IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 July 2024, accepted 1 August 2024, date of publication 9 August 2024, date of current version 19 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3440951

RESEARCH ARTICLE

Potential microRNA-Disease Association Prediction Using Node2vec and Singular Value Decomposition

YUNXIA LIU, JIAZHEN LIN, PIN LIANG, YAYU TIAN, AND XUAN HE
College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110819, China

Corresponding author: Xuan He (hexuan@bmie.neu.edu.cn)

This work was supported by the Fundamental Research Funds for the Central Universities under Grant N2319007.

ABSTRACT Many biological studies show that microRNAs (miRNAs) play an indispensable role in the regulation of various biological processes. MiRNAs are significant biomakers in disease diagnosis, aiding in the understanding of pathogenesis and facilitating the identification, diagnosis, and treatment of various diseases. However, the exact mechanism by which miRNAs influence the development of these diseases remains incompletely understood. Thus, it is crucial to develop a computational method to identify unknown miRNA-disease associations. In this study, we designed a computational framework based on singular value decomposition (SVD) and node2vec to predict unknown miRNA-disease associations (SNMDA). We use SVD technique to extract the linear features of miRNAs and diseases. The node2vec method is applied to learn the non-linear embeddings of miRNAs and diseases. We combine the linear feature and non-linear feature to get a new feature vector and feed it into the Gradient Boosting (GB) classifier for binary classification prediction. According to the experimental findings, SNMDA demonstrated an average area under the curve (AUC) of 0.9608 during five-fold cross-validation. Compared with the other Cutting-edge methods, SNMDA achieved the highest AUC value. Furthermore, the case studies on gastric cancer, malignant esophageal lesions, and lung tumors validate the effectiveness of SNMDA. The comprehensive experimental results demonstrate that SNMDA is effective in identifying unknown miRNA-disease associations.

INDEX TERMS Node2vec, singular value decomposition, miRNA-disease association prediction, linear feature, non-linear feature.

I. INTRODUCTION

MicroRNAs (miRNAs) are a class of non-coding RNAs (ncRNAs) that typically consist of 20-25 nucleotides in length. Their function involves post-transcriptional suppression of gene expression by binding to the 3' untranslated regions (UTRs) of target messenger RNAs [1], [2]. Since the discovery of the first miRNA, Lin-4, in C. elegans in 1993 [3], there has been extensive research demonstrating the involvement of miRNAs in a multitude of biological processes, such as cell proliferation [4], differentiation [5], viral infection [6], aging [7], among others. In addition, both overexpression and downregulation of miRNA expression in humans have been demonstrated to contribute to the development of a wide range of complex diseases [8], [9]. MiR-15 and miR-16, for instance, have a more pronounced impact on chronic lymphocytic leukemia (CLL) via controlling the antiapoptotic B-cell lymphoma protein BCL-2 in B cells [10]. The upregulation of miR17-5p expression has been shown to lead to increased pancreatic cancer cell proliferation and a significant increase in the number of invading cells [11]. When compared to normal oral tissue, abnormal expression of miRNAs such as miR-34b, miR-137, miR-193a, and miR-203 causes oral squamous cell carcinomas (OSCC) [12]. According to the aforementioned research, it has been demonstrated that miRNAs are strongly

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

linked as the progression of numerous complex human diseases. Therefore, it is imperative to uncover additional associations between miRNAs and diseases to gain insight into the pathogenesis of illnesses and significantly enhance the accuracy of disease diagnosis.

Traditional biological wet experimental methods used to investigate the potential miRNA-disease associations primarily encompass anchored polymerase chain reaction and reverse transcription polymerase chain reaction, etc [13]. In general, traditional biological wet experiments often confront quite a few bottlenecks, such as complicated experiments, time-consuming, pricey and the low recognition rate [14]. However, to address the limitations of traditional research methods, researchers have developed several trusted bioinformatic databases that store experimentally validated miRNA-disease associations. Simultaneously, with the rapid development of computer technology, numerous advanced computational methods have been proposed for predicting potential miRNA-disease associations. Computational methods are not only economical and effective, but also provide a new perspective for researchers to investigate the miRNAs ranked at the top and conduct relevant experiments to validate the potential associations predicted. According to previous studies summarized in [15], the existing computational methods can be broadly classified into traditional machine learning-based prediction models, deep learning-based prediction models, and matrix transformation-based prediction models. These computational methods are based on the assumption that functionally similar miRNAs are more likely to be allied with the phenotypically same or similar diseases and vice versa.

