

Article

HGSMDA: miRNA–Disease Association Prediction Based on HyperGCN and Sørensen-Dice Loss

Zhenghua Chang, Rong Zhu *, Jinxing Liu, Junliang Shang  and Lingyun Dai

School of Computer Science, Qufu Normal University, Rizhao 276826, China; changzhenghua313@126.com (Z.C.); sdcavell@126.com (J.L.); shangjunliang110@163.com (J.S.); dailinyun_1@163.com (L.D.)

* Correspondence: zhurongsd@qfnu.edu.cn

Abstract: Biological research has demonstrated the significance of identifying miRNA–disease associations in the context of disease prevention, diagnosis, and treatment. However, the utilization of experimental approaches involving biological subjects to infer these associations is both costly and inefficient. Consequently, there is a pressing need to devise novel approaches that offer enhanced accuracy and effectiveness. Presently, the predominant methods employed for predicting disease associations rely on Graph Convolutional Network (GCN) techniques. However, the Graph Convolutional Network algorithm, which is locally aggregated, solely incorporates information from the immediate neighboring nodes of a given node at each layer. Consequently, GCN cannot simultaneously aggregate information from multiple nodes. This constraint significantly impacts the predictive efficacy of the model. To tackle this problem, we propose a novel approach, based on HyperGCN and Sørensen-Dice loss (HGSMDA), for predicting associations between miRNAs and diseases. In the initial phase, we developed multiple networks to represent the similarity between miRNAs and diseases and employed GCNs to extract information from diverse perspectives. Subsequently, we draw into HyperGCN to construct a miRNA–disease heteromorphic hypergraph using hypernodes and train GCN on the graph to aggregate information. Finally, we utilized the Sørensen-Dice loss function to evaluate the degree of similarity between the predicted outcomes and the ground truth values, thereby enabling the prediction of associations between miRNAs and diseases. In order to assess the soundness of our methodology, an extensive series of experiments was conducted employing the Human MicroRNA Disease Database (HMDD v3.2) as the dataset. The experimental outcomes unequivocally indicate that HGSMDA exhibits remarkable efficacy when compared to alternative methodologies. Furthermore, the predictive capacity of HGSMDA was corroborated through a case study focused on colon cancer. These findings strongly imply that HGSMDA represents a dependable and valid framework, thereby offering a novel avenue for investigating the intricate association between miRNAs and diseases.



Citation: Chang, Z.; Zhu, R.; Liu, J.; Shang, J.; Dai, L. HGSMDA: miRNA–Disease Association Prediction Based on HyperGCN and Sørensen-Dice Loss. *Non-Coding RNA* **2024**, *10*, 9. <https://doi.org/10.3390/ncrna10010009>

Academic Editor: Sarah Rennie, Shizuka Uchida

Received: 15 December 2023

Revised: 19 January 2024

Accepted: 24 January 2024

Published: 26 January 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: miRNA–disease association; HyperGCN; Sørensen-Dice loss; prediction

1. Introduction

MiRNAs (microRNAs) belong to a class of endogenous non-coding single-stranded RNAs. They play a critical role in regulating biological phenomena through mechanisms such as cleavage or translational repression [1]. The dysregulation of miRNA expression can lead to alterations in the levels of gene expression for their target genes, thereby playing a vital role in the initiation and progression of specific illnesses. Hence, uncovering the correlation between miRNAs and diseases is of utmost significance for medical researchers, as it enables them to enhance their comprehension of the intricate pathological mechanisms underlying various diseases. Given that traditional biological experiments to infer the correlation between miRNAs and diseases are costly and take up a significant amount of time, computer-based methods have emerged as a viable and efficient strategy to tackle this challenging task. In recent years, computational methods commonly employed for

this purpose can be categorized into two groups: similarity-based predictive models and machine learning-based predictive models.

Fundamental predictive models, which rely on the assumption of similarity, propose that miRNAs sharing functional similarities are linked to diseases exhibiting comparable phenotypes, and the reverse is also true. These models are based on the principle that entities with similar characteristics or traits are inclined to have related functions or associations. In similarity-based prediction models, it is crucial to acquire similarity scores between various biomedical entities. Chen et al. [2] came up with a Randomized Wandering and Restart-based miRNA–disease Association (RWRMDA) method, which first constructs a functional similarity network of miRNAs associated with diseases. Zhong et al. [3] introduced a new method utilizing non-negative matrix decomposition to rank candidate disease miRNAs for prediction. To fully incorporate similarity information into miRNA–disease heteromorphic networks, MDHGI [4] developed a miRNA–disease heteromorphic map using matrix decomposition techniques. In the study by BNPMMDA [5], a dichotomous network recommendation algorithm was used to represent the relationship between miRNAs and diseases as a dichotomous network, and the topology of this network and the attribute information of the nodes were utilized to predict possible correlations between miRNAs and diseases. In order to accomplish this, transfer weights were assigned to establish connections between resource allocation and miRNAs and diseases, while also considering bias ratings. These means have yielded excellent results in predicting miRNA–disease correlations, but these similarity-based models tend to be overly reliant on known associations between miRNAs and diseases.

Machine learning-based predictive models are also commonly utilized to discover associations between miRNAs and diseases according to specific criteria. Feature extraction is a vital stage in most machine learning models, as it significantly influences the forecast outcomes of the classifier. Xu et al. [6] proposed a means referred to as MTDN that inputs characteristics derived from the network into an SVM classification model and then feeds the features back into a predictive model. IMCMDA [7] is a new inductive matrix-completion approach that utilizes known associations and integrated miRNA similarities and disease similarities to complete the association of missing miRNAs with diseases. HDMP [8] predicts built upon weighting the k most similar neighbors, combining the informational value of disease terms and the phenotypic similarity between diseases. ABMDA [9] enhances the precision of a given learning algorithm by incorporating weak classifiers capable of scoring samples. Shang et al. [10] proposed a weighted bi-level network based miRNA–disease association (BLNIMDA) method, which calculates association scores based on bidirectional information distribution strategies and association types to ensure comprehensive and accurate predictions. Machine learning-based models, greatly improve efficiency, but these models require high quality of features, such as miRNA expression levels, biological functions, biological pathways involved, and network properties of miRNAs associated with known diseases.

In the past few years, graph convolutional networks (GCNs) have demonstrated strong performance capabilities in learning graphical representations. The method has garnered significant attention and has found extensive application in fundamental research areas, for example the prediction of drug–target interactions [11–13] and tumor-associated gene prediction [14]. Also, given the excellent performance of graph convolutional networks in link prediction tasks, in recent years it has begun to be applied to inferring links between miRNAs and diseases. NIMCGCN [15] employs GCN, which fully extracts the nonlinear characteristics of miRNAs and diseases and predicts associations by achieving feature enhancement through multichannel attention. MMGCN [16] utilizes a GCN encoder to extract features. It further employs a multi-channel attention strategy to assign weights to different features for predicting associations. GCNA-MDA [17] utilizes an autoencoder to extract representations, while GCN is used to capture topological information to forecast miRNA–disease correlations. GCN-based models have the capability to aggregate more comprehensive node information, thereby enhancing the predictive ability of association.

