Prediction of miRNA-Disease Associations Based on Hybrid Gated GNN and Multi-Data Integration

Yeqiang Wang†
College of Information Engineering
Northwest A & F University
Yangling, China
wangyeqiang@nwafu.edu.cn

Sharen Yun†
College of Information Engineering
Northwest A & F University
Yangling, China
yunsr1215@nwafu.edu.cn

Yuchen Zhang*
College of Information Engineering
Northwest A & F University
Yangling, China
yczhang@nwafu.edu.cn

Xiujuan Lei*
School of Computer Science
Shaanxi Normal University
Xian, China
xjlei@snnu.edu.cn

2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822161

***Abstract*—It is well-established that miRNAs play a crucial role in the occurrence and development of diseases. Current miRNA-disease associations prediction research faces several challenges, including model bias due to data sparsity, information loss from overlooked complex relationships during feature fusion and insufficient capability of existing methods to capture the intricate relationships (between miRNAs, genes, lncRNAs and diseases), thereby limiting prediction accuracy. Based on hybrid gated GNN and multi-data fusion, a method (PMDGGM) for predicting miRNA-disease associations is proposed in this study. PMDGGM constructed seven similarity networks by comprehensively considering the relationships between miRNA and related genes, miRNA, lncRNA and diseases. It provides a solid foundation for feature fusion and information propagation. Subsequently, the method captures the complex relationships between heterogeneous nodes through a bilinear pooling layer and uses a gating mechanism to fuse multi-source heterogeneous features, thereby predicting miRNA-disease associations more accurately. The experimental results show that the method performs well and has significant advantages in predicting the miRNA-disease associations. Among various evaluation metrics, especially the AUCs of ROC and PR curves, the performance of method is outstanding, reaching a high level of 0.9413 and 0.9362. The study conducted case analyses on two diseases heart failure and acute myeloid leukemia. The predicted associated miRNAs can be validated by existing biomedical research efforts. The source code and data of PMDGGM can be publicly accessed on GitHub for further research and verification: https://github.com/WangYeQianger/PMDGGM**

Keywords—miRNA-Disease Associations, Gated GNN, Multi-Data Integration, Graph Attention, GraphSAGE

I. INTRODUCTION

MicroRNAs are crucial non-coding RNAs that play a vital role in the occurrence, development and mutation of diseases. For instance, the let-7 family inhibits cell proliferation during embryonic development [1], while the lin-4 promotes these processes [2]. Abnormal miRNA expression is linked to cancers, cardiovascular and neurological diseases [3, 4]. Identifying miRNA-disease associations is crucial for understanding pathological mechanisms and developing treatment strategies. However, experimental validation is resource-intensive, highlighting the urgent need for computational prediction methods.

Traditional machine learning methods, such as SVM, have been widely used in predicting miRNA-disease associations. Chen et al. [5] proposed a semi-supervised learning method. Wang et al. [6] introduced logistic model trees, integrating multi-source information like sequences and similarities for prediction. With the development of deep learning, Wen et al. [7] calculated multiple similarities between miRNAs and diseases to extract node features and used a two-layer graph convolutional network (GCN) to predict miRNA-disease associations. Zhao et al. proposed the NSAMDA model [8]. They integrated miRNA sequence similarity and constructed a miRNA-disease heterogeneous graph using the fused miRNA feature and integrated disease similarity information. In our previous studies, we also used matepath2vec to identify the associations between non-coding RNAs and diseases [9].

Existing graph models have made progress in bioinformatics but still face several challenges. Firstly, methods relying on low-order information to measure molecular similarity perform poorly. Second, using algebraic operations in Euclidean space to analyze irregular network data, though simple and intuitive, often disrupts the network's topological structure, affecting accurate identification of information propagation paths. What's more, cross-source network information transmission, node updating, and integrating multi-source heterogeneous features remain challenging, especially when combining miRNA, related genes, lncRNA and disease-associated information.

Given the existing challenges, the introduction of Gated GNN is essential. Its gating mechanism can better capture high-order information and complex associations, addressing the issue of insufficient low-order information and performing well in similarity measurements. Additionally, Gated GNN preserves network topology in non-Euclidean space, ensuring the accuracy of information propagation paths. It also efficiently integrates and updates multi-source heterogeneous features, overcoming the limitations of traditional methods. Based on the advantages of Gated GNN, this study focuses on multiple key factors to improve the prediction accuracy of the model. At the network construction level, based on our previous work [10], the interactions between miRNA, lncRNA, genes and diseases were integrated to build a more representative

† indicates that the authors have the same contribution.
* indicates corresponding authors.

This work was supported in part by National Natural Science Foundation of China (62402394, 62272288), the Natural Science Basic Research Program of Shaanxi Province (2024JC-YBQN-0645), the Fundamental Research Funds for the Central Universities, Shaanxi Normal University (GK202302006), the Chinese Universities Scientific Fund, Northwest A&F University (No.2452023023) and Chinese College Students' Innovative Entrepreneurial Training Plan Program (S202410712315).

similarity network. At the method level, this study captured the complex associations between heterogeneous nodes through the bilinear pooling layer and gating mechanism, ultimately constructing more reasonable miRNA and disease features. The framework of the method is shown in Fig.1. The contribution of this method lies in its ability to better preserve and utilize network topological structure information while overcoming the feature loss problem caused by excessive simplification in traditional methods.

