EasyChair Preprint

№ 9377

Predicting the Human miRNA-Disease Associations Based on Non-Linear Gaussian Profile Kernel Similarity

Zou Haitao, Xie Xiaolan, Ji Boya and Peng Shaoliang

November 28, 2022

Predicting the human miRNA-disease associations based on Non-linear Gaussian Profile Kernel Similarity

1st Haitao Zou
College of Information Science and Engineering
Guilin University of Technology
Guilin 541006, China
zht@glut.edu.cn

2nd Xiaolan Xie*
College of Information Science and Engineering
Guilin University of Technology
Guilin 541006, China
xie_xiao_lan@foxmail.com

3rd Boya Ji
College of Computer Science and Electronic Engineering
Hunan University
Changsha 410082, China
byj@hnu.edu.cn

4th Shaoliang Peng*
College of Computer Science and Electronic Engineering
Hunan University
Changsha 410082, China
slpeng@hnu.edu.cn

***Abstract*—In view of the fact that the traditional methods of determining potential miRNA-disease associations tend to be destructive, labor-intensive, time-consuming, and associated with practice effects, a number of computational methods are being developed to address the burden on biological researchers. In this study, we introduced the computational strategy of non-linear gaussian profile kernel similarity and proposed a novel deep-learning method called NGPKS[1] to engage the in-depth understanding of miRNA-disease associations. More specifically, NGPKS comprehensively integrates the miRNA functional similarity and disease semantic similarity information. Then, the gaussian interaction profile kernel similarity algorithm was utilized to capture the structural information between miRNAs and diseases. Finally, a deep learning framework was constructed for modeling the integration of two types of similarity features. We used three model validation strategies, including five-fold cross-validation, comparison with the state-of-the-art methods, and ablation experiments were used to check the predictive ability of our model. Besides, we conducted case studies for two common diseases. As a result, there are 50 (Colon Cancer), and 47(Lymphoma) among the top 50 predicted miRNAs validated through experiments. Therefore, we could conclude that NGPKS is an effective method to predict potential miRNA-disease associations.**

***Index Terms*—miRNA-disease matrix,machine-learning,non-linear gaussian profile kernel similarity**

I. Introduction

MicroRNAs (miRNAs) are endogenous non-coding small RNAs of approximately 18-25 nucleotides in length, which are transcribed under the action of RNA polymerase II and play a role in regulating gene expression [1]. Studies have shown that miRNAs can not only regulate gene expression but also participate in the emergence and development of various human diseases. Li L *et al*. [2] demonstrated that miRNA-146a plays an important role in innate immunity, inflammatory response, viral infection, and human diseases, and they also discussed the potential use of miRNA-146a as a biomarker for disease diagnosis, prevention, and treatment. In addition, emerging biological experiments have also shown that miRNAs are involved in the process of cell differentiation, biological development [3], and disease progression. Therefore, identifying the correlation between miRNAs and diseases will not only help medical personnel to classify pathologies but also understand the pathological mechanisms of various complex diseases. However, traditional biological experiments are time-consuming and expensive to research the association between miRNAs and disease. With the continuous improvement of computer performance, some computational methods can be used to predict the potential association of miRNA-disease (MDA) efficiently and economically. Therefore, a computer-based approach to reveal potential associations between miRNAs and diseases has become a research hotspot.

There are many computational methods to predict miRNA-disease correlations, and these methods are mainly divided into the following three categories:

(1)**Similarity measure-based prediction.** The method based on similarity measure believes that the higher the functional similarity between miRNAs, the higher the similarity of diseases based on the high similarity of phenotype. For example, You *et al*. [4] proposed a pathway-based miRNA-disease association prediction model (PBMDA) to predict miRNA-disease associations using a depth-first algorithm. Chen *et al*. [5] proposed a similarity measure based on super disease and super miRNA to predict MDA. Chen *et al*. [6] proposed a predictive model of graphlet interaction for the miRNA-disease association, which calculates miRNA-disease correlations by

[1]Availability and implementations:https://github.com/zht-code/NGPKS.git

measuring the graph interaction between two miRNAs or two diseases. Han *et al*. [7] proposed a method to learn embedding-enhanced MDA of miRNA-disease features through graph convolution and attention mechanism. Zhao *et al*. [8] proposed a similarity measure-based method to predict miRNA-disease associations with spy and super cluster strategy (SSCMDA). Chen *et al*. [9] proposed an improved restart random walk strategy to predict miRNA-disease associations. Li *et al*. [10] proposed a new ensemble network approach to calculate similarity by benchmarking lncRNA pairs that share a target miRNA.

(2)**Machine learning based prediction.** With the continuous advancement in the field of artificial intelligence and the emergence of more and more databases verified by biological experiments, it has become possible to predict miRNA-disease associations by means of machine learning. For example, Peng *et al*. [11] proposed a learning-based method to automatically identify the features of a three-layer similarity network through an autoencoder, and input the learned features into a convolutional neural network to predict MDA. Fu *et al*. [12] proposed a deep-learning model to predict MDA by extracting high-level features through stacked autoencoders. Chen *et al*. [13] proposed an extreme gradient boosting machine (EGBMMDA) prediction model is proposed, which forms an informative feature vector by aggregating each miRNA-disease statistics, graph theory, and matrix factorization results, and utilizes a boosted gradient regression tree to predict MDA. Chen *et al*. [14] proposed an ensemble computational model (ELLPMDA) that combines three classical similarity-based algorithms using ensemble learning to predict MDA. Qu *et al*. [15] proposed a KATZ-based prediction model, which predicts MDA by integrating similarity networks to construct heterogeneous networks. Chen *et al*. [16] proposed a ranking KNN-based algorithm to predict potentially relevant MDA. Chen *et al*. [17] proposed a two-region network projection method (BNPMDA), which feeds three similar networks into a clustering machine and predicts the MDA by a two-network recommendation algorithm.

(3)**Matrix-based prediction.** Matrix-based methods are roughly divided into two categories, one is matrix processing performed while processing data, and the other is matrix completion performed after machine learning. Li *et al*. [18] proposed a matrix completion algorithm by updating the known adjacency matrix to predict MDA. Chen *et al*. [19] proposed a computational model for MDA matrix factorization and heterogeneous graph reasoning, which integrates the similarity matrix into a heterogeneous network of sparse learning to predict MDA. Lan *et al*. [20] proposed a kerned bayesian matrix factorization method to infer latent MDAs by integrating similarity networks. Chen *et al*. [21] proposed an inductive matrix model to complete the missing miRNA-disease association by known miRNA functional similarity matrix, disease semantic similarity matrix, and association matrix. Shen *et al*. [22] proposed a collaborative matrix factorization method(CMFMDA) to predict MDA by a similarity network. Chen *et al*. [23] proposed a neighborhood constraint matrix completion method that uses neighborhood similarity to recover the missing correlations to predict MDA. Gao *et al*. [24] proposed a model based on graph laplacian regularized negative matrix factorization to predict MDA by computing the weighted K as the nearest known neighbors.