However, GCN is a locally aggregated graph convolution algorithm that can only consider the information of a node's first-order neighbors at each layer and cannot aggregate multiple nodes. If the number of layers of the GCN is increased to handle information from multiple nodes, then nodes in the GCN may begin to receive information from most or all other nodes in the graph. This may lead to an oversmoothing problem, where the feature representations of different nodes tend to be similar, making it difficult for the network to capture differences between nodes. Additionally, GCNs usually require the structure and feature information of the entire graph for forward propagation. This means that for very large graphs, GCN may be difficult to scale as it requires a lot of memory and computational resources.

This paper introduces a novel approach, named HGSMMA, for forecasting miRNA–disease associations. HGSMMA is built upon the foundation of HyperGCN and utilizes the Sørensen-Dice loss function. The contributions of the method are as follows:

- Information about miRNAs, information about diseases, and information about known miRNA–disease associations were integrated. By integrating this information into the HGSMMA model, we were able to more fully characterize the relationship between miRNAs and disease.
- HyperGCN was introduced to construct a miRNA–disease heterogeneous hypergraph using hypernodes, and GCNs were trained on the graph to aggregate information.
- The Sørensen-Dice loss function is employed to evaluate the likeness between the predicted outcomes and the actual values. This facilitates a more precise evaluation of the model's capability.

2. Data and Experiments

2.1. Datasets

The miRNA–disease association data utilized in the present study were sourced from the HMDD v3.2 [18]. This database is one of the largest human microRNA disease databases in the world and collects human microRNA sequencing data from all over the world. A total of 12,446 associations between 853 miRNAs and 591 diseases were carefully selected for this study [19–21]. These associations have been verified through experimentation and are considered reliable. Next, we labeled the identified miRNA–disease correlations as positive samples, indicated by the number 1. For other samples, the absence of correlations between miRNA and disease is indicated by the number 0. Predicting associations between miRNAs and diseases is dependent on the hypothesis that miRNAs exhibiting analogous functions are more inclined to be linked with diseases that possess analogous characteristics. Therefore, understanding the relationships between disease–disease and miRNA–miRNA becomes crucial in this context. In this research investigation, we employed a comprehensive approach that considered multiple factors to determine the resemblance between miRNA–miRNA and disease–disease. These factors include similarity in miRNA function, resemblance in miRNA sequence, semantic resemblance in disease, target-based disease resemblance, and nuclear resemblance in the Gaussian interaction spectrum between miRNA and disease.

2.2. Parametric Analysis

In our training process, we established the batch size as 128, the dimension of the node features was established as 128, and the number of hypernodes was established as 64. In addition, to mitigate the issue of overfitting, we set the dropout to 0.5 to randomly ignore certain neurons. We used several assessment metrics to provide a comprehensive assessment of the forecast results, including the area under the curve (AUC) for receiver operating characteristics, the area under the precision/recall curve (AUPRC), precision, F1-score, and accuracy. To ensure a reliable evaluation of the predicted outcomes, we performed multiple replications of all experiments. This ensures the stability and dependability of the assessment results. The model's predictive performance is influenced by hyperparameters. Hence, we employed Cross-validation with a 5-fold approach to analyze

the effect of certain parameters on the predicted outcomes and to conduct a comparison of the evaluation metrics values.

2.2.1. Dropout Parameter Settings

Dropout is critical for tuning deep learning models. During training, dropout uses a randomized approach to zero out the output of a portion of neurons. This reduces the model's dependence on specific neurons and thus reduces the risk of overfitting. Figure 1 illustrates the evaluation metric scores for different dropouts. For optimal performance, we chose a suitable dropout of 0.5 after tuning.

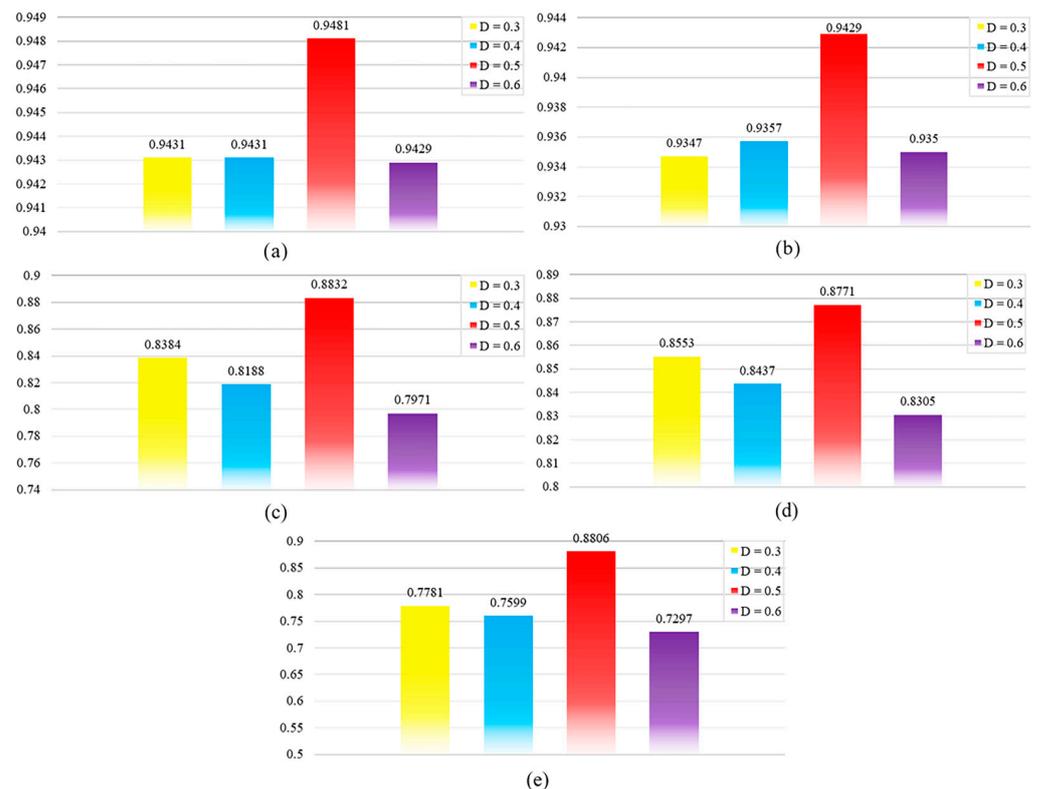


Figure 1. Parameter analysis of different dropouts. (a) AUC analysis for different dropout; (b) AUPRC analysis for different dropout; (c) Accuracy analysis for different drop-out; (d) F1-score analysis for different dropout; (e) Precision analysis for different dropout.