A. RELATED WORKS

Currently, some traditional traditional machine learning algorithms are used for modeling to predict the relationships between miRNA and diseases. These approaches used experimentally verified miRNA-disease associations, and unverified miRNA-disease pairs as training samples and then generate the relative training models. These models demonstrated good outcomes in predicting miRNA-disease associations. For example, Zhou et al. [16] developed a computational model, combining gradient boosting decision tree and logistic regression (GBDT-LR), to prioritize potential miRNAs involved in specific diseases. The model is capable of capturing non-linear features, which are then subjected to scoring through the implementation of logistic regression. Liu et al. [17] proposed a novel model called SMALF, which utilized stacked autoencoder to obtain latent feature from the original miRNA-disease association matrix and employed the XGBoost [18] algorithm to predict unknown relationships. Zhao et al. [19] utilized k-means clustering in data processing to balance the positive and negative sample and presented ABMDA implemented by boosting algorithm that iterates the weak classifier, decision tree, to improve the accuracy of classification to know the potential miRNA-disease interaction. You et al. [20] presented a model of path-based association prediction (PBMDA). They established a heterogeneous graph consisting of three interrelated subgraphs and then leveraged a Depth First Search algorithm (DFS) to predict potential miRNA-disease associations.

Meanwhile, deep learning algorithms are gradually utilized to observe potential relationships between miRNAs and diseases. In instance, Li et al. [21] employed a graph attention network to aggregate the neighbor information of nodes in each layer, and then fed the representation of the hidden layer into the structure-aware jumping knowledge network to obtain the global features of nodes. The output features of miRNAs and diseases are then concatenated and fed into a fully connected layer to score the potential associations. Li et al. [22] proposed a novel graph auto-encoder model (GAEMDA) that took the similarity between miRNAs and diseases as feature information, applied a graph neural networks-based encoder to produce the reduced-dimensional embeddings of miRNA and disease nodes. Finally, the embeddings of miRNA and disease nodes are inputted into a bilinear decoder, which is responsible for identifying potential links between miRNA and disease nodes. Wang et al. [15] introduced a base model that utilizes a multi-layer collaborative unsupervised training approach. They concatenated the low-dimensional representations learned by stacked graph autoencoder with the association features to derive the ultimate features for miRNA-disease pairs. Then, they used a multilayer perceptron (MLP) to predict scores for unknown miRNA-disease associations. Li et al. [23] proposed a novel deep learning model based on a hierarchical graph attention network for predicting miRNA-disease associations (HGANMDA), which wielded semantic-layer and node-layer attention to weight different importance of meta-paths for excavating unobserved interactions.

Furthermore, in recent years, several miRNA-disease association prediction algorithms based on matrix transformation have appeared. Zhong et al. [24] constructed a miRNA-disease double-layer network based on miRNA similarity, disease similarity and miRNA-disease association data, and then used the method of non-negative matrix factorization to rank the candidate miRNAs of diseases, so as to predict disease-related miRNAs. Cui et al. [25] considered that the lack of association information would have a negative impact on the prediction results, and used the k-Nearest Neighbor (KNN) method to preprocess the miRNA-disease matrix. Then, Collective Matrix Factorization (CMF) is used to infer potential miRNA-disease associations. The method adopts $L_{2,1} - norm$ to avoid over-fitting. The linear prediction method uses matrix factorization to map the miRNA-association matrix to a low-rank subspace to obtain linear features. According to the linear correlation between features, the miRNA-disease association is inferred. However, linear prediction methods only focus on the extraction of linear features, and cannot effectively extract nonlinear high-dimensional features in miRNA and disease