Although the above methods are effective for MDA prediction, the current findings still have some limitations. First, the similarity measure relies on known association information, but it has little impact on the prediction of new diseases. Furthermore, the prediction effect of machine learning on new diseases was significantly improved. However, in order to further improves the prediction accuracy, it is necessary to reconsider how to select feature representations and machine learning models. Finally, although the matrix-based method also achieves certain results, the matrix method also has certain limitations. For example, the similarity network structure affects the effect of matrix processing, and different network structures correspond to different matrix processing methods. However, not all matrix processing methods are helpful for feature extraction in machine learning.

To overcome the above limitations and deficiencies, we developed a novel non-linear gaussian profile kernel similarity(NGPKS) method for predicting miRNAs by combining existing miRNA functional similarity, disease semantic similarity, miRNA-disease association matrix, and machine learning association with disease. First, we calculated the non-linear gaussian profile kernel similarity matrix of miRNA and disease by the known miRNA functional similarity matrix and disease semantic similarity matrix, respectively. Second, the part of the miRNA functional similarity matrix and the disease semantic similarity matrix with a similarity of 0 was replaced by a non-linear gaussian profile kernel similarity score. Third, we randomly select 5430 positive and negative samples and choose 878 features for each sample, divide these samples into training and testing sets in a certain proportion, and then input the pre-processed training set into a three-layer fully connected layer network for training. Finally, the entire model is trained end-to-end using MSE loss function and backpropagation algorithm. In addition, we also evaluated the predictive performance of the NGPKS model based on 5-fold cross-validation. As a result, the mean area under the curve (AUC) of NGPKS was 0.95, the precision was 0.88, the recall was 0.88, and the F1-score was 0.88. To further validate the performance of our model in predicting unknown miRNA-disease, we performed case studies on colon cancer and lymphoma. The results show that our model can be used as an effective tool to predict potential miRNA-disease correlations.

II. Experimental materials

A. Human miRNA-disease associations

The HMDD v3.2[2] database collected 35,547 miRNA-disease association items from 19,280 papers, including 1,206 miRNA genes and 893 diseases. It has been verified by

[2]http://www.cuilab.cn/hmdd

biological experiments. Human miRNA-disease association data were downloaded from the HMDD v3.2 database, and we selected 495 miRNAs and 383 human diseases from the HMDD database and constructed a binary matrix MD. If miRNA i∈M is associated with disease j∈D, then MD(i,j)=1. MD(i,j)=0 if miRNA i is not associated with disease j or not observed.

B. MiRNA functional similarity

Wang *et al*. [25] built a method to calculate the functional similarity between miRNAs. The hypothesis of this method is that the higher the functional similarity between miRNAs, the higher the similarity based on diseases with high phenotypic similarity. We downloaded miRNA functional similarity scores from https://www.cuilab.cn/files/images/cuilab/misim.zip. We constructed a symmetric matrix MM with 495 rows and 495 columns to carry the miRNA functional similarity scores, where MM($\mathrm{m_i}, \mathrm{m_j}$) represents the functional similarity scores of miRNA i∈M and miRNA j∈M, and the scores are between 0 and 1.

C. Disease semantic similarity model

The disease semantic similarity information used in this paper was obtained from the MeSH database of the US National Library of Medicine [3]. In MeSH, the relationship between diseases is described as a directed acyclic graph (DAG), where nodes represent diseases and edges represent relationships between diseases. In DAG, disease d(i) is represented as DAG(d(i)) = (d(i), N(d(i)), E(d(i))), where N(d(i)) is the set of ancestral nodes of disease d(i) including disease d(i), and E(d(i)) is the set of edges connecting these diseases. Therefore, the semantic contribution of a disease t in DAG(d(i)) to disease d(i) can be calculated as follows:

$$\begin{cases} \mathrm{D}_1(d(i),t) = 1 \quad \text{if} \;\; t = d(i) \\ \mathrm{D}_1(d(i),t) = \max\{\gamma * \mathrm{D}_1(d(i),t') \mid t' \in children \; of \; t\} \;\; \text{if} \;\; t \neq d(i) \end{cases} \tag{1}$$

Here, γ is the semantic contribution factor.The contribution of disease t to the semantic value of disease d(i) decreases as the distance between them increases. Therefore, we can get the semantic value SD_1 of the disease d(i).

$$SD_1(d(i)) = \sum_{t \in N_{d(i)}} D_1(d(i),t) \tag{2}$$

Here, we assume that according to the more shared parts of DAGs among different diseases, the higher the semantic similarity is, and the semantic similarity model SSD_{V1} is constructed. The calculation formula is as follows:

$$SSD_{V1}(d(i),d(j)) = \frac{\sum_{t \in N_{d(i)} \cap N_{d(j)}} (D_1(d(i),t) + D_1(d(j),t))}{SD_1(d(i)) + SD_1(d(j))} \tag{3}$$

In the SSD_{V1} model, we consider the hierarchical relationship of different diseases in the DAG. However, the number of occurrences of different diseases in DAG is different. Therefore, we calculate another disease contribution value based on the number of diseases:

$$D_1'(d(i),t) = -\log\left(\frac{\mathrm{num}(\text{ DAGs }(t))}{\mathrm{num}\ (\text{ diseases }\ )}\right) \tag{4}$$

Here, num(DAGs(t)) represents the number of DAGs containing disease t, and num(diseases) represents the number of diseases. From this, we construct a second disease semantic similarity models SSD_{V2}:

$$SSD_{V2}(d(i),d(j)) = \frac{\sum_{t \in N_{d(i)} \cap N_{d(j)}} (D_1'(d(i),t) + D_1'(d(j),t))}{SD_1(d(i)) + SD_1(d(j))} \tag{5}$$

III. Methods

Based on miRNA-miRNA similarity network, disease-disease similarity network and experimentally validated miRNA-disease data, this study proposes a new NGPKS method that effectively addresses the problems associated with miRNA-disease association-related predictions.

A. Gaussian interaction profile kernel similarity for diseases and miRNAs

Gaussian Interaction profile kernel similarity (GIP) can calculate miRNA-disease information without semantic similarity in HMDD. First, In order to determine whether disease d(i) is associated with each miRNA, we used a binary vector V(d(i)) to represent the interaction profile of disease d(i). GIP similarity GD can be calculated as follows:

$$GD(d(i),d(j)) = \exp\left(-\theta_d \|V(d(i)) - V(d(j))\|^2\right) \tag{6}$$

where θ_d is the width parameter of the function, which is computed by normalizing the parameter:

$$\theta_d = \theta_d' / \left(\frac{1}{nd}\sum_{i=1}^{nd} \|V(d(i))\|^2\right) \tag{7}$$

Finally, GD is the gaussian interaction profile kernel similarity matrix of the disease, where the entity GD(d(i),d(j)) is the gaussian interaction profile kernel similarity between the diseases d(i) and d(j).