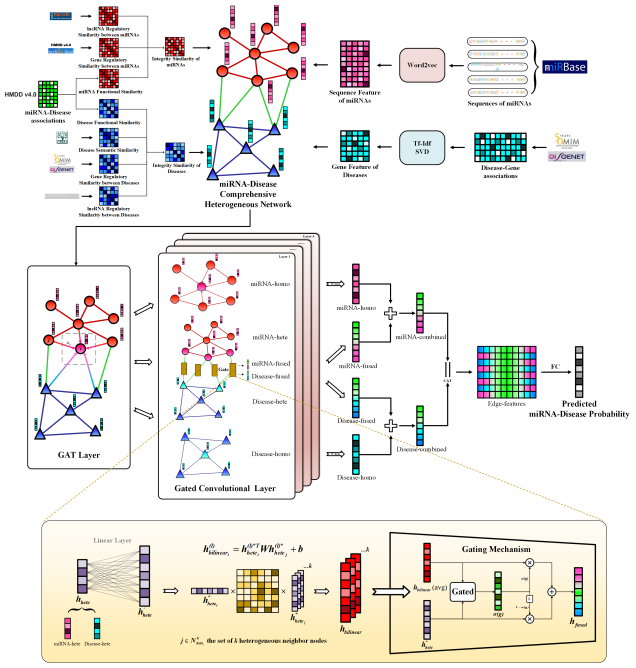

Fig. 1. The Flowchart of PMDGGM

II. Materials and Methods

A. Data preprocessing

The miRNA-disease associations database obtained in this study was mainly from The Human MicroRNA Disease Database (HMDD v4.0) [11]. This database includes 953 miRNAs and 483 diseases and 8,708 miRNA-disease associations data. Additionally, the MeSH IDs for diseases were obtained from Medical Subject Headings [12]. To further explore the relationships among miRNAs, this paper retrieved 12,695 miRNA-lncRNA associations from the ENCORI database [13], 101,704 miRNA-gene associations from HMDD v4.0 [11] and miRTarBase [14], and all miRNA sequences from the miRbase [15] database. Meanwhile, The study obtained 2,554 disease-lncRNA associations from LncRnadisease [16]. Extensive studies have demonstrated the critical role of gene diseases and their mutations, prompting us to gather 198,239 disease-associated genes from OMIM [17] and DisGeNET [18]. After deduplication and data ID organization, this paper finally obtained 812 miRNAs, 438 diseases and 7,820 miRNA-disease associations. TABLE I provides a summary of the statistical information for these datasets.

TABLE I. THE STATISTICAL OVERVIEW OF THE DATA PROCESSED

Data Type	*Number*	*Data Source*
miRNAs	812	HMDD v4.0 [11]
diseases	438	HMDD v4.0 [11]
disease-gene associations	198,239	OMIM [17], DisGeNET [18]
miRNA-lncRNA associations	12,695	ENCORI [13]
miRNA-gene associations	101,704	miRTarBase [14], HMDD 4.0 [11]
miRNA sequences	812	miRbase [15]
disease-lncRNA associations	2,554	LncRnadisease [16]

B. Construction of Disease Similarity Network

1) Disease Semantic Similarity

Many diseases are named by symptoms and affected organs. Exploring the semantic similarity of diseases can help to study miRNA-disease associations. This study adopted Wang's [19] method to extract biological information of diseases from MeSH [12] and convert it into a directed acyclic graph (DAG). Based on their relative positions and common ancestor information in the DAG, this paper obtained the disease semantic similarity matrix DS_s with dimensions 438×438 can be computed.

2) Gene Regulatory Similarity between Diseases

Genetic mutations, aberrant expression or other genetic variations are significant factors in the onset of diseases. Incorporating genetic information is crucial when assessing disease similarity. This study collected a large amount of disease-associated genetic data and obtained a gene-based disease similarity matrix DS_g using Pearson correlation coefficients (PCC) [20]. The specific calculation formula is as follows:

$$DS_{g_{ij}} = \left| \frac{\sum_{k=1}^{n}(d_{ik}-\bar{d}_i)(d_{jk}-\bar{d}_j)}{\sqrt{\sum_{k=1}^{n}(d_{ik}-\bar{d}_i)^2}\sqrt{\sum_{k=1}^{n}(d_{jk}-\bar{d}_j)^2}} \right| \quad (1)$$

where d_{ik} and d_{jk} represent the association values (0 or 1) of diseases i and j with $gene_k$, respectively. The total number of genes is expressed as n, while $\bar{d}_i$ and $\bar{d}_j$ are the average association number of diseases i and j across all genes.