2.2.2. Feature Embedding Dimension

The dimension of the learnable parameter matrix associated with a feature is influenced by its embedding size. The evaluation metric scores for different feature dimensions, specifically {32, 64, 128, 256}, are depicted in Figure 2. We can see that there are significant differences in precision between different feature dimensions. We chose a suitable feature dimension of 128 for optimal performance after tuning.

2.2.3. The Number of Hypernodes

The structure of the hypergraph is influenced by the number of hypernodes. Hence, we conducted further analysis to examine the impact of the quantity of hypernodes on the prediction outcomes. Figure 3 displays the evaluation metric scores for different numbers of hypernodes, namely {8, 16, 32, 64, 128}. It is important to note that having too many hypernodes can hinder the extraction of useful information and potentially complicate the graph structure due to the random initialization of hypernode representations. Consequently, we have determined that the optimal number of hypernodes is 64.

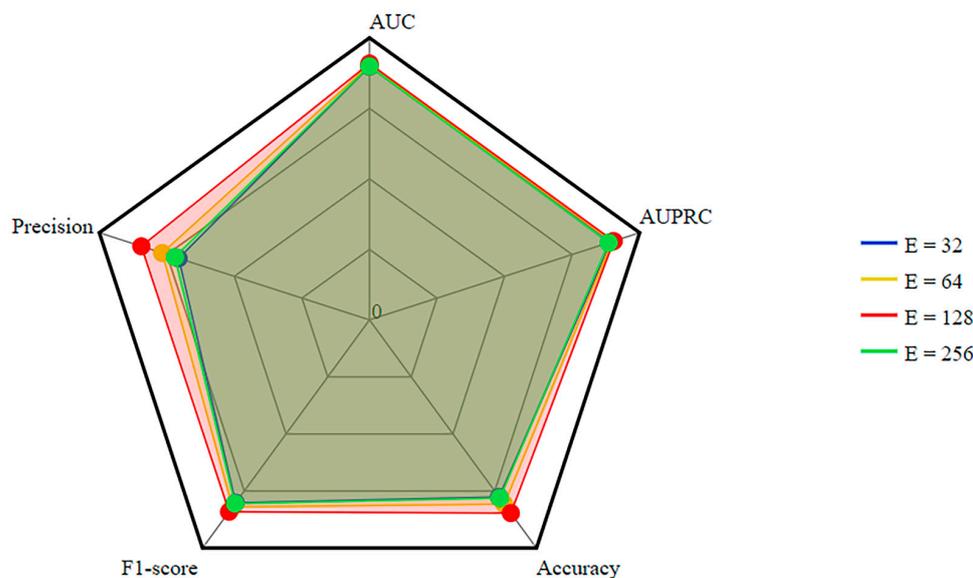


Figure 2. Parameter analysis for different feature embedding dimensions.

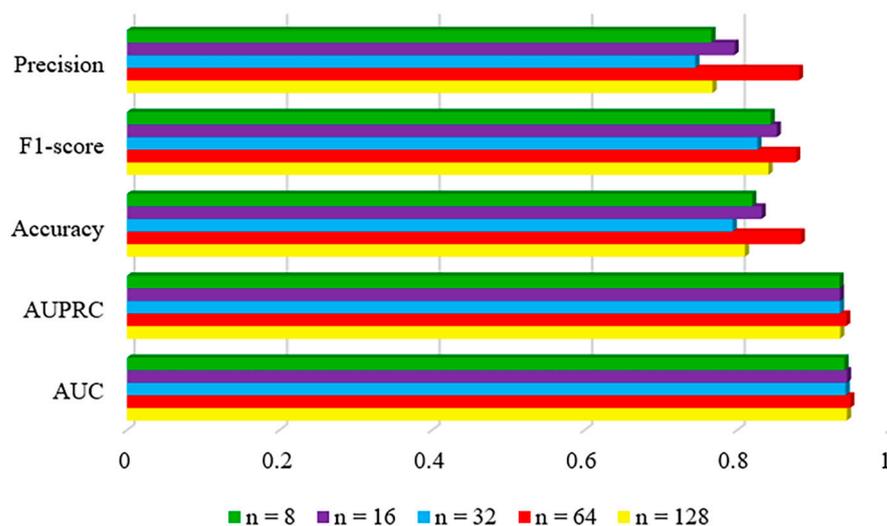


Figure 3. Parameter analysis for different numbers of hypernodes.

3. Results and Discussion

3.1. Comparison with Other Methods

We conducted a comparison between our model and eight other methods on HMDD v3.2, namely NIMCGCN [15], MMGCN [16], ERMDA [22], HGANMDA [23], AGAEMD [24], MINIMDA [25], MAGCN [26] and AMHMDA [27]. For all comparison methods, the initial similarity data used is the one already available for each respective method. Furthermore, we maintained the settings, parameters, and the portion that dynamically acquires specific similarities for the other methods.

1. NIMCGCN [15] uses GCN to learn miRNA and disease potential characteristic representations. It inputs the learned characteristics into the NIMC model to produce a matrix of association complements to forecast miRNA–disease associations.
2. MMGCN [16] uses a GCN encoder to obtain characteristics under different similarity views separately and forecasts miRNA–disease associations by utilizing multi-channel attention to enhance the learned potential representation of association prediction.
3. ERMDA [22] proposed a resampling strategy to construct multiple subsets and applied feature selection methods to increase the diversity among these subsets. It then uses soft voting to forecast the connections of miRNAs with diseases.

4. HGANMDA [23] constructs a heterogeneous graph, applies node-level attention to learn neighboring nodes, applies semantic-level attention to learn meta-paths, and lastly employs a bilinear decoder to reconstruct miRNA–disease associations.
5. AGAEMD [24] creates heterogeneous matrices and uses autoencoders in miRNA–disease networks to polymerize information and reconstruct miRNA–disease association networks.
6. MINIMDA [25] constructs disease similarity networks to obtain embedding representations by mixing higher-order neighborhood information, which is fed into a multilayer perceptron (MLP) to forecast potential connections between miRNAs and diseases.
7. MAGCN [26] used lncRNA–miRNA interactions to predict novel miRNA–disease correlations through graph convolutional networks with attentional mechanisms and convolutional neural network combiners.
8. AMHMDA [27] constructed multiple similarity networks, introduced virtual nodes to construct heterogeneous hypergraphs, and used the output of graph convolutional networks to predict associations.