Likewise, the GIP similarity GR of miRNAs to each other can be calculated as follows:

$$GR(r(i),r(j)) = \exp\left(-\theta_r \|V(r(i)) - V(r(j))\|^2\right) \tag{8}$$

$$\theta_r = \theta_r' / \left(\frac{1}{nr}\sum_{i=1}^{nr} \|V(r(i))\|^2\right) \tag{9}$$

Here, the miRNA interaction profiles GR(r(i), r(j)) is defined to represent the gaussian interaction profile kernel similarity between r(i) and r(j). θ_r is obtained by normalizing a new

[3]http://www.ncbi.nlm.nih.gov/

bandwidth parameter θ_r' to the average number of associated diseases for all miRNAs.

The gaussian interaction profile kernel similarity is obtained by normalizing a new bandwidth parameter with all miRNAs and diseases to replace the miRNA-disease information without semantic similarity in HMDD. However, this calculation only linearly combines the normalization of the underlying features, which may not be enough to capture the subtle interaction between miRNA features and diseases. To overcome this limitation, we propose a non-linear gaussian profile kernel similarity method in this study to capture the underlying miRNA-disease associations.

B. Non-linear gaussian profile kernels similarity for miRNA

In the non-linear gaussian profile Kernels similarity method, the feature matrix NGM of miRNA and the feature matrix NGD of disease are encoded by two different non-linear gaussian profile kernels similarities, respectively. In the next section, the encoding process of the feature matrix NGM and the feature matrix NGD will be described.

First, the gaussian profile kernel bandwidth parameter θ_r is obtained by dividing the derivative of the gaussian profile kernel bandwidth parameter θ_r' by dividing the normalized miRNA vector. Second, we multiply the bandwidth parameters obtained in the first step with the normalized parameters of the corresponding miRNA features after inversion. In order to measure the weights of observed and unobserved miRNA entries, we added another parameter $\alpha \in$(0,1). This enables better discovery of potential feature relationships. Finally, the calculated non-linear gaussian profile kernel of miRNA is saved into the NGM matrix as a feature. The specific miRNA non-linear gaussian profile kernel similarity formula is as follows:

$$NGM(r(i), r(j)) = \exp((1-\alpha)/2(-\theta_r * \|V(r(i)) - V(r(j))\|^2 + \alpha/2(-\theta_r * \|V(r(i)) - V(r(j))\|^2)) \tag{10}$$

$$\theta_r = \theta_r' / \left(\frac{1}{nr} \sum_{i=1}^{nr} \|V(r(i))\|^2 \right) \tag{11}$$

C. Non-linear gaussian profile kernels similarity for disease

The feature matrix NGD encoding process of the disease is described below.

First, the gaussian profile kernel bandwidth parameter θ_d is obtained by dividing the derivative of the gaussian profile kernel bandwidth parameter θ_d' by dividing the normalized disease vector. Second, we multiply the bandwidth parameters obtained in the first step with the normalized parameters of the corresponding disease features after inversion. In order to measure the weights of observed and unobserved disease entries, we added another parameter $\alpha \in$(0,1). This enables better discovery of potential feature relationships. Finally, the non-linear gaussian profile kernel of the calculated disease is saved into the NGD matrix as a feature. The formula for the specific disease non-linear gaussian profile kernel similarity is as follows:

$$NGD(d(i), d(j)) = \exp((1-\alpha)/2(-\theta_d * \|V(d(i)) - V(d(j))\|^2 + \alpha/2(-\theta_d * \|V(d(i)) - V(d(j))\|^2)) \tag{12}$$

$$\theta_d = \theta_d' / \left(\frac{1}{nd} \sum_{i=1}^{nd} \|V(d(i))\|^2 \right) \tag{13}$$

D. Integration of non-linear gaussian profile kernels for miRNAs-disease similarity

In this paper, we finally use the descriptors including disease similarity, miRNA similarity, miRNA non-linear gaussian profile kernel matrix, disease non-linear gaussian profile kernel matrix, and miRNA-disease correlation matrix.

For diseases, we construct a semantic similarity model SSD_{V1}, SSD_{V2}, disease non-linear gaussian profile kernel matrix NGD. From this, we get the disease feature matrix KD:

$$KD(d(i), d(j)) = \begin{cases} \frac{SSD_{V1}(d(i),d(j)) + SSD_{V2}(d(i),d(j))}{2}, \\ \text{if } there\ is\ semantic\ similarity\ between\ d(i)\ \text{ and }\ d(j) \\ NGD(d(i), d(j)), \text{ else} \end{cases} \tag{14}$$

For miRNAs, we combine the functional similarity matrix MM with the miRNA non-linear gaussian profile kernel matrix NGM to form the miRNA feature matrix KM, which is expressed as follows:

$$KM(r(i), r(j)) = \begin{cases} MM(r(i), r(j)), \\ \text{if there is semantic similarity between } d(i) \text{ and } d(j) \\ NGM(r(i), r(j)), \text{ else} \end{cases} \tag{15}$$

E. NGPKS with fully connected nuralnetwork for miRNA-disease association prediction

we will introduce how the NGPKS model uses the feature matrix for sample screening, and how to input it into the fully connected neural network for end-to-end model training and miRNA-disease association prediction. An overview of the framework of NGPKS is shown in Figure 1.

First, we filtered 5430 positive and negative samples indexed from the miRNA-disease correlation matrix, and then took out the features corresponding to their indexes to form a total of 10,860 positive and negative samples, which each sample had 878 features. Second, the selected samples are divided into a training set and test set according to a certain proportion, and the training set is input into a three-layer fully connected neural network for end-to-end model training. The definition of t-layer fully connected for model training on the training set is as follows:

$$Score^{(t)}(X) = \text{relu}\left(W^{(t)} \text{ relu}\left(\cdots \text{Sigmoid}\left(W^{(1)}X + b^1\right)\cdots\right) + b^t\right) \tag{16}$$

Among them, $\mathbf{W}^{(t)}(t \in \{1, 2, 3, \ldots\})$ represents the weight matrix of the t-th layer in the non-linear fully connected layer, and $\mathbf{X}^t \in \mathbb{R}^{f^t \times f^{t+1}}$ represents the input feature matrix of the