similarity data. It is difficult to make full use of the valuable feature information in the similarity data, and the prediction effect is not ideal for new diseases or diseases with incomplete association information. Machine learning methods can learn the rich intrinsic representation of data through nonlinear functions and can make more efficient use of the feature information of miRNA and diseases. However, they only focus on the nonlinear feature learning of miRNA and diseases, do not consider the linear features in the associated data, and ignore the fusion of linear features and nonlinear features. It is not conducive to potential miRNA-disease association prediction. Ding et al. [26] used NMF and variational graph autoencoder (VGAE) to fuse linear and nonlinear features of miRNA and disease, and predicted miRNA-disease association on the basis of association matrix, and achieved good results.

B. METHOD OVERVIEW

In order to explore the complex potential factors hidden under the miRNA-disease association matrix and make full use of the similarity information between miRNA and disease, we design an integrated feature extraction model that combines SVD and node2vec to predict potential miRNA-disease associations. Specifically, the main contributions of our study included the following parts:

1) We integrated both linear features and non-linear features to construct the final features and could better learn the potential information in miRNA-disease pairs.
2) We proposed an integrated feature extraction model prediction framework that combined SVD and Node2Vec. SVD method can obtain linear features by mapping miRNA-disease association matrix to bottom subspace, so as to dig deep correlation information between miRNA and disease. Node2vec method can obtain non-linear features by automatically learning miRNA and disease similarity information and achieve efficient prediction of miRNA-disease association.
3) We used Gradient Boosting (GB) for prediction of the potential miRNA-disease associations eventually, which demonstrates a high level of fault tolerance and is capable of swiftly and efficiently learning feature information from miRNA-disease pairs. As a result, it significantly enhances the prediction performance of the model.

We used five-fold cross validation to evaluate accuracy of the SNMDA(Singular value decomposition and Node2vec MiRNA-Disease Associations), which obtained AUC(Area Under the Curve) of 0.9608. Case studies on gastric cancer, lung cancer and esophageal neoplasms were also carried out to prove the prediction ability of the model. Consequently, most of the predicted miRNAs associated with these diseases were verified by mir2disease database [27] and dbDEMC v2.0 database [28]. In conclusion, SNMDA can efficiently predict potential miRNA-disease association. In addition, the source codes can be found at https://github.com/xiaoyanzi1124/SNMDA.git.

II. MATERIALS AND METHODS

A. BENCHMARK DATASET

In the present investigation, we retrieved miRNA-disease associations from the HMDD v2.0 database. Additionally, we directly downloaded experimentally verified miRNA-disease associations from https://www.cuilab.cn/hmdd [29]. After eliminating duplications and inconsistent entries, we obtained 5430 experimentally validated miRNA-disease associations, covering 495 miRNAs and 383 diseases. On this basis, we created a binary matrix, $A \in \mathbb{R}^{I \times J}$, where I and J represent the number of miRNAs and diseases, respectively. The matrix mapping the associations between miRNAs and diseases can be mathematically defined as follows:

$$A(i,j) = \begin{cases} 1, & \text{if miRNA } i \text{ is associated to disease } j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

In addition, matrix A is a sparse matrix, with an associated density of 0.0286. A miRNA is associated with a maximum of 125 diseases and a minimum of 1 disease, and on average one miRNA is associated with 11 diseases. A disease is associated with a maximum of 213 miRNAs and a minimum of 0 miRNAs, and on average a disease is associated with 14 miRNAs. There were 495 miRNAs associated with diseases and 384 diseases associated with miRNAs [30], [31].