3) lncRNA Regulatory Similarity and Functional Similarity of Diseases

Long non-coding RNAs (lncRNAs) play a crucial role in regulating gene expression and cellular functions. This study utilized cosine similarity method [21] to assess the similarity of disease regulation on lncRNAs DS_l. The formula for calculating is as follows:

$$DS_l(d_i,d_j) = \frac{d_i \cdot d_j}{\|d_i\|\|d_j\|} \quad (2)$$

where d_i and d_j represent the association value vectors of disease i and disease j with all lncRNAs, respectively. Each element of these vectors indicates the associations of a specific disease with a lncRNA, typically being 0 or 1. Similarly, based on the collection of 7,820 miRNA-disease associations entries, the Jaccard similarity calculation [22] method was utilized to evaluate the functional similarity of diseases DS_f. Finally, the disease integrated similarity DS is as follows:

$$DS = \alpha DS_s + \beta DS_g + \gamma DS_l + (1-\alpha-\beta-\gamma)DS_f \quad (3)$$

C. Construction of miRNA Similarity Network

1) Gene Regulatory Similarity and Functional Similarity of miRNA

miRNAs are integral components of the gene expression regulatory network. Thus, in constructing the miRNA similarity network, this study incorporates miRNA-gene associations. The approach is similar to that for calculating gene-based disease similarity, using PCC [20] to assess miRNA similarity at the gene level, providing insights into the role of miRNAs in the gene regulatory network MS_g. Similarly, based on the collection of 7,820 miRNA-disease associations entries, the functional similarity matrix MS_f among miRNAs was assessed using cosine similarity [21].

2) lncRNA Regulatory Similarity between miRNA

Extensive research has demonstrated the complex regulatory interactions between lncRNAs and miRNAs. Therefore, this study also used Spearman's rank [23] correlation coefficient based on miRNA-lncRNA associations data to determine the similarity of miRNAs. For each miRNA m_i, this study extracted its associations vector VM with all lncRNAs from the miRNA-lncRNA associations matrix, with a length of n.

$$VM(m_i) = \{VM_{i1}, VM_{i2}, ..., VM_{in}\} \tag{4}$$

For miRNAs m_1 and m_2, their associations vectors are $VM(m_1)$ and $VM(m_2)$, respectively. Next, this paper ranked the elements in these two vectors, resulting in the rank vectors $R(VM(m_1))$ and $R(VM(m_2))$. This study calculated the rank difference d_i for each lncRNA association.

$$d_i = R(VM(m_1))_i - R(VM(m_2))_i \tag{5}$$

After that, this study calculated the Spearman correlation coefficient ρ_{m_1,m_2} between miRNAs m_1 and m_2:

$$\rho_{m_1,m_2} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{6}$$

where n is the length of the vector, which represents the number of lncRNAs. $\sum_{i=1}^{n} d_i^2$ is the sum of the squared rank differences for all lncRNA associations. This paper can calculate an 812 × 812 miRNA association matrix MS_l regulated by lncRNAs. Finally, the miRNA integrated similarity MS is as follows:

$$MS = \alpha MS_f + \beta MS_g + (1-\alpha-\beta)MS_l \tag{7}$$

D. Construction of Heterogeneous Networks

The previous sections detailed the calculation methods for various similarities. Based on this foundation, this paper integrated miRNA-disease associations MDA_s, miRNA similarity MS and disease similarity DS to construct a miRNA-disease comprehensive heterogeneous network A. Specifically, this network was constructed by processing and concatenating these three components as follows:

$$A = \begin{bmatrix} MS & MDA_s \\ MDA_s^T & DS \end{bmatrix} \tag{8}$$

where A is the adjacency matrix of a miRNA-disease comprehensive network.

E. miRNA and Disease Feature Extraction

Research shows that miRNA sequences are highly specific and conserved. Analyzing these sequences can reveal potential target genes and roles in gene regulatory networks. Diseases often involve abnormal expression or mutations in multiple genes, reflecting molecular mechanisms and pathogenic pathways. This study selected miRNA sequences and disease-related genes as biological features and used different methods for calculation.

1) Sequence Features of miRNAs

To better extract biological features of miRNAs, the study chose word2vec [24] for in-depth analysis. This approach uses neural networks to map sequences into a vector space, where similar items are projected to nearby locations. This process effectively extracts the functional and structural features of miRNA sequences. To capture local patterns within the sequences, each miRNA is first segmented into smaller fragments. Then, a sliding window is used to generate all subsequences of length k fragments. Then, a sliding window is used to generate all subsequences of length k (with k set to 3). For example, given $M = AGGACC$, the generated k-mers are: AGG, GGA, GAC and ACC. These fragments are then trained using the Skip-gram model to capture the semantic relationships between them. Next, the mean of the word vectors for all fragments within the sequence is calculated to obtain a mean vector that represents the global feature of M. Finally, sequence M is represented as a fixed-dimensional feature vector that encapsulates the semantic information of the entire sequence. The sequence features of miRNA MF were obtained with dimensions 812×360.