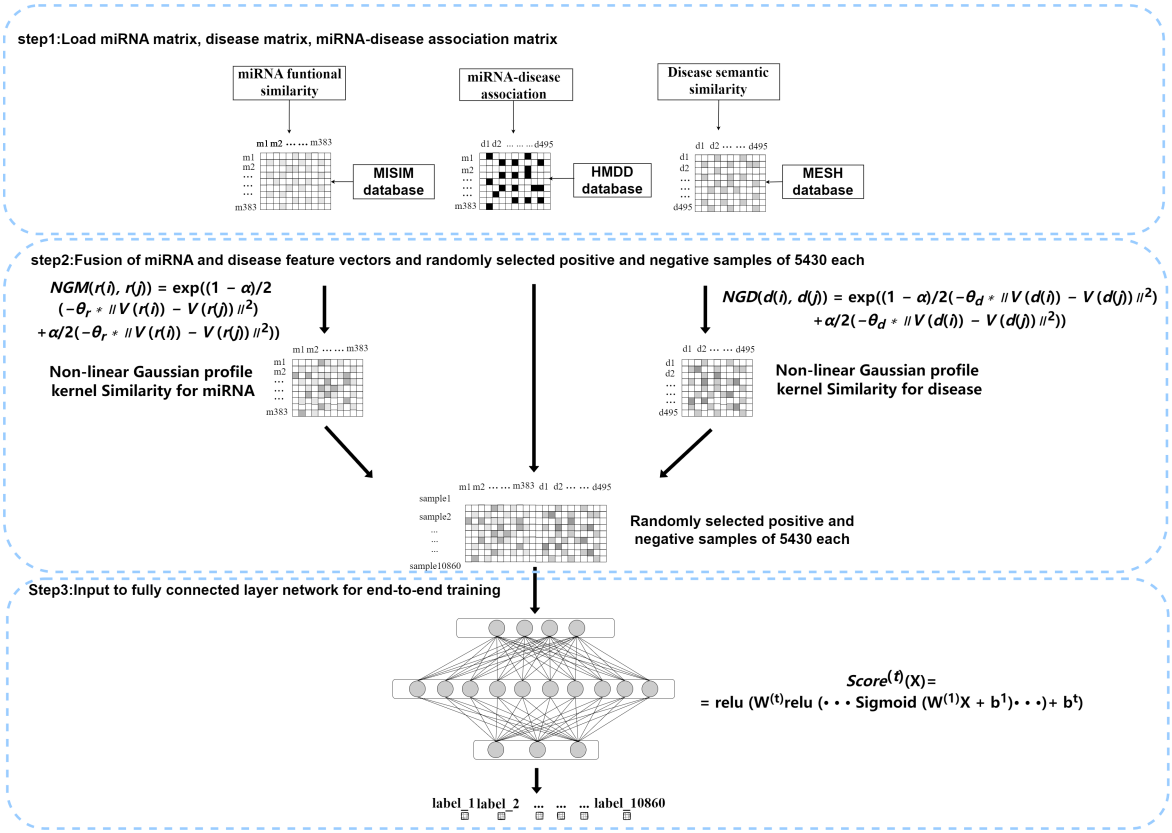

Fig. 1. The framework of NGPKS. Firstly, NGPKS utilized miRNA functional similarity based on miRNA-disease association data and disease similarity scores, and then calculated disease similarity based on MeSH disease terms. Secondly, a gaussian interaction profile kernel similarity algorithm is used to capture the structural information between miRNAs and diseases. Finally, the integration of the two types of similarity features is modeled using a deep learning framework.

t-th layer, where f^t and f^{t+1} represents the input and output dimensions of the t-th layer. b^t is the bias term of the t-th layer, and relu(·) and Sigmoid are the non-linear activation functions of the corrected linear unit.

Finally, the model parameters are optimized and the loss value between the predicted probability and the true label is calculated by MSE loss function, which is defined as follows:

$$\text{Loss} = \frac{1}{dr} \sum_{i=1,j=1}^{dr} \|score(d(i), r(j)) - MD(d(i), r(j))\|^2 \tag{17}$$

IV. Results and Discussion

A. Performance evaluation

In the experiments, we use 5-fold cross-validation (5FCV) and 10-fold cross-validation (10FCV) to evaluate the predictive performance of NGPKS. In 5FCV and 10FCV, we take all known miRNA-disease association samples as positive samples, and randomly divide the positive samples into 5 subsamples or 10 subsamples, and each subsample is used as a test sample in turn, and the rest are used as training samples. When we predicted the score was higher than 0.5, we considered the model to successfully predict miRNA-disease. At training time, we s*et al*l miRNA-disease association labels in the association matrix MM to 0.

In order to fairly evaluate the proposed model, we follow common evaluation criteria including area under the receiver operating characteristic curve (ROC), the area under precision and recall (AUPR), precision, recall, and F1-score. The ROC curve is based on the false positive rate (FPR) under different thresholds as the abscissa and the true rate (TPR) under different thresholds as the ordinate. AUPR is the area under the PR curve, and the PR curve is a graph of recall and accuracy. Accuracy is the proportion of the true minority class among all samples that we predicted to be the minority class. The recall rate is the proportion of all samples whose true value is 1 that we predict correctly. The F1-score is that the harmonic mean between precision and recall tends to be closer to the smaller of the two numbers.

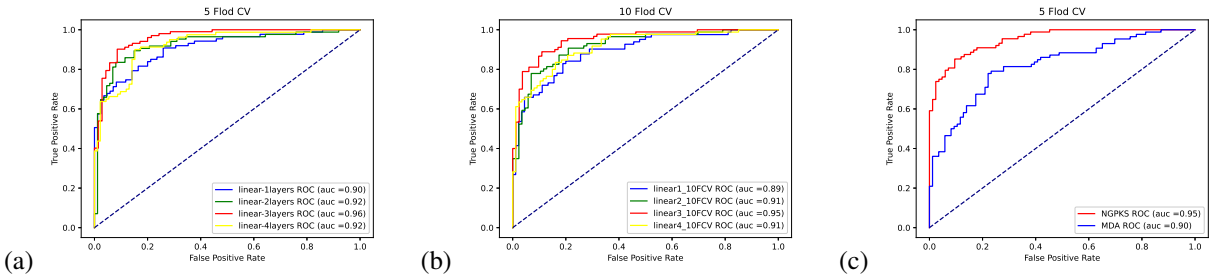

Fig. 2. (a) Comparison of the prediction performance of different fully connected layers after 5FCV;(b)Comparison of the prediction performance of different fully connected layers after 10FCV;(c)Comparing the ROC performance of two variants of NGPKS (NGPKS and MDA)

B. Ablation experiments

In this section, we will introduce the components that significantly affect the performance of NGPKS. One is the effect of t-layers of linear encoders, and the other is the effect of non-linear gaussian profile kernels.

1) Effect of t-layers of linear encoders: In this section, we need to determine the number of linear connection layers in the fully connected layer. Based on the miRNA-disease association data set, we take the linear layers as 1, 2, 3, and 4, and perform 5-fold cross-validation and 10-fold cross-validation on different layers of linear in turn. The results are shown in Fig.2 (a) and Fig.2 (b). In Figure 2 (a) and (b), when the number of linear layers is 3, the model performance is optimal. However, when the number of linear layers is increased again, the performance of the model decreases. This suggests that deepening the number of fully connected layers may cause over-smoothing, resulting in a decrease in the overall predictive performance of the model. Therefore, the number of linear layers in the fully connected layer of the model in this paper is 3 layers.