B. SIMILARITY NETWORKS

1) DISEASE SEMANTIC SIMILARITY

According to earlier research [32], similarity in disease semantics can be calculated via utilizing the arborescence attribute of disease in MeSH database, which is obtained from https://www.nlm.nih.gov/mesh. Based on the MeSH database, we marked each disease node using the Directed Acyclic Graph (DAG). Consequently, we can adopt $DAG(d_i) = (d_i, T(d_i), E(d_i))$ to describe the given disease d_i, where $T(d_i)$ represents the ancestor nodes of d_i and $E(d_i)$ denotes all the edges that directly connect ancestor nodes to descendant nodes. Later, based on the above information, we can compute the semantic function of disease d_k to d_i as shown below:

$$D1_{d_i}(d_k) = \begin{cases} 1 & \text{if } d_k = d_i \\ \max\left\{\Delta * D1_{d_i}\left(d'_k\right) \mid d'_k \in \text{children of } d_k\right\} & \text{if } d_k \neq d_i \end{cases} \tag{2}$$

where Δ is the semantic contribution factor, which is assigned a value of 0.5 based on a earlier research [33]. The contribution score of disease d_i to itself is 1, and the contribution score of disease d_k to disease d_i will decrease as the distance. Hence the contribution score of disease d_i can be defined as below:

$$DS1(d_i) = \sum_{d_k \in T(d_i)} D1_{d_i}(d_k) \tag{3}$$

If two diseases share greater parts of their DAGs, two diseases can be considered more similar. We constructed a 383×383 matrix $SS1$ to store the initial type of disease semantic similarity. We can compute the similarity score for disease d_i to d_j as follows:

$$SS1\left(d_i, d_j\right) = \frac{\sum_{d_t \in T(d_i) \cap T(d_j)} \left(D1_{d_i}\left(d_t\right) + D1_{d_j}\left(d_t\right)\right)}{DS1\left(d_i\right) + DS1\left(d_j\right)} \quad (4)$$

However, the aforementioned calculation method has a drawback, namely, it does not take into account the different contributions of two diseases on the same layer of the DAG, and diseases with low frequency should contribute more than those with high frequency. We obtain the similarity between two diseases by using another method of calculating the semantic similarity of diseases [34]. Then, we can describe the semantic impact of disease d_k to d_i as follows:

$$D2_{d_i}\left(d_k\right) = -\log\left(\frac{\text{the number of DAGs including } d_k}{\text{the number of diseases}}\right) \quad (5)$$

Correspondingly, the second kind of semantic significance of disease d_i can be described using

$$DS2\left(d_i\right) = \sum_{d_k \in T(d_i)} D2_{d_i}\left(d_k\right) \quad (6)$$

and the similarity in disease semantics $SS2(d_i, d_j)$ between disease d_i and d_j can be computed as:

$$SS2\left(d_i, d_j\right) = \frac{\sum_{d_t \in T(d_i) \cap T(d_j)} \left(D2_{d_i}\left(d_t\right) + D2_{d_j}\left(d_t\right)\right)}{DS2\left(d_i\right) + DS2\left(d_j\right)} \quad (7)$$

where $SS2$ is a matrix with dimensions of 383×383 to store the semantic similarity values for the second type of diseases.

In order to calculate the semantic similarity between diseases, we combine the above two aspects of information to compute the final disease semantic similarity based on previous study [35]. The disease semantic similarity between diseases d_i and d_j can be computed using the following formula:

$$SS\left(d_i, d_j\right) = \frac{SS1\left(d_i, d_j\right) + SS2\left(d_i, d_j\right)}{2} \quad (8)$$

2) MIRNA FUNCTIONAL SIMILARITY

According to the hypothesis of Wang et al. [33], the theoretical basis for miRNA functional similarity is that diseases exhibiting similar phenotypes are more likely to be associated with miRNAs that have similar functions, and conversely. We can download miRNA functional similarity information from https://www.cuilab.cn/files/images/cuilab/misim.zip [33]. Then, we constructed the matrix FS to store the miRNA functional similarity for the convenience and efficiency of subsequent calculation, where the element $FS(m_i, m_j)$ denotes the miRNA functional similarity score between miRNA m_i and m_j.The range of miRNA functional similarity values is [0, 1]. The calculation process will involve the semantic similarity of known associations and diseases. The calculation formula is as follows:

