let-7c	dbDEMC
hsa-mir-106b	dbDEMC	hsa-mir-373	dbDEMC
hsa-mir-30a	dbDEMC	hsa-mir-205	dbDEMC
hsa-mir-15b	dbDEMC	hsa-mir-25	dbDEMC
hsa-mir-206	dbDEMC	hsa-mir-16-1	miRCancer,HMDDv3.2
hsa-mir-18a	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-424	dbDEMC
hsa-mir-182	dbDEMC	hsa-mir-144	dbDEMC

TABLE 7
Top 50 Related miRNAs of Stomach Cancer Predicted by HGCNMDA on HMDDv3.2 Dataset

Top1-25 miRNA	Evidence	Top26-50 miRNA	Evidence
hsa-mir-17	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-20b	dbDEMC,miRCancer,HMDDv3.2
hsa-mir-146a	miRCancer,HMDDv3.2	hsa-mir-145	dbDEMC,miRCancer,HMDDv3.2
hsa-mir-21	dbDEMC,miRCancer,HMDDv3.2	hsa-let-7c	dbDEMC,miRCancer
hsa-mir-155	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-224	dbDEMC,miRCancer,HMDDv3.2
hsa-mir-142	PMID:30178386,HMDDv3.2	hsa-let-7 d	dbDEMC,miRCancer
hsa-mir-20a	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-144	miRCancer,HMDDv3.2
hsa-mir-15a	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-373	dbDEMC,miRCancer,HMDDv3.2
hsa-mir-34a	dbDEMC,miRCancer,HMDDv3.2	hsa-let-7i	dbDEMC,miRCancer,HMDDv3.2
hsa-mir-222	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-16-1	miRCancer,HMDDv3.2
hsa-mir-106b	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-424	dbDEMC
hsa-mir-93	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-26b	dbDEMC,miRCancer,HMDDv3.2
hsa-mir-15b	dbDEMC,HMDDv3.2	hsa-mir-146b	miRCancer,HMDDv3.2
hsa-mir-31	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-34b	miRCancer,HMDDv3.2
hsa-mir-30a	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-200b	dbDEMC,miRCancer,HMDDv3.2
hsa-mir-126	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-186	dbDEMC,miRCancer,HMDDv3.2
hsa-mir-221	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-148a	dbDEMC,miRCancer,HMDDv3.2
hsa-mir-106a	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-210	miRCancer,HMDDv3.2
hsa-mir-150	miRCancer,HMDDv3.2	hsa-let-7 g	dbDEMC,miRCancer,HMDDv3.2
hsa-let-7b	miRCancer,HMDDv3.2	hsa-mir-182	dbDEMC,miRCancer,HMDDv3.2
hsa-mir-30b	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-27b	dbDEMC,miRCancer,HMDDv3.2
hsa-mir-18a	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-335	miRCancer, HMDDv3.2
hsa-mir-223	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-206	dbDEMC,miRCancer,HMDDv3.2
hsa-mir-195	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-183	dbDEMC,miRCancer,HMDDv3.2
hsa-mir-122	dbDEMC,miRCancer,HMDDv3.2	hsa-let-7e	dbDEMC,miRCancer
hsa-mir-143	dbDEMC,miRCancer,HMDDv3.2	hsa-mir-16-2	PMID:18449891,HMDDv3.2

differentially expressed miRNAs in human cancers) collects cancer-related miRNAs by analyzing miRNA differential expressions between cancer and normal samples. miRCancer [33] is a microRNA-cancer association database constructed by text mining on literature. Lymphoma starts in the immune system cells and is one of the ten deadly diseases. Table 6 shows the top 50 Lymphoma -related miRNAs detected by HGCNMDA on the HMDDv3.2 data set. Based on the dbDEMC and miRCancer database records, we found that all the top 50 miRNAs except one miRNA(hsa-mir-142) are related to Lymphoma. hsa-mir-142 is also reported to control the development of Lymphoma in PMID:28031239 [34].