2) Effect of non-linear gaussian profile kernels: In this paper, the model NGPKS introduces non-linear gaussian profile kernel similarity, fusing the basic features and non-linear features of miRNA and disease to predict miRNA-disease associations. In order to verify the validity of the non-linear gaussian profile kernel similarity, we employ two variants of NGPKS (NGPKS and MDA) as comparison methods. Specifically, five-fold cross-validation experiments were performed for NGPKS and MDA under the same baseline model with the same parameters. NGPKS is a prediction model using non-linear gaussian profile kernel similarity. MDA is an experiment that removes non-linear gaussian profile kernel similarity and uses common miRNA and disease features for prediction.

As can be seen from Figure 2(c), the metrics for calculating similarity using features of common miRNAs and diseases are lower than those using non-linear gaussian profile kernel similarity. This is because miRNAs with high similarity are often associated with similar diseases, and the non-linear gaussian profile kernel similarity fuses potential associations between different similarity information, which can better capture the characteristics between miRNA-diseases, thereby improving the miRNA-disease association prediction accuracy. Figure 2(c) shows the performance of NGPKS and its variant model. It can be observed that the ROC of NGPKS is better than that of the variant model.

C. Comparison with other methods

In recent years, researchers have proposed many methods to predict miRNA-disease associations. But the datasets used and the forecasting methods vary. In order to compare the performance of our model NGPKS with existing methods, we conduct experimental validations from the following aspects.

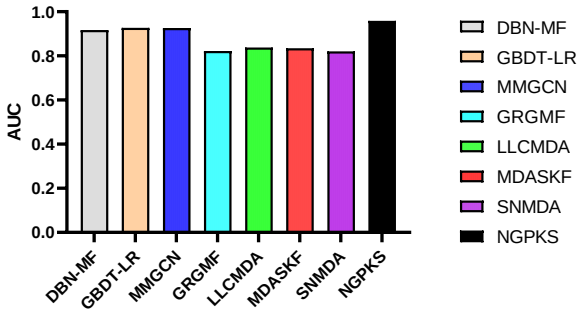

Fig. 3. Comparison of 5FCV with other ten methods on the same database

First, we compare the prediction results of NGPKS with seven methods in different aspects of the same dataset. The seven competing approaches include a matrix factorization model based on deep belief networks (DBN-MF) [26], a model combining gradient boosted decision trees and logistic regression (GBDT-LR) [27], a multi-view multi-channel attention graph convolutional network model (MMGCN) [28], a model based on graph regularized generalized matrix factorization (GRGMF) [29], a model based on local constrained linear coding (LLCMDA) [30], a laplacian regularized least squares method (MDA-SKF) [31]and a sparse neighborhood-based method (SNMDA) [32]. The comparison results are shown in table 1, it is obvious that our method performs well in AUC, AUPR, precision, recall, and F1-score. Therefore, the prediction performance of NGPKS outperforms other competing methods.

Second, we compare the AUC values of NGPKS with 10 other methods by 5FCV on the same dataset. The methods compared include BNPMDA [17], DNRLCNN [33], LR-GNN [34], MCMDA [18], MDHGI [19], MDSCMF [35], NIMGSA

TABLE I
TABLE 1 THE COMPARISON BETWEEN NGPKS AND OTHER SEVEN METHODS ON AUC, AUPR, PRECISION, RECALL, F1-SCORE VALUE ON THE SAME DATABASE BY 5-CV.

Methods	AUC	AUPR	precision	recall	F1-score
DBN-MF	0.9169	0.9043	0.8377	0.8526	0.8451
GBDT-LR	0.9274	0.9014	0.8315	0.8273	0.8302
MMGCN	0.9266	0.2589	0.3338	0.2989	0.3146
GRGMF	0.8217	0.0769	0.1213	0.2757	0.1567
LLCMDA	0.8378	0.1022	0.1653	0.2261	0.1886
MDASKF	0.8344	0.1079	0.1794	0.2185	0.1932
SNMDA	0.8204	0.1115	0.1770	0.2857	0.2126
NGPKS	**0.9584**	**0.9116**	**0.8873**	**0.8863**	**0.8874**

[36] , NMFMC [37], SAEMDA [38] and SGNNMD [39]. Figure 3 shows the AUC values of all methods. It can be seen from the figure that the AUC of our method is significantly better than other methods.

Finally, the proposed strategy does improve the performance of the model to some extent, but it is still evident from Table 1 that the model has a lot of room for improvement. Observing Figure 2, our ROC curve shows the optimal, but the advantage is not so obvious. Therefore, we still have to further improve the prediction accuracy on future work.

D. Case Studies

To further evaluate the performance of NGPKS in predicting novel miRNA-disease association information, we performed a specific case study. We still use the data selected above in the HMDD database for the dataset, which has 5430 known association entries and 184155 unknown association entries. We made predictions for colon cancer and lymphoma separately and validated all predicted top 50 potentially relevant miRNAs. We used the miRCancer database [40] and the dbDEMC database [41] to validate miRNAs potentially associated with these two diseases. If our predicted potentially associated miRNAs are included in at least one database, it means that our model (NGPKS) successfully predicts the association of miRNAs with this disease, further demonstrating the validity of our model for predicting miRNA-disease. Two case studies are presented below.

Colon cancer is a common malignant tumor in the gastrointestinal tract, and its incidence is second only to gastric cancer and esophageal cancer. Because colon cancer has no obvious early symptoms, it is often missed or misdiagnosed clinically. Therefore, there is an urgent need for more sensitive and specific molecular biomarkers to facilitate the early diagnosis of colon cancer. Here, colon cancer is the first case study in this model. Similarly, the colon cancer-related miRNAs predicted by this model were validated by dbDEMC and miRCancer databases. The experimental results are shown in Table 2, in which the top 50 candidate miRNAs have all been verified by the dataset. Some typical miRNA-related works are introduced below. Long *et al*. [42] showed that a-mir-214 can inhibit the growth of colon cancer cells by inhibiting ARL2 (ADP-ribosylating factor-like protein 2), and also pointed out that this miRNA may be the future treatment of colon cancer an important goal. Valeri *et al*. [43] showed that miR-135b upregulation is associated with tumor stage and poor clinical outcome and has been identified as a critical downstream effector of oncogenic pathways and a potential target for colon cancer therapy.