$$S(dt, D\,T) = \max_{1 \le i \le k}\left(SS\left(dt, d\ t_i\right)\right) \quad (9)$$

where DT is a set of k diseases, and dt represents a single disease. This formula can determine the maximum similarity value between disease dt and the diseases in DT. The functional similarity of miRNAs is calculated based on the following formula:

$$FS\left(m_1, m_2\right) = \frac{\sum_{1 \le i \le m} S\left(dt_{1i}, D\,T_2\right) + \sum_{1 \le j \le n} S\left(dt_{2j}, D\,T_1\right)}{m + n} \quad (10)$$

where m_1 and m_2 represent two miRNAs, DT_1 represents the set of diseases associated with m_1, and DT_2 represents the set of diseases associated with m_2. DT_1 contains m diseases, and DT_2 contains n diseases. After calculating the similarity of each pair of miRNAs, the matrix representing the functional similarity of miRNAs is denoted by FS.

3) GAUSSIAN INTERACTION PROFILE KERNEL SIMILARITY FOR MIRNAS AND DISEASES

Matrix of miRNA functional similarity and matrix of disease semantic similarity have a large number of sparse values. To complement the similarity information pertaining to miRNAs and diseases, we can compute the similarity based on the Gaussian interaction profile (GIP) kernel to represent miRNA similarity and disease similarity based on a assumption that miRNAs with similar characteristics are more likely to be associated with similar diseases [35]. Specifically, we use a binary vector $IP(m_i)$ to represent associations between miRNA m_i and each disease, which is located in the ith column of matrix A. Specifically, the Gaussian kernel similarity for miRNA $KM(m_i, m_j)$ between miRNA m_i and m_j can be defined as:

$$\mathrm{KM}\left(m_i, m_j\right) = \exp\left(-\gamma_m \left\|IP\left(m_i\right) - IP\left(m_j\right)\right\|^2\right) \quad (11)$$

where the adjustment parameter γ_m is utilized to regulate the bandwidth of the kernel. Its value can be determined using the formula provided below:

$$\gamma_m = \gamma_m' \Big/ \left(\frac{1}{nm}\sum_{i=1}^{nm}\left\|IP\left(m_i\right)\right\|^2\right) \quad (12)$$

where γ_m' denotes the normalizing original kernel bandwidth that is set to 1 referring to the prior research [35]. In addition, nm nm denotes the total count of miRNAs, which is equivalent to 495 in our study. In the same manner, the Gaussian kernel similarity for disease $KD(d_i, d_j)$ between disease d_i and d_j can be calculated according to the following two equations:

$$\mathrm{KD}\left(d_i, d_j\right) = \exp\left(-r_d \left\|\mathrm{IP}\left(d_i\right) - \mathrm{IP}\left(d_j\right)\right\|^2\right) \quad (13)$$

$$r_d = r_d' \Big/ \left(\frac{1}{nd}\sum_{i=1}^{nd}\left\|\mathrm{IP}\left(d_i\right)\right\|^2\right) \quad (14)$$

where binary vector $IP(d_i)$, located in the ith row of matrix A, is constructed to describe the associations between disease di and each individual miRNA. In our study, nd refers to the general amount of diseases, which is 383. Moreover, we have set the value of γ'_d to 1.

4) INTEGRATED SIMILARITY FOR MIRNAS AND DISEASES

To obtain the comprehensive miRNA similarity network, we combined the miRNA functional similarity FS and the miRNA Gaussian interaction kernel similarity KM. On account of the previous research [35], we describe the aggregated similarity for miRNAs $SM(m_i, m_j)$ between miRNA m_i and m_j as follows:

$$SM\left(m_i, m_j\right) = \begin{cases} FS\left(m_i, m_j\right) & \text{if } m_i \text{ and } m_j \text{ have functional similarity} \\ KM\left(m_i, m_j\right) & \text{otherwise} \end{cases} \tag{15}$$