Stomach cancer develops in the stomach or gastric. It may spread to other organs. Stomach cancer is among the most prevalent cancers worldwide and is the third most leading cause of cancer-related death globally. Stomach cancer remains difficult to cure, primarily because the actual cause of stomach cancer is not yet known. Recent studies report that abnormalities in miRNA expression are associated with the initiation and progression of stomach cancer [35], [36], [37]. We apply our model HGCNMDA to rank miRNAs for stomach cancer. Table 7 lists the top 50 of the most potential miRNAs. We observed that 48 of 50 miRNAs are stomach cancer-related in the dbDEMC and miRCancer database. The remaining two miRNAs hsa-mir-142 and hsa-mir-16-2 are found to be associated with stomach tumor growth in literature [35], [36].

6 Conclusion

This work proposes a multi-relational graph convolutional network model to predict miRNA-disease associations (HGCNMDA) from a miRNA-gene-disease heterogeneous network. HGCNMDA introduces an additional gene layer and constructs a miRNA-gene-disease heterogeneous network. To collect direct and indirect neighbors' features, it prepares two types attributes for every nodes in the heterogeneous network. One is the initial attributes generated by passing the similarity data into a nonlinear transformer and the other is inductive attributes aggregating local neighbor's attributes via GCN network. Then the HGCNMDA runs a multi-relational graph convolutional network model on the heterogeneous network to encoder miRNAs and diseases with finer-grained edge type information. Finally, we adopt the learned feature embeddings to reveal the miRNA-disease associations. We test our model on human miRNA-disease association dataset HMDDv3.2. Compared with the state-of-the art methods, our model leads to outstanding performance on identifying missing miRNA-disease associations and also performs well on recommending related miRNAs/diseases to new diseases/ miRNAs. The ablation studies show that HGCNMDA can improve its performance by considering gene information and it successfully improves the miRNA-disease association detection by integrating multiple neighbors features related to the association prediction task.

References

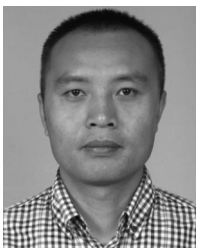
[1] Q. Zou, J. Li, L. Song, X. Zeng, and G. Wang, "Similarity computation strategies in the microRNA-disease network: A survey," *Brief. Funct. Genomic.*, vol. 15, no. 1, pp. 55–64, 2016.
[2] C. Yan, F.-X. Wu, J. Wang, and G. Duan, "PESM: Predicting the essentiality of miRNAs based on gradient boosting machines and sequences," *BMC Bioinf.*, vol. 21, no. 1, pp. 1–9, 2020.
[3] Y. Ding, X. Lei, B. Liao, and F.-X. Wu, "MLRDFM: A multi-view laplacian regularized deepfm model for predicting miRNA-disease associations," *Brief. Bioinf.*, vol. 23, no. 3, 2022, Art. no. bbac079.