Lymphoma is a type of cancer that arises from lymphoid tissue. Hodgkin lymphoma and non-Hodgkin lymphoma are the two main subtypes of lymphoma. Hodgkin lymphoma and non-Hodgkin lymphoma occur in children, adolescents, and adults. Lymphoma cancer becomes more common as people age. Unlike most cancers, lymphoma rates are highest among adolescents and young adults (ages 15 to 39) and older adults (aged 75 or older). Whites are more likely to develop lymphoma than blacks, and men are more likely to develop lymphoma than women. We used the NGPKS model to predict miRNAs with potential associations with lymphoma and then selected the top 50 miRNAs from the predicted results for validation with miR2Disease and dbDEMC. The validation results are shown in Table 3. The validation results showed that 47 of the top 50 miRNAs associated with lymphoma were confirmed. For example, Akao *et al*. [44] found that 8 of 9 patients with B-cell lymphoma had extremely low levels of miR-145 expression. The miR-145 is consistently expressed at low levels in human Burkitt's lymphoma cell lines and inversely correlates with the cell proliferation observed in EBV (Epstein-Barvirus) transformed B-cell lines. The oncogene Myc plays an important role in the development of B-cell lymphoma, especially in Burkitt's lymphoma. Cheson *et al*. [45] confirmed that the oncogene Myc regulated by miR-125b is related to B-cell lymphoma and found to be downregulated in Burkitt's lymphoma in kitt lymphoma. Craig *et al*. [46] found that among the Myc-repressed miRNAs downregulated in malignant lymphomas, miR-34a exhibited the strongest antiproliferative properties when overexpressed in diffuse large B-cell lymphoma cells.

V. CONCLUSION

MiRNAs are associated with a variety of human diseases, and identifying miRNA-disease associations can help in clinical trials and the treatment of diseases. In recent years, there have been more and more miRNA-disease-related databases, which also help researchers to further predict unknown miRNA-diseases through computational methods. In this study, we propose a novel non-linear gaussian profile kernel similarity method (NGPKS) to predict miRNA-disease associations. First, we calculated the non-linear gaussian profile kernel similarity matrix of miRNA and disease by the known miRNA functional similarity matrix and disease semantic similarity matrix, respectively. Second, the part of the miRNA functional similarity matrix and the disease semantic similarity matrix with a similarity of 0 was replaced by a non-linear gaussian profile kernel similarity score. Third, we randomly select 5430 positive and negative samples and choose 878 features for each sample, divide these samples into training and testing sets in a certain proportion, and then input the pre-processed training set into a three-layer

TABLE II
THE TOP 50 VERIFIED ASSOCIATIONS ASSOCIATED WITH COLON CANCER.

Rank	miRNAs	Evidence	Rank	miRNAs	Evidence
1	hsa-mir-125a	dbDEMC;miRCancer	26	hsa-mir-650	dbDEMC
2	hsa-mir-198	dbDEMC	27	hsa-mir-744	dbDEMC
3	hsa-mir-29b	dbDEMC	28	hsa-mir-100	dbDEMC;miRCancer
4	hsa-mir-15a	dbDEMC;miRCancer	29	hsa-mir-214	dbDEMC;miRCancer
5	hsa-mir-208b	dbDEMC	30	hsa-mir-484	dbDEMC;miRCancer
6	hsa-mir-151a	dbDEMC	31	hsa-mir-503	dbDEMC;miRCancer
7	hsa-mir-191	dbDEMC	32	hsa-let-7d	dbDEMC
8	hsa-mir-192	dbDEMC;miRCancer	33	hsa-mir-106b	dbDEMC
9	hsa-mir-193b	dbDEMC;miRCancer	34	hsa-mir-132	dbDEMC
10	hsa-mir-204	dbDEMC;miRCancer	35	hsa-mir-15b	dbDEMC;miRCancer
11	hsa-mir-205	dbDEMC;miRCancer	36	hsa-mir-222	dbDEMC
12	hsa-mir-223	dbDEMC;miRCancer	37	hsa-mir-301b	dbDEMC
13	hsa-mir-449a	dbDEMC	38	hsa-mir-376a	dbDEMC
14	hsa-mir-449b	dbDEMC	39	hsa-mir-376b	dbDEMC
15	hsa-mir-99b	dbDEMC	40	hsa-mir-376c	dbDEMC
16	hsa-mir-101	dbDEMC;miRCancer	41	hsa-mir-424	dbDEMC;miRCancer
17	hsa-mir-196b	dbDEMC	42	hsa-mir-675	dbDEMC;miRCancer
18	hsa-mir-30b	dbDEMC	43	hsa-mir-410	dbDEMC;miRCancer
19	hsa-mir-30c	dbDEMC	44	hsa-mir-128	dbDEMC
20	hsa-mir-320c	dbDEMC	45	hsa-mir-153	dbDEMC;miRCancer
21	hsa-mir-374b	dbDEMC;miRCancer	46	hsa-mir-181c	dbDEMC
22	hsa-mir-421	dbDEMC	47	hsa-mir-590	dbDEMC;miRCancer
23	hsa-mir-433	dbDEMC;miRCancer	48	hsa-mir-485	dbDEMC
24	hsa-mir-452	dbDEMC	49	hsa-mir-509	dbDEMC
25	hsa-mir-519a	dbDEMC	50	hsa-mir-765	dbDEMC

TABLE III
THE TOP 50 VERIFIED ASSOCIATIONS ASSOCIATED WITH LYMPHOMA.

Rank	miRNAs	Evidence	Rank	miRNAs	Evidence
1	hsa-mir-125a	dbDEMC	26	hsa-mir-205	dbDEMC
2	hsa-mir-499a	dbDEMC	27	hsa-mir-215	dbDEMC
3	hsa-mir-29a	dbDEMC;miRCancer	28	hsa-mir-221	dbDEMC
4	hsa-mir-29b	dbDEMC;miRCancer	29	hsa-mir-223	dbDEMC;miRCancer
5	hsa-let-7a	dbDEMC	30	hsa-mir-25	dbDEMC
6	hsa-mir-141	dbDEMC	31	hsa-mir-26b	dbDEMC;miRCancer
7	hsa-mir-143	dbDEMC	32	hsa-mir-31	dbDEMC
8	hsa-mir-145	dbDEMC;miRCancer	33	hsa-mir-34b	dbDEMC
9	hsa-mir-1	dbDEMC	34	hsa-mir-429	**unconfirmed**
10	hsa-mir-133a	dbDEMC	35	hsa-mir-449a	dbDEMC
11	hsa-mir-208b	dbDEMC	36	hsa-mir-449b	**unconfirmed**
12	hsa-mir-103a	**unconfirmed**	37	hsa-mir-93	dbDEMC
13	hsa-mir-106a	dbDEMC;miRCancer	38	hsa-mir-95	dbDEMC
14	hsa-mir-10b	dbDEMC	39	hsa-mir-99b	dbDEMC
15	hsa-mir-151a	dbDEMC	40	hsa-let-7e	dbDEMC
16	hsa-mir-152	dbDEMC	41	hsa-mir-1246	dbDEMC
17	hsa-mir-181b	dbDEMC	42	hsa-mir-125b	dbDEMC
18	hsa-mir-182	dbDEMC	43	hsa-mir-146b	dbDEMC;miRCancer
19	hsa-mir-183	dbDEMC	44	hsa-mir-148a	dbDEMC
20	hsa-mir-191	dbDEMC	45	hsa-mir-148b	dbDEMC
21	hsa-mir-192	dbDEMC	46	hsa-mir-196b	dbDEMC
22	hsa-mir-193b	dbDEMC;miRCancer	47	hsa-mir-219	dbDEMC
23	hsa-mir-194	dbDEMC	48	hsa-mir-27a	dbDEMC
24	hsa-mir-195	dbDEMC	49	hsa-mir-27b	dbDEMC
25	hsa-mir-204	dbDEMC	50	hsa-mir-34a	dbDEMC