Figure 4 displays the results of multiple experiments for comparative analysis. HGSMMDA had an AUC of 94.81%, AUPRC of 94.29%, ACC of 88.32%, and a recall of 87.36%. HGSMMDA shows good performance on HMDD v3.2 compared to alternative approaches.

3.2. Ablation Experiments

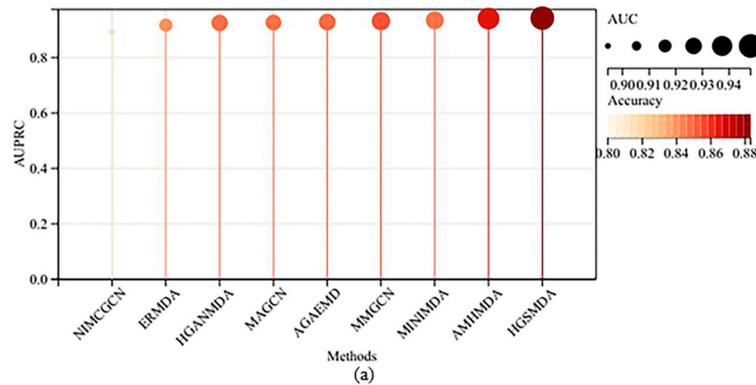
To evaluate the significance of each module in HGSMMDA, we created two different versions of the model, namely HGSMMDA-L and HGSMMDA-H, for comparative analysis. Specifically, to assess the impact of the Sørensen-Dice loss function on model performance, HGSMMDA-L replaces the Sørensen-Dice loss function in the original modeling framework with the BCE loss function frequently used in classification tasks. To assess the impact of HyperGCN on model performance, HGSMMDA-H replaces the miRNA–disease heteromorphic hypergraph with a miRNA–disease heteromorphic graph, and the model loses the ability to model hyperedges, thus no longer constituting HyperGCN.

Figure 5 shows a comparison of the evaluation metrics for HGSMMDA-L, HGSMMDA-H, and HGSMMDA. We can find that the loss function Sørensen-Dice has a positive effect on the model predictions. This shows that the Sørensen-Dice loss function improves the performance of the model by operating on the intersection and union of the predictions with the true labels. The introduction of HyperGCN contributes to the prediction results of the model. This suggests that by utilizing hyperedges to capture higher-order relationships between nodes, HyperGCN is better able to capture the correlation and similarity between nodes, thus improving the performance of the model.

3.3. Case Study

Colon cancer, a malignant neoplasm affecting the gastrointestinal tract within the colon region, can be effectively detected at an early stage through the utilization of colonoscopy [28]. The global incidence of cancer has been progressively rising in recent times [29]. Consequently, the identification of prognostic and predictive biomarkers linked to colon cancer holds immense importance in the realm of colon cancer treatment. The investigation of miRNAs has garnered significant attention due to their involvement in colon cancer cell proliferation and resistance to therapeutic agents. For example, reduced let-7a expression is closely associated with colon carcinogenesis [30]. High expression of the gene encoding mir-21 is associated with low expression of the gene encoding the tumor suppressor protein PDCD4 [31]. The known miRNA–disease associations in the HMDD v3.2 dataset were employed as a training dataset. Subsequently, the HGSMMDA model was employed to predict potential miRNAs for colon cancer. Following this, we screened the top 10 predicted miRNAs and validated these predicted associations based on the authoritative miRNA–disease association dataset dbDEMC [32]. The results are shown in Figure 6, and the obtained results were validated in the database. This comprehensive

demonstration underscores the effectiveness and dependability of HGSM DA in accurately predicting miRNA–disease associations.



Methods	AUC	AUPRC	Accuracy	F1-score	Recall	Precision
NIMCGCN	0.8945	0.8926	0.8131	0.8148	0.8220	0.8076
MMGCN	0.9321	0.9320	0.8550	0.8544	0.8512	0.8580
ERMDA	0.9171	0.9173	0.8399	0.8396	0.8382	0.8412
HGANMDA	0.9265	0.9253	0.8489	0.8481	0.8433	0.8529
AGAEMD	0.9270	0.9286	0.8502	0.8507	0.8544	0.8481
MINIMDA	0.9304	0.9350	0.8481	0.8482	0.8529	0.8505
MAGCN	0.9245	0.9268	0.8483	0.8473	0.8425	0.8533
AMHMDA	0.9422	0.9411	0.8669	0.8653	0.8549	0.8763
HGSM DA	0.9481	0.9429	0.8832	0.8771	0.8736	0.8806

Note: The best scores are in bold.

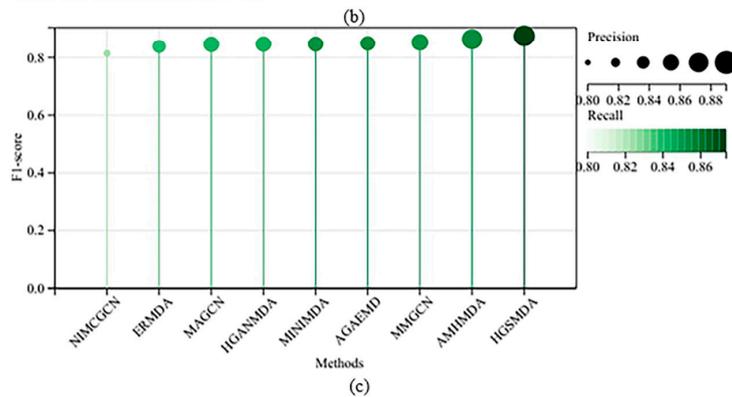


Figure 4. Comparison with other methods on HMDD v3.2. (a) AUC, AUPRC and Accuracy analysis for different method; (b) Detailed data on individual methods; (c) F1-score, Recall and Precision analysis for different method.

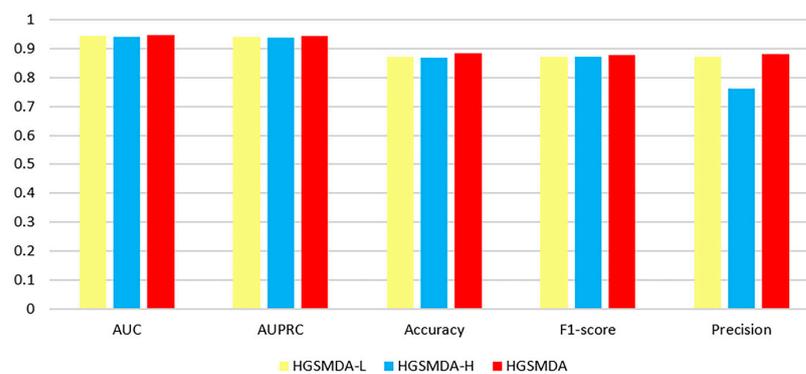


Figure 5. Results of ablation experiments for three different models.