2) Gene Features of Diseases

This study utilized disease-associated genes as the reference standard to evaluate the biological features of diseases, considering the importance of genes in disease mechanisms. First, Term Frequency-Inverse Document Frequency (TF-IDF) [25] values were calculated to integrate both the direct associations between diseases and genes and the relative importance of genes across all diseases. Term Frequency (TF) was used to quantify the occurrence of gene g in a specific disease d, reflecting the gene's relative importance within that disease and highlighting its associations. Next, Inverse Document Frequency (IDF) was calculated to measure the importance of gene g across all diseases. Finally, Singular Value Decomposition (SVD) [26] was applied for dimensionality reduction, transforming the high-dimensional TF-IDF matrix into a lower-dimensional representation, which produced the disease feature matrix DF.

3) Feature Initialization

After deriving the biological features of miRNAs and diseases, these features can be vertically concatenated to obtain an initial feature matrix $h^{(0)}$ with dimensions 1250×360. The calculation process for the initial feature matrix is as follows:

$$h^{(0)} = \begin{bmatrix} MF \\ DF \end{bmatrix} \tag{9}$$

where MF represents the biological features of miRNAs with dimensions 812×360 and DF represents the biological features of diseases with dimensions 438×360.

F. PMDGGM Method

To capture the complex relationships between miRNA and disease nodes, PMDGGM integrates graph attention networks (GAT) [27], hybrid gated convolutional layers (customized graph sampling and aggregation convolutional layers based on bilinear pooling and gating mechanisms) to predict the associations between miRNAs and diseases. During training, the sample edge set used includes negative samples obtained through random walk with restarts [28] method from the miRNA-disease heterogeneous network A, and a set of the most valuable miRNA-disease associations sample edges.

1) Graph Attention Layer

After feature initialization and the construction of the miRNA-disease comprehensive heterogeneous network, this method first inputs the features of all nodes into an attention layer to capture important interactions between nodes and identify which nodes contribute the most to the feature representation of the target node. Higher weights are then assigned to these nodes. Specifically, the attention

layer calculates the feature representation of node i using the following formula:

$$h_i^{(1)} = \text{LeakyReLU}(\sum_{j \in N(i)} \alpha_{ij} W h_j^{(0)}) \tag{10}$$

where $h_j^{(0)}$ is the initial feature representation of node j, $N(i)$ represents the set of neighbor nodes of node i, α_{ij} is the attention weight between node i and node j, and W is the linear transformation matrix. The attention weight α_{ij} is calculated using the following formula:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(a^T [W h_i \,||\, W h_j]\right)\right)}{\sum_{k \in N(i)} \exp\left(\text{LeakyReLU}\left(a^T [W h_i \,||\, W h_k]\right)\right)} \tag{11}$$

where a_{ij} is a learnable weight vector and $||$ denotes the vector concatenation. In this way, the attention mechanism can effectively focus on the important neighbor node features, thereby improving the quality of feature representation.

2) Gated Convolutional Layer

Node features are processed through one layer of gated convolutional layers, which aggregate, fuse and update features from homogeneous and heterogeneous nodes. The aggregation involves combining features from both types of neighbors and integrating them with a bilinear pooling layer using a gating mechanism.

The gated convolutional layer modifies the traditional GraphSAGE convolution [29] layer's feature aggregation method to handle different node types in a heterogeneous network. This study designs a method to aggregate features from homogeneous and heterogeneous nodes separately. First, features are aggregated from neighboring nodes of the same type. The specific process is as follows:

$$h_{homo_i}^{(l)} = \sum_{j \in N_{homo_i}} \frac{h_j^{(l)}}{\sqrt{\deg(i) \cdot \deg(j)}} \tag{12}$$

where $h_{homo_i}^{(l)}$ represents the homogeneous node features of node i at the l layer, N_{homo_i} denotes the set of homogeneous neighbor nodes of node i, and $\deg(i)$ and $\deg(j)$ are the degrees of node i and node j, respectively. Then features are aggregated from neighboring nodes of the different type. The feature aggregation formula for heterogeneous neighbor nodes is as follows:

$$h_{hete_i}^{(l)} = \sum_{j \in N_{hete_i}} \frac{h_j^{(l)}}{\sqrt{\deg(i) \cdot \deg(j)}} \tag{13}$$

where $h_{hete_i}^{(l)}$ represents the heterogeneous node features of node i at the l layer, N_{hete_i} denotes the set heterogeneous neighbor nodes of node i. After obtaining the aggregated features, this study transformed them through a linear layer. The specific process is as follows:

$$\begin{aligned} h_{homo_i}^{(l)*} &= W_{homo} \cdot h_{homo_i}^{(l)} \\ h_{hete_i}^{(l)*} &= W_{hete} \cdot h_{hete_i}^{(l)} \end{aligned} \tag{14}$$

where W_{homo} and W_{hete} are the linear transformation matrices.