fully connected layer network for training. Finally, the entire model is trained end-to-end using MSE loss function and backpropagation algorithm. As a result, the AUC of NGPKS has the highest value of 0.95, confirming that our method significantly outperforms the existing methods. To comprehensively evaluate the performance of our method, we compare it with 8 state-of-the-art computational models under 5FCV. The results show that our method can effectively predict miRNA-disease correlations. In addition, we have conducted case studies. After validation, our method can effectively predict unknown miRNA-disease associations.

The excellent predictive performance of NGPKS is attributed to several important factors. First, NGPKS applies to diseases with unknown associated miRNAs, which greatly improves the utility of this method. Second, the non-linear gaussian profile kernel similarity method can enhance the features between miRNA-diseases, and the combination between new features can be better captured when calculating similarity. Finally, NGPKS can obtain more miRNA-disease feature representations from heterogeneous networks than simply linearly combining features from different levels.

Although NGPKS has good performance in predicting novel miRNA-disease associations, it has some limitations. For example, the calculation of similarity needs further improvement. Furthermore, we simply take the final result as the average of the prediction scores from different similarity networks, which may lead to suboptimal results. Therefore, we remain eager to develop new computational models to overcome these limitations.

ACKNOWLEDGMENT

This work was supported by NSFC Grants (62262011, U19A2067); Natural Science Foundation of Guangxi (2021JJA170130); Science Foundation for Distinguished Young Scholars of Hunan Province (2020JJ2009); National Key R&D Program of China (2017YFB0202602, 2018YFC0910405, 2017YFC1311003, 2016YFC1302500); Science Foundation of Changsha (Z20206942-0652, kq2004010; JZ20195242029, JH20199142034); The Funds of State Key Laboratory of Chemo/Biosensing and Chemometrics and Peng Cheng Lab.

REFERENCES

[1] V. Ambros, "The functions of animal micrornas," *Nature*, vol. 431, no. 7006, pp. 350–355, 2004.

[2] L. Li, X.-P. Chen, and Y.-J. Li, "Microrna-146a and human disease," *Scandinavian journal of immunology*, vol. 71, no. 4, pp. 227–231, 2010.

[3] A. L. Kasinski and F. J. Slack, "Microrna en route to the clinic: progress in validating and targeting microrna for cancer therapy," *Nature Reviews Cancer*, vol. 11, no. 12, pp. 849–864, 2011.

[4] Z.-H. You, Z.-A. Huang, Z. Zhu, G.-Y. Yan, Z.-W. Li, Z. Wen, and X. Chen, "Pbmda: A novel and effective path-based computational model for mirna-disease association prediction," *PLoS computational biology*, vol. 13, no. 3, p. e1005455, 2017.

[5] X. Chen, Z.-C. Jiang, D. Xie, D.-S. Huang, Q. Zhao, G.-Y. Yan, and Z.-H. You, "A novel computational model based on super-disease and mirna for potential mirna–disease association prediction," *Molecular bioSystems*, vol. 13, no. 6, pp. 1202–1212, 2017.

[6] X. Chen, N.-N. Guan, J.-Q. Li, and G.-Y. Yan, "Gimda: graphlet interaction-based mirna-disease association prediction," *Journal of cellular and molecular medicine*, vol. 22, no. 3, pp. 1548–1561, 2018.

[7] H. Han, R. Zhu, J.-X. Liu, and L.-Y. Dai, "Predicting mirna-disease associations via layer attention graph convolutional network model," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1–8, 2022.

[8] Q. Zhao, D. Xie, H. Liu, F. Wang, G.-Y. Yan, and X. Chen, "Sscmda: spy and super cluster strategy for mirna-disease association prediction," *Oncotarget*, vol. 9, no. 2, p. 1826, 2018.

[9] X. Chen, Z.-H. You, G.-Y. Yan, and D.-W. Gong, "Irwrlda: improved random walk with restart for lncrna-disease association prediction," *Oncotarget*, vol. 7, no. 36, p. 57919, 2016.

[10] J. Li, Y. Zhao, S. Zhou, Y. Zhou, and L. Lang, "Inferring lncrna functional similarity based on integrating heterogeneous network data," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 27, 2020.

[11] J. Peng, W. Hui, Q. Li, B. Chen, J. Hao, Q. Jiang, X. Shang, and Z. Wei, "A learning-based framework for mirna-disease association identification using neural networks," *Bioinformatics*, vol. 35, no. 21, pp. 4364–4371, 2019.

[12] L. Fu and Q. Peng, "A deep ensemble model to predict mirna-disease association," *Scientific reports*, vol. 7, no. 1, pp. 1–13, 2017.

[13] X. Chen, L. Huang, D. Xie, and Q. Zhao, "Egbmmda: extreme gradient boosting machine for mirna-disease association prediction," *Cell death & disease*, vol. 9, no. 1, pp. 1–16, 2018.

[14] X. Chen, Z. Zhou, and Y. Zhao, "Ellpmda: ensemble learning and link prediction for mirna-disease association prediction," *RNA biology*, vol. 15, no. 6, pp. 807–818, 2018.

[15] Y. Qu, H. Zhang, C. Liang, and X. Dong, "Katzmda: prediction of mirna-disease associations based on katz model," *Ieee Access*, vol. 6, pp. 3943–3950, 2017.

[16] X. Chen, Q.-F. Wu, and G.-Y. Yan, "Rknnmda: ranking-based knn for mirna-disease association prediction," *RNA biology*, vol. 14, no. 7, pp. 952–962, 2017.

[17] X. Chen, D. Xie, L. Wang, Q. Zhao, Z.-H. You, and H. Liu, "Bnpmda: bipartite network projection for mirna–disease association prediction," *Bioinformatics*, vol. 34, no. 18, pp. 3178–3186, 2018.

[18] 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, p. 21187, 2017.

[19] X. Chen, J. Yin, J. Qu, and L. Huang, "Mdhgi: matrix decomposition and heterogeneous graph inference for mirna-disease association prediction," *PLoS computational biology*, vol. 14, no. 8, p. e1006418, 2018.