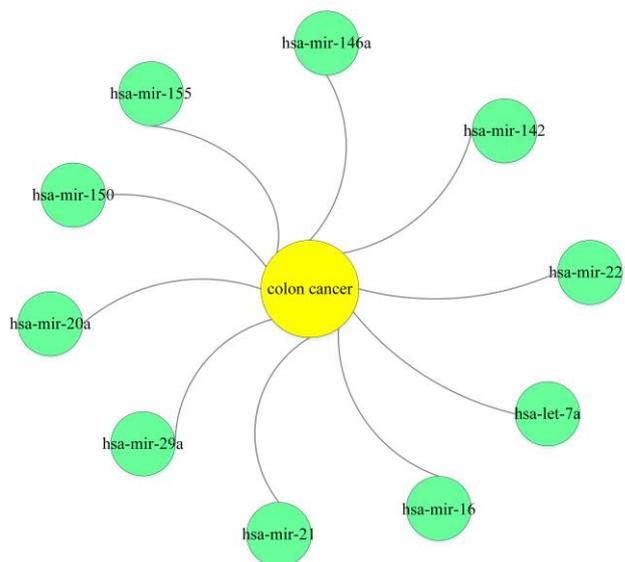


Figure 6. Prediction of the top 10 relevant miRNAs for predicting colon cancer.

4. Methodology

4.1. HGSM DA Framework

Inspired by the AMHMDA [27] research, we introduced a new approach called HGSM DA to predict the association of miRNAs with diseases. HGSM DA is built upon the foundation of HyperGCN and utilizes the Sørensen-Dice loss function, as described in Figure 7.

The HGSM DA process is comprised of three primary stages:

1. Extracting features: We constructed multiple miRNA and disease similarity networks and used GCN for extracting information from different perspectives.
2. HyperGCN: We introduce HyperGCN to construct a miRNA–disease heteromorphic hypergraph using hypernodes, and train GCN on the graph to aggregate information.
3. Measuring the degree of similarity: We leverage the attention mechanism to fuse the output of the HyperGCN layer in combination with a CNN (Convolutional Neural Network) for classification. We then use the Sørensen-Dice loss function to scale the degree of similarity between the predictions and the true values.

4.2. Extraction of Features

To investigate similarity data that aids in making association predictions, we constructed three different miRNA–miRNA networks and three different disease–disease networks. The collections of adjacency matrix for miRNAs and diseases are denoted as stated below:

$$B_m = \{ B_{mf} [H_{mf}], B_{ms} [H_{ms}], B_{mg} [H_{mg}] \} \tag{1}$$

$$B_d = \{ B_{ds} [H_{ds}], B_{dt} [H_{dt}], B_{dg} [H_{dg}] \}, \tag{2}$$

where B_{mf} denotes the adjacency matrix of miRNA functional similarity matrices, B_{ms} denotes the adjacency matrix of miRNA sequence similarity matrices, and B_{mg} denotes the adjacency matrix of Gaussian interaction profile kernel similarity matrices. Similarly, where B_{ds} denotes the adjacency matrix of the disease semantic similarity matrices, B_{dt} denotes the adjacency matrix of target-based disease similarity matrices, and B_{dg} denotes the adjacency matrix of Gaussian interaction profile kernel similarity matrices.

After constructing the interaction network, the GCN extracts the information in the miRNA and disease networks by iteratively propagating the information of the nodes and updating the representation of the nodes. The initial embedding of the GCN is achieved by generating random feature vectors with specified dimensions for each node. On this

basis, the adjacency matrix of miRNA–miRNA and disease–disease interaction network was normalized:

$$J_m = \tilde{P}_m^{-\frac{1}{2}} \tilde{B}_m \tilde{P}_m^{-\frac{1}{2}} \tag{3}$$

$$J_d = \tilde{P}_d^{-\frac{1}{2}} \tilde{B}_d \tilde{P}_d^{-\frac{1}{2}}, \tag{4}$$

where \tilde{B} denotes the adjacency matrix with the unit matrix added, and \tilde{P} is the degree matrix of \tilde{B} .