[4] X. Zeng, X. Zhang, Y. Liao, and L. Pan, "Prediction and validation of association between microRNAs and diseases by multipath methods," *Biochimica et Biophysica Acta Gen. Subjects*, vol. 1860, no. 11, pp. 2735–2739, 2016.
[5] Z.-H. You *et al.*, "PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction," *PLoS Comput. Biol.*, vol. 13, no. 3, 2017, Art. no. e1005455.
[6] W. Peng, W. Lan, Z. Yu, J. Wang, and Y. Pan, "A framework for integrating multiple biological networks to predict microRNA-disease associations," *IEEE Trans. Nanobiosci.*, vol. 16, no. 2, pp. 100–107, Mar. 2017.
[7] W. Peng, W. Lan, J. Zhong, J. Wang, and Y. Pan, "A novel method of predicting microRNA-disease associations based on microRNA, disease, gene and environment factor networks," *Methods*, vol. 124, pp. 69–77, 2017.
[8] W. Peng, M. Li, L. Chen, and L. Wang, "Predicting protein functions by using unbalanced random walk algorithm on three biological networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 2, pp. 360–369, Mar./Apr. 2017.
[9] C. Yan, G. Duan, N. Li, L. Zhang, F.-X. Wu, and J. Wang, "PDMDA: Predicting deep-level miRNA–disease associations with graph neural networks and sequence features," *Bioinformatics*, vol. 38, no. 8, pp. 2226–2234, 2022.
[10] Z. Fang and X. Lei, "Prediction of miRNA-circRNA associations based on k-NN multi-label with random walk restart on a heterogeneous network," *Big Data Mining Analytics*, vol. 2, no. 4, pp. 261–272, 2019.
[11] X. Chen and G.-Y. Yan, "Semi-supervised learning for potential human microRNA-disease associations inference," *Sci. Rep.*, vol. 4, no. 1, pp. 1–10, 2014.
[12] W. Lan, J. Wang, M. Li, J. Liu, F.-X. Wu, and Y. Pan, "Predicting microRNA-disease associations based on improved microRNA and disease similarities," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 6, pp. 1774–1782, Nov./Dec. 2018.
[13] Q. Xiao, J. Luo, C. Liang, J. Cai, and P. Ding, "A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations," *Bioinformatics*, vol. 34, no. 2, pp. 239–248, 2018.
[14] C. Yan, J. Wang, P. Ni, W. Lan, F.-X. Wu, and Y. Pan, "DNRLMF-MDA: Predicting microRNA-disease associations based on similarities of microRNAs and diseases," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 1, pp. 233–243, Jan./Feb. 2019.
[15] W. Peng, J. Du, W. Dai, and W. Lan, "Predicting miRNA-disease association based on modularity preserving heterogeneous network embedding," *Front. Cell Devlop. Biol.*, vol. 9, 2021, Art. no. 893.
[16] Y. Ding, X. Lei, B. Liao, and F.-X. Wu, "Predicting miRNA-disease associations based on multi-view variational graph auto-encoder with matrix factorization," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 1, pp. 446–457, Jan. 2022.
[17] J.-Q. Li, Z.-H. Rong, X. Chen, G.-Y. Yan, and Z.-H. You, "MCMDA: Matrix completion for miRNA-disease association prediction," *Oncotarget*, vol. 8, no. 13, 2017, Art. no. 21187.
[18] X. Chen, L. Wang, J. Qu, N.-N. Guan, and J.-Q. Li, "Predicting miRNA–disease association based on inductive matrix completion," *Bioinformatics*, vol. 34, no. 24, pp. 4256–4265, 2018.
[19] X. Chen, L.-G. Sun, and Y. Zhao, "NCMCMDA: MiRNA–disease association prediction through neighborhood constraint matrix completion," *Brief. Bioinf.*, vol. 22, no. 1, pp. 485–496, 2021.
[20] W. Peng, Q. Tang, W. Dai, and T. Chen, "Improving cancer driver gene identification using multi-task learning on graph convolutional network," *Brief. Bioinf.*, vol. 23, no. 1, 2022, Art. no. bbab432.
[21] J. Li, S. Zhang, T. Liu, C. Ning, Z. Zhang, and W. Zhou, "Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction," *Bioinformatics*, vol. 36, no. 8, pp. 2538–2546, 2020.
[22] L. Li, Y.-T. Wang, C.-M. Ji, C.-H. Zheng, J.-C. Ni, and Y.-S. Su, "GCAEMDA: Predicting miRNA-disease associations via graph convolutional autoencoder," *PLoS Comput. Biol.*, vol. 17, no. 12, 2021, Art. no. e1009655.
[23] Q. Jiang *et al.*, "Prioritization of disease microRNAs through a human phenome-micrornaome network," *BMC Syst. Biol.*, vol. 4, no. 1, pp. 1–9, 2010.