[20] 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 transactions on computational biology and bioinformatics*, vol. 15, no. 6, pp. 1774–1782, 2016.

[21] 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.

[22] Z. Shen, Y.-H. Zhang, K. Han, A. K. Nandi, B. Honig, and D.-S. Huang, "mirna-disease association prediction with collaborative matrix factorization," *Complexity*, vol. 2017, 2017.

[23] X. Chen, L.-G. Sun, and Y. Zhao, "Ncmcmda: mirna–disease association prediction through neighborhood constraint matrix completion," *Briefings in bioinformatics*, vol. 22, no. 1, pp. 485–496, 2021.

[24] Z. Gao, Y.-T. Wang, Q.-W. Wu, J.-C. Ni, and C.-H. Zheng, "Graph regularized l2, 1-nonnegative matrix factorization for mirna-disease association prediction," *BMC bioinformatics*, vol. 21, no. 1, pp. 1–13, 2020.

[25] 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, 2010.

[26] Y. Ding, F. Wang, X. Lei, B. Liao, and F.-X. Wu, "Deep belief network–based matrix factorization model for microrna-disease associations prediction," *Evolutionary Bioinformatics*, vol. 16, p. 1176934320919707, 2020.

[27] S. Zhou, S. Wang, Q. Wu, R. Azim, and W. Li, "Predicting potential mirna-disease associations by combining gradient boosting decision tree with logistic regression," *Computational Biology and Chemistry*, vol. 85, p. 107200, 2020.

[28] X. Tang, J. Luo, C. Shen, and Z. Lai, "Multi-view multichannel attention graph convolutional network for mirna–disease association prediction," *Briefings in Bioinformatics*, vol. 22, no. 6, p. bbab174, 2021.

[29] Z.-C. Zhang, X.-F. Zhang, M. Wu, L. Ou-Yang, X.-M. Zhao, and X.-L. Li, "A graph regularized generalized matrix factorization model

for predicting links in biomedical bipartite networks," *Bioinformatics*, vol. 36, no. 11, pp. 3474–3481, 2020.
[30] Y. Qu, H. Zhang, C. Lyu, and C. Liang, "Llcmda: a novel method for predicting mirna gene and disease relationship based on locality-constrained linear coding," *Frontiers in genetics*, vol. 9, p. 576, 2018.
[31] L. Jiang, Y. Ding, J. Tang, and F. Guo, "Mda-skf: similarity kernel fusion for accurately discovering mirna-disease association," *Frontiers in Genetics*, vol. 9, p. 618, 2018.
[32] Y. Qu, H. Zhang, C. Liang, P. Ding, and J. Luo, "Snmda: A novel method for predicting micro rna-disease associations based on sparse neighbourhood," *Journal of Cellular and Molecular Medicine*, vol. 22, no. 10, pp. 5109–5120, 2018.
[33] J. Zhong, W. Zhou, J. Kang, Z. Fang, M. Xie, Q. Xiao, and W. Peng, "Dnrlcnn: A cnn framework for identifying mirna–disease associations using latent feature matrix extraction with positive samples," *Interdisciplinary Sciences: Computational Life Sciences*, pp. 1–16, 2022.
[34] C. Kang, H. Zhang, Z. Liu, S. Huang, and Y. Yin, "Lr-gnn: A graph neural network based on link representation for predicting molecular associations," *Briefings in Bioinformatics*, vol. 23, no. 1, p. bbab513, 2022.
[35] J. Ni, L. Li, Y. Wang, C. Ji, and C. Zheng, "Mdscmf: Matrix decomposition and similarity-constrained matrix factorization for mirna–disease association prediction," *Genes*, vol. 13, no. 6, p. 1021, 2022.
[36] C. Jin, Z. Shi, K. Lin, and H. Zhang, "Predicting mirna-disease association based on neural inductive matrix completion with graph autoencoders and self-attention mechanism," *Biomolecules*, vol. 12, no. 1, p. 64, 2022.
[37] X. Zheng, C. Zhang, and C. Wan, "Mirna-disease association prediction via non-negative matrix factorization based matrix completion," *Signal Processing*, vol. 190, p. 108312, 2022.
[38] C.-C. Wang, T.-H. Li, L. Huang, and X. Chen, "Prediction of potential mirna–disease associations based on stacked autoencoder," *Briefings in Bioinformatics*, vol. 23, no. 2, p. bbac021, 2022.
[39] G. Zhang, M. Li, H. Deng, X. Xu, X. Liu, and W. Zhang, "Sgnnmd: signed graph neural network for predicting deregulation types of mirna-disease associations," *Briefings in Bioinformatics*, vol. 23, no. 1, p. bbab464, 2022.
[40] 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.
[41] Z. Yang, F. Ren, C. Liu, S. He, G. Sun, Q. Gao, L. Yao, Y. Zhang, R. Miao, Y. Cao, *et al.*, "dbdemc: a database of differentially expressed mirnas in human cancers," in *BMC genomics*, vol. 11, pp. 1–8, Springer, 2010.
[42] L.-M. Long, B.-F. He, G.-Q. Huang, Y.-H. Guo, Y.-S. Liu, and J.-R. Huo, "microrna-214 functions as a tumor suppressor in human colon cancer via the suppression of adp-ribosylation factor-like protein 2," *Oncology letters*, vol. 9, no. 2, pp. 645–650, 2015.
[43] N. Valeri, C. Braconi, P. Gasparini, C. Murgia, A. Lampis, V. Paulus-Hock, J. R. Hart, L. Ueno, S. I. Grivennikov, F. Lovat, *et al.*, "Microrna-135b promotes cancer progression by acting as a downstream effector of oncogenic pathways in colon cancer," *Cancer cell*, vol. 25, no. 4, pp. 469–483, 2014.
[44] Y. Akao, Y. Nakagawa, Y. Kitade, T. Kinoshita, and T. Naoe, "Down-regulation of microrna-143 and-145 in b-cell malignancies," *Cancer science*, vol. 98, no. 12, pp. 1914–1920, 2007.
[45] B. D. Cheson, B. Pfistner, M. E. Juweid, R. D. Gascoyne, L. Specht, S. J. Horning, B. Coiffier, R. I. Fisher, A. Hagenbeek, E. Zucca, *et al.*, "Revised response criteria for malignant lymphoma," *Journal of clinical oncology*, vol. 25, no. 5, pp. 579–586, 2007.
[46] V. J. Craig, S. B. Cogliatti, J. Imig, C. Renner, S. Neuenschwander, H. Rehrauer, R. Schlapbach, S. Dirnhofer, A. Tzankov, and A. Müller, "Myc-mediated repression of microrna-34a promotes high-grade transformation of b-cell lymphoma by dysregulation of foxp1," *Blood, The Journal of the American Society of Hematology*, vol. 117, no. 23, pp. 6227–6236, 2011.