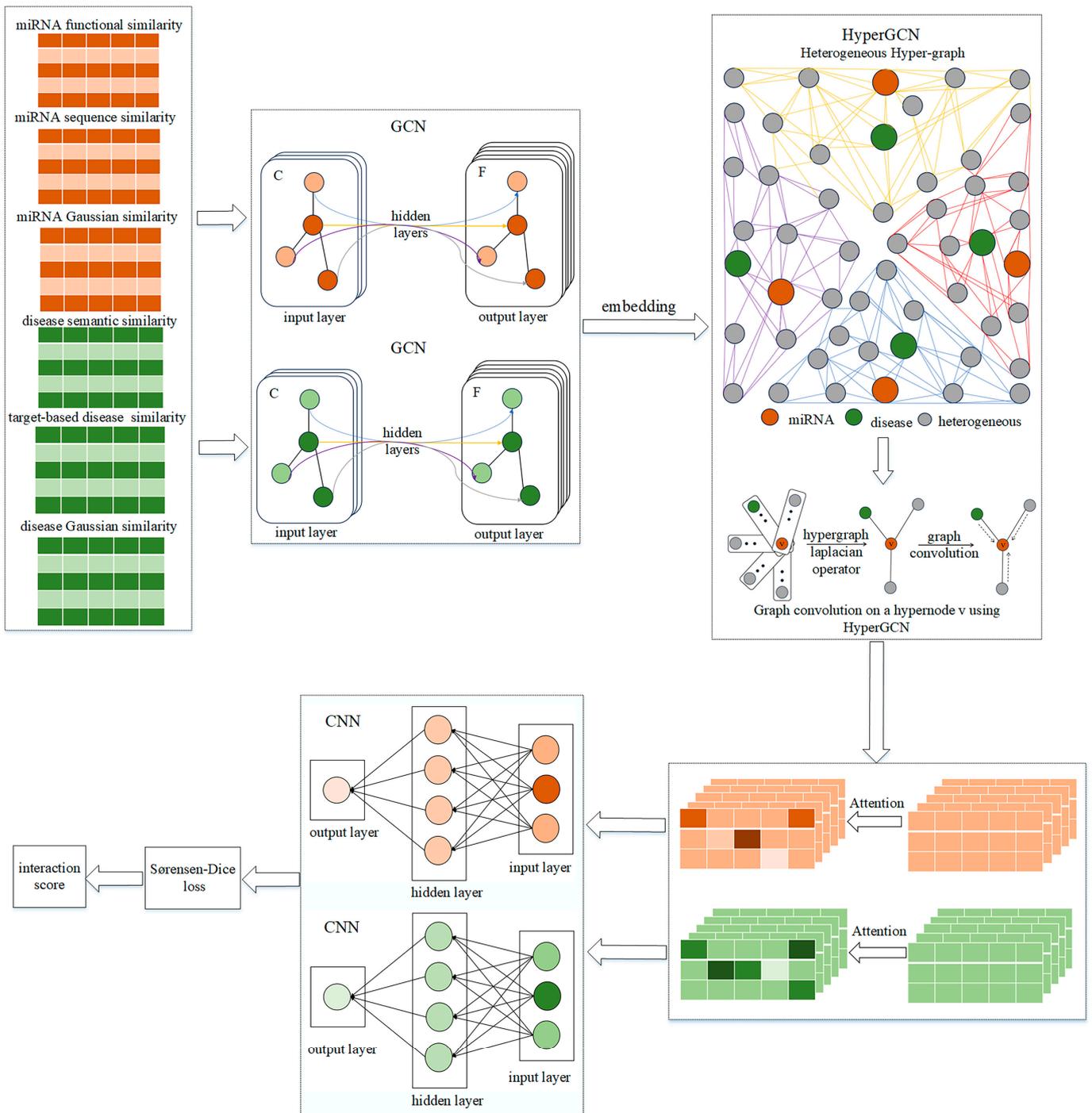


Figure 7. Architecture of the HGSM DA.

4.3. HyperGCN

After obtaining an embedded representation that aggregates various similarity information of miRNAs and diseases, we construct a hypergraph of miRNAs and diseases through hypernodes. Within the hypergraph, hypernodes can connect to multiple nodes or other hypernodes. These hypernodes are the connection source of all miRNA nodes and disease nodes, which provides a new idea for miRNA–disease node research. The introduction of hypernodes allows us to incorporate unknown miRNA–disease associations into the hypergraph, thus exploring potential miRNA–disease associations more comprehensively. Furthermore, hypernodes have the ability to autonomously acquire node representations within the network, thus aggregating richer information in the hypergraph and improving the quality of connections between miRNAs and diseases. This approach avoids the use of formulas to directly calculate the association score, but instead computes the cosine similarity to obtain the adjacency matrix of the hypergraph, which better describes the association relationship between miRNAs and diseases, and can be described as:

$$C_{M_i, v_k} = \frac{M_i \cdot v_k}{\|M_i\|_2 \|v_k\|_2} \quad (5)$$

$$C_{D_j, v_k} = \frac{D_j \cdot v_k}{\|D_j\|_2 \|v_k\|_2}, \quad (6)$$

where M_i represents the embedding of a previously obtained miRNA node i and D_j represents the embedding of a previously obtained disease node j . v is a hypernode whose node characteristics are randomly initialized. Thus, for miRNA i –disease j pairs, the correlation between them can be quantified using variables C_{M_i, v_k} and C_{D_j, v_k} .

We performed GCN on the graph. In the neural information propagation framework, we denote the information update of hypernode v as:

$$Z_v^{(\tau+1)} = \sigma \left(\left(\Theta^{(\tau)} \right)^T \sum_{u \in N(v)} \left(\left[\bar{A}_S^{(\tau)} \right]_{v,u} \cdot h_u^{(\tau)} \right) \right), \quad (7)$$

where τ is the epoch, $Z_v^{(\tau+1)}$ is the new hidden layer node representation of the node v , and $N(v)$ is the neighborhood of v , $\left[\bar{A}_S^{(\tau)} \right]_{v,u}$ is the weight of the regularized edge $\{v, u\}$, and $h_u^{(\tau)}$ is the implicit representation of the neighboring nodes in the previous stage. We employ the conventional graph convolution operation on the variable v , taking into consideration only the simple edges that are incident to it. For each node, its features are weighted and averaged according to the features of its neighboring nodes to update the node's representation. We perform operations on each super node $v \in V$ during each training period τ until convergence is achieved.

4.4. Measuring Similarity

We believe that the information of nodes in different layers contributes differently to the prediction results, after acquiring the representations of nodes from various layers, we utilize the layer-level attention mechanism to acquire the node representations at different layers, each assigned with varying degrees of significance. The layer-level attention mechanism assigns different importance to node representations in different layers by calculating the weights of node representations in each layer. This allows the model to pay more attention to node representations that have a greater impact on the prediction of miRNA–disease associations. In this way, the layer-level attention mechanism contributes to the final node representation, enabling the model to better capture information from different layers of node representations:

$$\hat{W}_i = \text{cnn}(\text{att}(W_i^c)) \quad (8)$$

$$\hat{F}_j = \text{cnn}\left(\text{att}\left(F_j^c\right)\right), \tag{9}$$

where \hat{W}_i indicates the ultimate embedding of miRNA, where \hat{F}_j indicates the ultimate embedding of disease. Afterward, element-level multiplication was carried out on the embeddings of miRNA nodes and disease nodes. After that, we utilize a Feedforward Neural Network (FNN) to forecast the likelihood of association between miRNA–disease pairs.

To enhance the model’s performance, we utilized the Sørensen-Dice loss function to compute the loss during the training of the model:

$$L = 1 - \frac{X + \epsilon}{U - X + \epsilon}, \tag{10}$$

where X is acquired:

$$X = \sum_1^N t_i q_i, \tag{11}$$

where U is obtained by:

$$U = \sum_1^N \left(t_i^2 + q_i^2\right), \tag{12}$$

where q_i is the network prediction value, which is the value obtained after sigmoid and takes the value between (0, 1). t_i is the target value, which can only be either 0 or 1.

ϵ is a smoothing factor that serves two purposes. First, it prevents the denominator from predicting zero. Generally, the output of the segmentation network goes through sigmoid or softmax, and there is no case where the output is absolutely 0. The smoothing factor is added here to prevent some extreme cases where the number of output bits is too small and causes the compiler to lose digits. Second, smoothing coefficients can operate to smooth out losses and gradients.

The Sørensen-Dice loss function pays more attention to the overlap of the predictions when calculating the loss. It works on the principle that when calculating the Sørensen-Dice loss function, the intersection and union of the predictions with the true labels need to be calculated in order to assess the degree of overlap of the predictions and consequently enhance the precision and resilience of the model. In binary classification problems, the Sørensen-Dice loss function focuses directly on the identification of positive samples, which is particularly important for miRNA–disease association prediction, as researchers are often more concerned with accurately identifying associations that are present than absent associations. In addition, the application of the Sørensen-Dice loss function reduces the loss value. We take the commonly used loss function Binary Cross Entropy Loss (BCE Loss) and compare it with the Sørensen-Dice loss function by taking the loss values after the first 100 batches as shown in Figure 8.