After the above processing, PMDGGM employs a bilinear pooling layer to fuse features from different types of nodes. The bilinear pooling layer achieves feature fusion by computing the outer product of two feature vectors and applying a linear transformation. For node i, its associated k heterogeneous nodes are considered for bilinear pooling, and then the average value is taken to obtain the bilinear pooling characteristics of the node. It can capture not only the first-order associations between miRNA and disease nodes, but also higher-order complex associations. The specific process is as follows:

$$\begin{aligned} h_{bilinear_i}^{(l)} &= \frac{1}{\left|N_{hete_i}^k\right|} \sum_{j \in N_{hete_i}^k} \text{Bilinear}(h_{hete_i}^{(l)*}, h_{hete_j}^{(l)*}) \\ &= \frac{1}{\left|N_{hete_i}^k\right|} \sum_{j \in N_{hete_i}^k} \left(h_{hete_i}^{(l)*\mathrm{T}} W h_{hete_j}^{(l)*} + b\right) \end{aligned} \tag{15}$$

where $N_{hete_i}^k$ represents the set of k heterogeneous neighbor nodes of node i, $h_{hete_i}^{(l+1)}$ and $h_{hete_j}^{(l+1)}$ represent the heterogeneous features of node i and node j at the l layer, respectively, W is the learnable weight matrix of the bilinear pooling layer and b is the bias.

To further optimize the feature fusion process, this study introduces a gating mechanism. The core of this mechanism lies in its ability to dynamically adjust the weights of various features during feature fusion. By controlling the flow of information, it enables a more flexible and effective combination of features during information aggregation, the gating mechanism achieves feature fusion through the following formula:

$$\begin{aligned} h_{fused_i}^{(l)} = \sigma(g) \cdot \text{LeakyReLU}(h_{bilinear_i}^{(l)}) + \\ \left(1 - \sigma(g)\right) h_{hete_i}^{(l)} \end{aligned} \tag{16}$$

where σ denotes the sigmoid function, $g \in \mathbb{R}^{d_{out}}$ is the gating weight vector and d_{out} is the dimension of the output feature. The gating mechanism dynamically adjusts the weights of the fused features and the neighbor features, allowing for more flexible feature combinations. This process is reflected in the gating mechanism part of Fig.1.

Finally, the fused features are combined with the features of homogeneous neighbor nodes to generate the final node feature representation.

$$h_{combined_i}^{(l)} = h_{fused_i}^{(l)} + h_{homo_i}^{(l)*} \tag{17}$$

For miRNA and disease nodes, their final features are denoted as h_{miRNA} and $h_{disease}$. Through this feature combination strategy, the method can gather useful information from multiple aspects, integrating the fused features with neighboring features from the homogeneous nodes to form a more complete node representation. This approach also retains more useful information during feature aggregation. Especially when dealing with the heterogeneous features of miRNA and disease nodes, it not only effectively handles the feature differences between different types of nodes but also achieves feature fusion without losing the unique information of each type of node.

3) Predicting miRNA-Disease Associations

For each miRNA-disease pair, the method concatenated their learned feature vectors in the final GCN layer. After that, the method predicts the association probability of each miRNA-disease pair.

$$Z_{ij} = [h_{miRNA_i}, h_{disease_j}] \tag{18}$$

where Z_{ij} is the pair feature vector of miRNA i and disease j. Then, concatenated feature vectors are input into a fully connected layer and sigmoid activation function, which is used to predict the association probability. The method is trained using the binary cross-entropy loss function and applies L2 regularization to the model parameters to prevent overfitting. The loss function is defined as follows:

$$L = -\frac{1}{N_s}\sum_{(i,j)\in\Gamma}\left[y_{ij}\log \mathrm{FC}\left(Z_{ij}\right)+(1-y_{ij})\log\left(1-\mathrm{FC}\left(Z_{ij}\right)\right)\right]+\frac{\lambda}{2}\|\theta\|^2 \quad (19)$$

where y_{ij} denotes the actual association between miRNA and disease, $FC(Z_{ij})$ represents the predicted association probability. Γ is the training set, N_s is the number of training samples, θ is the set of all model parameters, including weight matrices, bias vectors, and the embedding vectors for miRNAs and diseases, and λ controls the strength of the regularization term.

III. Experiments and Results

To achieve more accurate predictions, this study meticulously adjusted various parameters. In Equation (3), while calculating disease similarity, parameters α, β, and γ were all set to 0.2. In Equation (7), for the computation of miRNA integrated similarity, parameter α was set to 0.5, and parameter β was set to 0.25. In Equation (15), when selecting the number of heterogeneous neighbor nodes for the bilinear pooling process, the parameter k is set to 3. Additionally, this study conducted a comprehensive assessment of the model's performance using ablation studies, and comparison experiments to evaluate the predictive performance. The evaluation metrics used include True Positive Rate (TPR), False Positive Rate (FPR), Receiver Operating Characteristic (ROC) curve, Area Under the Curve (AUC), precision, F1-score and recall. In this study, a parameter sensitivity analysis was conducted within the PMDGGM to determine the optimal parameter configuration. The PMDGGM includes the following key parameters: (1) Learning rate: set to 0.0001; (2) Dropout rate: set to 0.5; (3) Learning rate adjustment strategy: after every 400 iterations, the learning rate is adjusted to 0.5 times its original value; (4) Hidden layer dimensions: set to 256 and 128.