[24] H. Shi *et al.*, "Walking the interactome to identify human miRNA-disease associations through the functional link between mirna targets and disease genes," *BMC Syst. Biol.*, vol. 7, no. 1, pp. 1–12, 2013.
[25] J. Peng *et al.*, "A learning-based framework for miRNA-disease association identification using neural networks," *Bioinformatics*, vol. 35, no. 21, pp. 4364–4371, 2019.
[26] Z. Li, K. Jiang, S. Qin, Y. Zhong, and A. Elofsson, "GCSENET: A GCN, CNN and senet ensemble model for microRNA-disease association prediction," *PLoS Comput. Biol.*, vol. 17, no. 6, 2021, Art. no. e1009048.
[27] H. Chen, Z. Zhang, and D. Feng, "Prediction and interpretation of miRNA-disease associations based on miRNA target genes using canonical correlation analysis," *BMC Bioinf.*, vol. 20, no. 1, pp. 1–8, 2019.
[28] Z. Huang *et al.*, "HMDD V3. 0: A database for experimentally supported human microRNA–disease associations," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1 013–D1 017, 2019.
[29] H.-Y. Huang *et al.*, "mirtarbase 2020: Updates to the experimentally validated microRNA–target interaction database," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D148–D154, 2020.
[30] J. Piñero *et al.*, "DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D833–D839, 2017.
[31] H. J. Lowe and G. O. Barnett, "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches," *Jama*, vol. 271, no. 14, pp. 1103–1108, 1994.
[32] Z. Yang *et al.*, "DBDEMC: A database of differentially expressed miRNAs in human cancers," in *Proc. BMC Genomics*, vol. 11, no. 4, pp. 1–8, 2010.
[33] B. Xie, Q. Ding, H. Han, and D. Wu, "mircancer: A microRNA–cancer association database constructed by text mining on literature," *Bioinformatics*, vol. 29, no. 5, pp. 638–644, 2013.
[34] C. Fernandez *et al.*, "MicroRNAs 142–3p, miR-155 and miR-203 are deregulated in gastric malt lymphomas compared to chronic gastritis," *Cancer Genomic. Proteomic.*, vol. 14, no. 1, pp. 75–82, 2017.
[35] J. Yan, B. Yang, S. Lin, R. Xing, and Y. Lu, "Downregulation of miR-142–5p promotes tumor metastasis through directly regulating CYR61 expression in gastric cancer," *Gastric Cancer*, vol. 22, no. 2, pp. 302–313, 2019.
[36] L. Xia *et al.*, "miR-15b and miR-16 modulate multidrug resistance by targeting BCL2 in human gastric cancer cells," *Int. J. Cancer*, vol. 123, no. 2, pp. 372–379, 2008.
[37] Z.-X. Yang, C.-Y. Lu, Y.-L. Yang, K.-F. Dou, and K.-S. Tao, "MicroRNA-125b expression in gastric adenocarcinoma and its effect on the proliferation of gastric cancer cells," *Mol. Med. Rep.*, vol. 7, no. 1, pp. 229–232, 2013.

Wei Peng received the PhD degree in computer science from Central South University, China, in 2013. Currently, she is a professor with the Kunming University of Science and Technology, China. Her research interests include bioinformatics and data mining.

Zicheng Che rececived the BE degree from the Chengdu College of Arts and Sciences, in 2019. He is currently working toward the master degree with the Kunming University of Science and Technology, China. His research interests include bioinformatics and feature extraction.

Wei Dai received the PhD degree in computer application from the University of Chinese Academy of Sciences, China, in 2018. Currently, he is an associate professor with the Kunming University of Science and Technology. His research interests include bioinformatics, distributed and cloud computing, data mining.



Shoulin Wei received the PhD degree in Astronomical Technique and Method from the University of Chinese Academy of Sciences, China, in 2017. Currently, he is an associate professor with the Kunming University of Science and Technology. His research interests include bioinformatics, distributed and cloud computing, data mining.

Wei Lan received the PhD degree in computer science from Central South University, China, in 2016. He is currently a associate professor with the School of Computer, Electronics and Information, Guangxi University, Nanning, Guangxi, China. Her main research interests include bioinformatics and data mining.