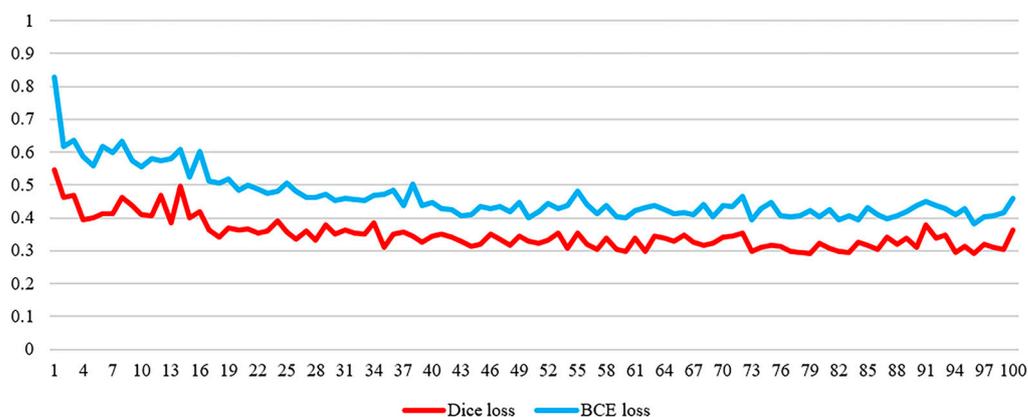


Figure 8. Loss values of the Sørensen-Dice loss function vs. the BCE loss function.

As seen in Figure 8, the loss value of the Sørensen-Dice loss function is lower than that of the BCE loss function. This is because the Sørensen-Dice loss function is much stricter. It takes into account not only their intersection but also their concatenation when calculating the similarity of two sets. Therefore, the Sørensen-Dice loss function requires more accuracy in the model's prediction results and is more capable of reducing the disparity between the predicted consequences and the true consequences compared to the BCE loss function.

5. Conclusions

Improving the prediction of the relationship between diseases and miRNAs can greatly enhance human research on the pathogenesis of diseases. Presently, the predominant methods employed for predicting disease associations rely on Graph Convolutional Network (GCN) techniques. However, the Graph Convolutional Network algorithm, which is locally aggregated, solely incorporates information from the immediate neighboring nodes of a given node at each layer. Consequently, GCN cannot simultaneously aggregate information from multiple nodes. This constraint significantly impacts the predictive efficacy of the model. To tackle this problem, we propose a novel approach, based on HyperGCN and Sørensen-Dice loss (HGSMDA), for predicting associations between miRNAs and diseases. In the initial phase, we developed multiple networks to represent the similarity between miRNAs and diseases and employed GCNs to extract information from diverse perspectives. Subsequently, we draw into HyperGCN to construct a miRNA–disease heteromorphic hypergraph using hypernodes and train GCN on the graph to aggregate information. Finally, we utilized the Sørensen-Dice loss function to evaluate the degree of similarity between the predicted outcomes and the actual values, thereby enabling the prediction of associations between miRNAs and diseases. To comprehensively and objectively evaluate the predictive performance of the HGSMDA model, we employed methods such as five-fold cross-validation and case analysis to assess the model from different perspectives. The evaluation results indicate that HGSMDA exhibits excellent performance and can be used for miRNA–disease association prediction.

However, HGSMDA still has potential for further improvement. Above all, calculating miRNA and disease similarity scores using available information is challenging. For example, similarity between miRNAs can be assessed based on aspects such as their functions, sequences, or interaction networks, while similarity between diseases can be assessed based on aspects such as their phenotypes, gene expression profiles, or clinical features. Therefore, how to effectively integrate this multi-source information and translate it into comparable similarity scores is a complex issue. Second, our method may face computational efficiency problems when dealing with large-scale hypergraphs. The presence of a large number of nodes and hyperedges in hypergraphs makes the representation, storage, and computation of hypergraphs very complex. Especially in graph convolutional networks, processing large-scale hypergraphs may lead to a great consumption of computational and storage resources, thus affecting the efficiency of model training and inference. Therefore, future research could further improve the methodology to increase the accuracy and efficiency of the predictions. This method provides an innovative approach to thinking and a tool to study the association between miRNAs and diseases, which are conducive to gaining insight into the pathogenesis of diseases as well as developing new therapeutic approaches and applying them to more biomedical fields.