A. Ablation Study

PMDGGM is composed of two key modules: the graph attention network and gated convolutional layers. To validate the necessity of each module in PMDGGM and the superiority of the improved model, this study conducted experiments on different configurations, including GAT, GCN, SAGE, GAT+GCN, GAT+SAGE, GCN+SAGE. Among them, GAT+GCN is set to one layer of GAT and one layer of GCN; GAT+SAGE is set to one layer of GAT and one layer of SAGE; GCN+SAGE is a cross setting, with a total of two layers.

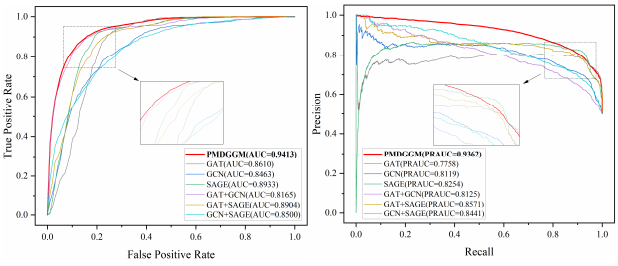

Fig. 2. ROC(a) and PR(b) curves for the prediction results on different combinations of convolutional layers

Based on the analysis of the ablation experiment results (Fig.2), the PMDGGM achieved the highest AUC and PRAUC scores, reaching 0.9413 and 0.9362, respectively, demonstrating optimal performance. This indicates that the various modules of the model and the improvements made have significantly contributed to its overall performance, further highlighting the effectiveness.

B. Compare with Other Methods

To demonstrate the superiority of PMDGGM, this study compared it with several graph neural network related methods, AGAEMD [30], GRPAMDA [31], GIN [32], Mixhop [33], FastRGCN [34], Transformer [35], Cheb [36], and TAG [37]. Among them, AGAEMD is a method based on node-level attention graph autoencoder. GRPAMDA is predicted by combining DropFeature's graph random propagation network and attention network. The experimental results are shown in Fig.3. PMDGGM outperforms the latest prediction techniques and other methods focused on predicting ncRNA-disease associations, exhibiting superior performance across two key evaluation metrics, namely AUC and PRAUC with values of 0.9413 and 0.9362.

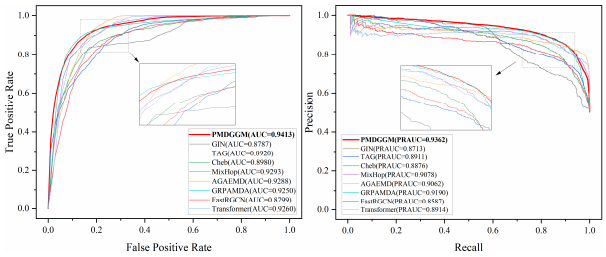

Fig. 3. ROC(a) and PR(b) curves for the prediction results on different graph neural network methods

C. Case Study

To validate the capability of the PMDGGM model in predicting miRNA-disease associations in real-world scenarios, this paper conducted case studies covering multiple diseases. Here, we analyzed the experimental validation situation for two diseases, heart failure and acute myeloid leukemia. For heart failure, this study conducted a search for experimental validation evidence of the miRNAs predicted by PMDGGM. All of the predicted miRNAs have been experimentally confirmed. In Fig.4, the regulatory pathways of three predicted miRNAs (hsa-mir-150 [38], hsa-mir-132 [39], and hsa-mir-182 [40]) are presented through a knowledge graph. Red nodes represent diseases, yellow nodes represent miRNAs, blue nodes represent genes and green nodes represent lncRNAs. These three miRNAs influence heart failure by commonly regulating 9 lncRNAs and 24 genes.

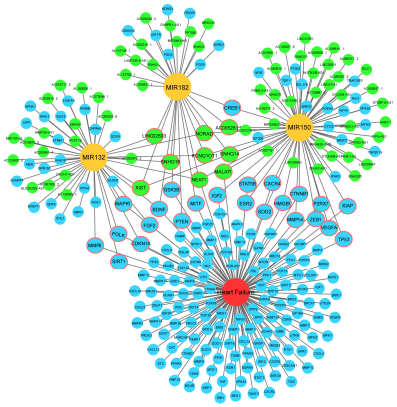

Fig.4. A relationship map of predicted miRNAs of heart failure

In TABLE II, this paper validated the top 20 miRNAs with potential association scores related to acute myeloid leukemia. By querying the PubMed database, we confirmed 19 miRNAs associated with acute myeloid leukemia. Only one remaining miRNA hsa-mir-421, has not been validated.