Author Contributions: Z.C. wrote the first draft; R.Z., J.L., J.S. and L.D. contributed to the preparation of the final version of our manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the Shandong Social Science Planning Fund Program (Grant No. 21BTQJ02).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This work utilizes data from publically available data set, the HMDD. The miRNA-disease database analyzed in the study is from HMDD v3.2, <http://www.cuilab.cn/hmdd>, accessed on 23 January 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bartel, D.P. MicroRNAs: Target recognition and regulatory functions. *Cell* **2009**, *136*, 215. [[CrossRef](#)] [[PubMed](#)]
2. Chen, X.; Liu, M.X.; Yan, G.Y. RWRMDA: Predicting novel human microRNA-disease associations. *Mol. Biosyst.* **2012**, *8*, 2792–2798. [[CrossRef](#)] [[PubMed](#)]
3. Zhong, Y.; Xuan, P.; Wang, X.; Zhang, T.; Li, J.; Liu, Y.; Zhang, W. A non-negative matrix factorization based method for predicting disease-associated miRNAs in miRNA-disease bilayer network. *Bioinformatics* **2018**, *34*, 267–277. [[CrossRef](#)] [[PubMed](#)]
4. Chen, X.; Yin, J.; Qu, J.; Huang, L. MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction. *PLoS Comput. Biol.* **2018**, *14*, e1006418. [[CrossRef](#)] [[PubMed](#)]
5. Chen, X.; Xie, D.; Wang, L.; Zhao, Q.; You, Z.H.; Liu, H. BNPMDA: Bipartite Network Projection for MiRNA-Disease Association prediction. *Bioinformatics* **2018**, *34*, 3178–3186. [[CrossRef](#)] [[PubMed](#)]
6. Xu, J.; Li, C.-X.; Lv, J.-Y.; Li, Y.-S.; Xiao, Y.; Shao, T.-T.; Huo, X.; Li, X.; Zou, Y.; Han, Q.-L.; et al. Prioritizing Candidate Disease miRNAs by Topological Features in the miRNA Target-Dysregulated Network: Case Study of Prostate Cancer. *Mol. Cancer Ther.* **2011**, *10*, 1857–1866. [[CrossRef](#)] [[PubMed](#)]
7. Chen, X.; Wang, L.; Qu, J.; Guan, N.N.; Li, J.Q. Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* **2018**, *34*, 4256–4265. [[CrossRef](#)] [[PubMed](#)]
8. Xuan, P.; Han, K.; Guo, M.; Guo, Y.; Li, J.; Ding, J.; Liu, Y.; Dai, Q.; Li, J.; Teng, Z.; et al. Prediction of microRNAs Associated with Human Diseases Based on Weighted k Most Similar Neighbors. *PLoS ONE* **2013**, *8*, e70204. [[CrossRef](#)]
9. Zhao, Y.; Chen, X.; Yin, J. Adaptive boosting-based computational model for predicting potential miRNA-disease associations. *Bioinformatics* **2019**, *35*, 4730–4738. [[CrossRef](#)]
10. Shang, J.; Yang, Y.; Li, F.; Guan, B.; Liu, J.-X.; Sun, Y. BLNIMDA: Identifying miRNA-disease associations based on weighted bi-level network. *BMC Genom.* **2022**, *23*, 686. [[CrossRef](#)]
11. Sun, M.; Zhao, S.; Gilvary, C.; Elemento, O.; Zhou, J.; Wang, F. Graph convolutional networks for computational drug development and discovery. *Brief. Bioinform.* **2020**, *21*, 919–935. [[CrossRef](#)]
12. Zhao, T.; Hu, Y.; Valsdottir, L.R.; Zang, T.; Peng, J. Identifying drug-target interactions based on graph convolutional network and deep neural network. *Brief. Bioinform.* **2021**, *22*, 2141–2150. [[CrossRef](#)]
13. Li, Y.; Qiao, G.; Wang, K.; Wang, G. Drug-target interaction predication via multi-channel graph neural networks. *Brief. Bioinform.* **2022**, *23*, bbab346. [[CrossRef](#)] [[PubMed](#)]
14. Peng, W.; Tang, Q.; Dai, W.; Chen, T. Improving cancer driver gene identification using multi-task learning on graph convolutional network. *Brief. Bioinform.* **2022**, *23*, bbab432. [[CrossRef](#)] [[PubMed](#)]
15. Li, J.; Zhang, S.; Liu, T.; Ning, C.; Zhang, Z.; Zhou, W. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. *Bioinformatics* **2020**, *36*, 2538–2546. [[CrossRef](#)]
16. Tang, X.; Luo, J.; Shen, C.; Lai, Z. Multi-view Multichannel Attention Graph Convolutional Network for miRNA-disease association prediction. *Brief. Bioinform.* **2021**, *22*, bbab174. [[CrossRef](#)]
17. Liao, Q.; Ye, Y.; Li, Z.; Chen, H.; Zhuo, L. Prediction of miRNA-disease associations in microbes based on graph convolutional networks and autoencoders. *Front. Microbiol.* **2023**, *14*, 1170559. [[CrossRef](#)] [[PubMed](#)]
18. Huang, Z.; Shi, J.; Gao, Y.; Cui, C.; Zhang, S.; Li, J.; Zhou, Y.; Cui, Q. HMDD v3.0: A database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res.* **2019**, *47*, D1013–D1017. [[CrossRef](#)]
19. Wang, D.; Wang, J.; Lu, M.; Song, F.; Cui, Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* **2010**, *26*, 1644–1650. [[CrossRef](#)]
20. Xiao, Q.; Luo, J.; Liang, C.; Cai, J.; Ding, P. A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* **2018**, *34*, 239–248. [[CrossRef](#)]
21. Lee, I.; Blom, U.M.; Wang, P.I.; Shim, J.; Marcotte, E.M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **2011**, *21*, 1109–1121. [[CrossRef](#)]
22. Dai, Q.; Wang, Z.; Liu, Z.; Duan, X.; Song, J.; Guo, M. Predicting miRNA-disease associations using an ensemble learning framework with resampling method. *Brief. Bioinform.* **2022**, *23*, bbab543. [[CrossRef](#)]
23. Li, Z.; Zhong, T.; Huang, D.; You, Z.-H.; Nie, R. Hierarchical graph attention network for miRNA-disease association prediction. *Mol. Ther.* **2022**, *30*, 1775–1786. [[CrossRef](#)]
24. Zhang, H.; Fang, J.; Sun, Y.; Xie, G.; Lin, Z.; Gu, G. Predicting miRNA-Disease Associations via Node-Level Attention Graph Auto-Encoder. *Ieee-Acm Trans. Comput. Biol. Bioinform.* **2023**, *20*, 1308–1318. [[CrossRef](#)] [[PubMed](#)]
25. Lou, Z.; Cheng, Z.; Li, H.; Teng, Z.; Liu, Y.; Tian, Z. Predicting miRNA-disease associations via learning multimodal networks and fusing mixed neighborhood information. *Brief. Bioinform.* **2022**, *23*, bbac159. [[CrossRef](#)]
26. Wang, W.; Chen, H. Predicting miRNA-disease associations based on lncRNA-miRNA interactions and graph convolution networks. *Brief. Bioinform.* **2022**, *24*, bbac495. [[CrossRef](#)]

27. Ning, Q.; Zhao, Y.; Gao, J.; Chen, C.; Li, X.; Li, T.; Yin, M. AMHMDA: Attention aware multi-view similarity networks and hypergraph learning for miRNA-disease associations identification. *Brief. Bioinform.* **2023**, *24*, 1–11. [[CrossRef](#)] [[PubMed](#)]
28. Liou, J.-M.; Lin, J.-T.; Huang, S.-P.; Chiu, H.-M.; Wang, H.-P.; Lee, Y.-C.; Lin, J.-W.; Shun, C.; Liang, J.-T.; Wu, M.-S. Screening for colorectal cancer in average-risk Chinese population using a mixed strategy with sigmoidoscopy and colonoscopy. *Dis. Colon Rectum* **2007**, *50*, 630–640. [[CrossRef](#)] [[PubMed](#)]
29. Uddin, A.; Chakraborty, S. Role of miRNAs in lung cancer. *J. Cell. Physiol.* **2018**, 1–10. [[CrossRef](#)]
30. Akao, Y.; Nakagawa, Y.; Naoe, T. let-7 microRNA functions as a potential growth suppressor in human colon cancer cells. *Biol. Pharm. Bull.* **2006**, *29*, 903–906. [[CrossRef](#)]
31. Lu, Z.; Liu, M.; Stribinskis, V.; Klinge, C.M.; Ramos, K.S.; Colburn, N.H.; Li, Y. MicroRNA-21 promotes cell transformation by targeting the programmed cell death 4 gene. *Oncogene* **2008**, *27*, 4373–4379. [[CrossRef](#)] [[PubMed](#)]
32. Yang, Z.; Wu, L.; Wang, A.; Tang, W.; Zhao, Y.; Zhao, H.; Teschendorff, A.E. dbDEMC 2.0: Updated database of differentially expressed miRNAs in human cancers. *Nucleic Acids Res.* **2017**, *45*, D812–D818. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.