TABLE II. THE MIRNA PREDICTION RESULTS RELATED TO ACUTE MYELOID LEUKEMIA

miRNA	*Evidence (PMID)*	*miRNA*	*Evidence (PMID)*
hsa-mir-150	30502345	hsa-mir-203a	33812413
hsa-mir-132	30542553	hsa-mir-32	31884339
hsa-mir-182	28079885	hsa-mir-18a	29996811
hsa-mir-23a	30246348	hsa-mir-153	32728112
hsa-mir-19b	28987820	hsa-mir-29b	32423796
hsa-mir-9	28831388	hsa-mir-142	27480083
hsa-mir-20b	32380791	hsa-mir-15a	34068078
hsa-mir-143	37149661	hsa-mir-106b	27351222
hsa-mir-195	34368922	hsa-mir-449a	32059753
hsa-mir-421	Not Confirmed	hsa-mir-193a	31837329

IV. CONCLUSION AND DISCUSSION

This study developed a predictive model named PMDGGM, which integrates graph attention layer and gated convolutional layer. Initially, this study constructs a comprehensive miRNA-disease heterogeneous network. Subsequently, PMDGGM leverages bilinear pooling layers to capture the complex relationships between heterogeneous nodes and employs a gating mechanism to effectively integrate multi-source heterogeneous features, which allows for more accurate prediction of molecular interactions. PMDGGM demonstrates excellent performance across benchmark datasets, accurately predicting novel miRNA-disease associations.

REFERENCES

[1] Y. Ma, N. Shen, M. S. Wicha, and M. Luo, "The Roles of the Let-7 Family of MicroRNAs in the Regulation of Cancer Stemness," *Cells,* vol. 10, no. 9, p. 2415, 2021, doi: 10.3390/cells10092415.

[2] E. G. Moss, R. C. Lee, and V. Ambros, "The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA," *Cell,* vol. 88, no. 5, pp. 637-646, 1997, doi: 10.1016/s0092-8674(00)81906-6.

[3] J. A. Chan, A. M. Krichevsky, and K. S. Kosik, "MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells," *Cancer research,* vol. 65, no. 14, pp. 6029-6033, 2005, doi: 10.1158/0008-5472.CAN-05-0137.

[4] I. Asangani, "Rasheed sA, Nikolova DA, Leupold JH, Colburn NH, Post s and Allgayer H: microRNA-21 (miR-21) post-transcriptionally downregulates tumor suppressor Pdcd4 and stimulates invasion, intravasation and metastasis in colorectal cancer," *Oncogene,* vol. 27, pp. 2128-2136, 2008, doi: 10.1038/sj.onc.1210856.

[5] X. Chen and G.-Y. Yan, "Semi-supervised learning for potential human microRNA-disease associations inference," *Scientific reports,* vol. 4, no. 1, p. 5501, 2014, doi: 10.1038/srep05501.

[6] L. Wang *et al.*, "LMTRDA: Using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities," *PLoS computational biology,* vol. 15, no. 3, p. e1006865, 2019, doi: 10.1371/journal.pcbi.1006865.

[7] W. Cao *et al.*, "Metapath-aggregated multilevel graph embedding for miRNA–disease association prediction," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Istanbul, Turkey, Dec 5-8 2023, pp. 468-473, doi: 10.1109/bibm58861.2023.10385762.

[8] H. Zhao, Z. Li, Z. H. You, R. Nie, and T. Zhong, "Predicting Mirna-Disease Associations Based on Neighbor Selection Graph Attention Networks," *IEEE/ACM Trans Comput Biol Bioinform,* vol. 20, no. 2, pp. 1298-1307, 2023, doi: 10.1109/TCBB.2022.3204726.

[9] Y. Zhang, X. Lei, Z. Fang, and Y. Pan, "CircRNA-disease associations prediction based on metapath2vec++ and matrix factorization," *Big Data Mining and Analytics,* vol. 3, no. 4, pp. 280-291, 2020, doi: 10.26599/BDMA.2020.9020025.

[10] X. Lei, W. zhang, and L. Liu, "Prediction of circRNA-disease associations based on multiple biological data (in Chinese)," *SCIENTIA SINICA Informationis,* vol. 51, no. 6, pp. 927-939, 2021, doi: 10.1093/bib/bbac388.

[11] C. Cui, B. Zhong, R. Fan, and Q. Cui, "HMDD v4.0: a database for experimentally supported human microRNA-disease associations," *Nucleic Acids Res,* vol. 52, no. D1, pp. D1327-D1332, 2024, doi: 10.1093/nar/gkad717.

[12] L. M. Schriml *et al.*, "Disease Ontology: a backbone for disease semantic integration," *Nucleic acids research,* vol. 40, no. D1, pp. D940-D946, 2012, doi: 10.1093/nar/gkr972.

[13] J. H. Li, S. Liu, H. Zhou, L. H. Qu, and J. H. Yang, "starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data," *Nucleic Acids Res,* vol. 42, no. Database issue, pp. D92-7, 2014, doi: 10.1093/nar/gkt1248.

[14] S.-D. Hsu *et al.*, "miRTarBase: a database curates experimentally validated microRNA–target interactions," *Nucleic Acids Research,* vol. 39, no. suppl_1, pp. D163-D169, 2011, doi: 10.1093/nar/gkq1107.

[15] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, "miRBase: from microRNA sequences to function," *Nucleic Acids Res,* vol. 47, no. D1, pp. D155-D162, 2019, doi: 10.1093/nar/gky1141.

[16] G. Chen *et al.*, "LncRNADisease: a database for long-non-coding RNA-associated diseases," *Nucleic Acids Res,* vol. 41, no. Database issue, pp. D983-6, 2013, doi: 10.1093/nar/gky905.

[17] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, "OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders," *Nucleic Acids Res,* vol. 43, no. Database issue, pp. D789-798, 2015, doi: 10.1093/nar/gku1205.

[18] J. Pinero *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, doi: 10.1093/nar/gkw943.

[19] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics,* vol. 26, no. 13, pp. 1644-1650, Jul 1 2010, doi: 10.1093/bioinformatics/btq241.

[20] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson Correlation Coefficient," in *Noise Reduction in Speech Processing*, (Springer Topics in Signal Processing, 2009, ch. Chapter 5, pp. 1-4.

[21] F. Rahutomo, T. Kitasuka, and M. Aritsugi, "Semantic cosine similarity," in *The 7th international student conference on advanced science and technology ICAST*, New York, NY, USA, 2012, vol. 4, no. 1, pp. 1-2.

[22] S. Bag, S. K. Kumar, and M. K. Tiwari, "An efficient recommendation generation using relevant Jaccard similarity," *Information Sciences,* vol. 483, pp. 53-64, 2019, doi: 10.1016/j.ins.2019.01.023.

[23] P. Sedgwick, "Spearman's rank correlation coefficient," *Bmj,* vol. 349, p. g7327, 2014, doi: 10.1136/bmj.g7327

[24] K. W. Church, "Word2Vec," *Natural Language Engineering,* vol. 23, no. 1, pp. 155-162, 2017, doi: 10.1017/S1351324916000334.

[25] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, Los Angeles, California, USA, 2003, vol. 242, no. 1, pp. 29-48.

[26] K. Baker, "Singular value decomposition tutorial," *The Ohio State University,* vol. 24, p. 22, 2005, doi: 10.1037/met0000105.

[27] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *The Sixth International Conference on Learning Representations*, Vancouver, Canada, Apr. 30-May. 3 2018.

[28] H. Tong, C. Faloutsos, and J.-Y. Pan, "Fast random walk with restart and its applications," in *Sixth international conference on data mining (ICDM'06)*, Hong Kong, China, 2006, pp. 613-622.

[29] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in neural information processing systems*, Long Beach, CA, USA, Dec. 4-9 2017.

[30] H. Zhang, J. Fang, Y. Sun, G. Xie, Z. Lin, and G. Gu, "Predicting miRNA-Disease Associations via Node-Level Attention Graph Auto-Encoder," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 20, no. 2, pp. 1308-1318, 2023, doi: 10.1109/TCBB.2022.3170843.

[31] T. Zhong, Z. Li, Z. H. You, R. Nie, and H. Zhao, "Predicting miRNA-disease associations based on graph random propagation network and attention network," *Brief Bioinform,* vol. 23, no. 2, Mar 10 2022, Art no. bbab589, doi: 10.1093/bib/bbab589.

[32] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, " How Powerful are Graph Neural Networks?," in *The 7th International Conference on Learning Representations*, New Orleans, LA, USA, May 6-9 2019.

[33] S. Abu-El-Haija *et al.*, "Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing," in *International conference on machine learning*, Long Beach, California, USA, 2019, pp. 21-29.

[34] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *The semantic web: 15th international conference (ESWC 2018)*, Heraklion, Crete, Greece, , June 3–7 2018: Springer, pp. 593-607.

[35] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, and Y. J. a. p. a. Sun, "Masked label prediction: Unified message passing model for semi-supervised classification," 2020, doi: 10.48550/arXiv.2009.03509.

[36] M. Defferrard, X. Bresson, and P. J. A. i. n. i. p. s. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *30th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain., Dec 5-10 2016, vol. 29.

[37] J. Du, S. Zhang, G. Wu, J. M. Moura, and S. J. a. p. a. Kar, "Topology adaptive graph convolutional networks," 2017, doi: 10.48550/arXiv.1710.10370

[38] D. Scrutinio, F. Conserva, P. Guida, A. J. M. C. Passantino, and Angiology, "Long-term prognostic potential of microRNA-150-5p in optimally treated heart failure patients with reduced ejection fraction: a pilot study," vol. 70, no. 4, pp. 439-446, 2020, doi: 10.23736/S2724-5683.20.05366-9.

[39] X. Liu, Z. Tong, K. Chen, X. Hu, H. Jin, and M. J. B. R. I. Hou, "The Role of miRNA - 132 against apoptosis and oxidative stress in heart failure," vol. 2018, no. 1, p. 3452748, 2018, doi: 10.1155/2018/3452748.

[40] F. Fang, X. Zhang, B. Li, and S. J. J. o. C. S. Gan, "miR-182-5p combined with brain-derived neurotrophic factor assists the diagnosis of chronic heart failure and predicts a poor prognosis," vol. 17, no. 1, p. 88, 2022, doi: 10.1186/s13019-022-01802-0.