Similarly, the aggregated similarity for diseases $SD(d_i, d_j)$ between disease d_i and d_j can be defined as follows:

$$SD\left(d_i, d_j\right) = \begin{cases} SS\left(d_i, d_j\right) & \text{if } d_i \text{ and } d_j \text{ have semantic similarity} \\ KD\left(d_i, d_j\right) & \text{otherwise} \end{cases} \tag{16}$$

C. OVERVIEW OF SNMDA

To forecast the potential association between miRNAs and diseases, we propose a prediction model that combines the SVD method and the node2vec method (SNMDA). We employ the SVD method to obtain linear features of miRNA and disease and apply the node2vec method to obtain non-linear features of miRNA and disease. By combining different features of each node, we construct a comprehensive feature vector that fuses linear features with interaction information and non-linear features with similarity information. The classifier is trained using these integrated feature vectors, and subsequently, the corresponding prediction score is generated. SNMDA can be outlined as a series of five steps (see Figure 1):

Step 1: Data processing and construction of miRNA similarity matrix SM, disease similarity matrix SD and miRNA-disease association matrix A.

Step 2: Adopt the SVD method to obtain the linear vectors of miRNA and disease entities on the miRNA-disease interaction matrix A. The linear representations are only based on the miRNA-disease association adjacency matrix A.

Step 3: Exploit the node2vec method to obtain the non-linear vectors of miRNA and disease entities from comprehensive similarity networks SM and SD, respectively.

Step 4: Combination of linear and non-linear features.

Step 5: Use GB classifier to predict miRNA-disease associations that are currently unknown.

1) LINEAR REPRESENTATIONS BY SVD

SVD is a mathematical technique based on linear algebra, used for processing and analyzing matrix data [36]. It achieves this by decomposing the original matrix into the product of three specific matrices: the left singular vector matrix, the diagonal singular value matrix, and the transpose of the right singular vector matrix. In this process, the features extracted by SVD are essentially linear transformations of the original data, as they are represented by the product of the original features with the left singular vectors, the transpose of the right singular vectors, and the singular values. These features are considered linear because they adhere to the principles of linear algebra, where any feature can be expressed as a weighted sum of other features. SVD does not involve any nonlinear operations, so it does not alter the linear properties of the data. Moreover, SVD reduces the dimensionality of the data by retaining the feature vectors corresponding to the largest singular values, which represent the main directions of variation in the data, and these directions are typically linear. Thus, we use the SVD method for capturing linear features of miRNAs and diseases. In SVD, the miRNA-disease association matrix is represented as $A \in \mathbb{R}^{m \times n}$, and the matrix A is a factorization of three matrices as follows:

$$A = U \Sigma V^T \tag{17}$$

where $U \in \mathbb{R}^{m \times m}$ represent miRNA feature matrix, $\Sigma \in \mathbb{R}^{m \times n}$ represent feature weight matrix, and $V^T \in \mathbb{R}^{n \times n}$ represent disease feature matrix. Among the obtained matrices, U is a real matrix, Σ is a diagonal matrix with non-negative square roots of the eigenvalues of the product $A^T A$ on the diagonal, and V^T a real matrix. The diagonal elements λ_i are called singular values of matrix A.

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0 \tag{18}$$

In the singular value diagonal matrix Σ, the singular values are arranged from largest to smallest in terms of importance. Typically, the singular values decrease rapidly as their indices increase. Based on the distribution of singular values, one can identify the point where they begin to drop sharply. This point can serve as the threshold k. This threshold is not fixed and varies depending on the dataset. By retaining the k largest singular values, smaller singular values can be disregarded, reducing the impact of noise and thereby enhancing the quality of data representation. Moreover, retaining the k largest singular values can reduce the dimensionality of features, preserving the most important features in the dataset, significantly reducing computational load, and making the algorithm faster and easier to handle. In summary, we only take the top k features with the largest values in matrix Σ, and A can be re-described as:

$$A_{m \times n} \approx U_{m \times k} \cdot \Sigma_{k \times k} \cdot V^T_{k \times n} \tag{19}$$

Figure 2 shows an approximate SVD representation of the interaction matrix A. According to the SVD principle, the U_i

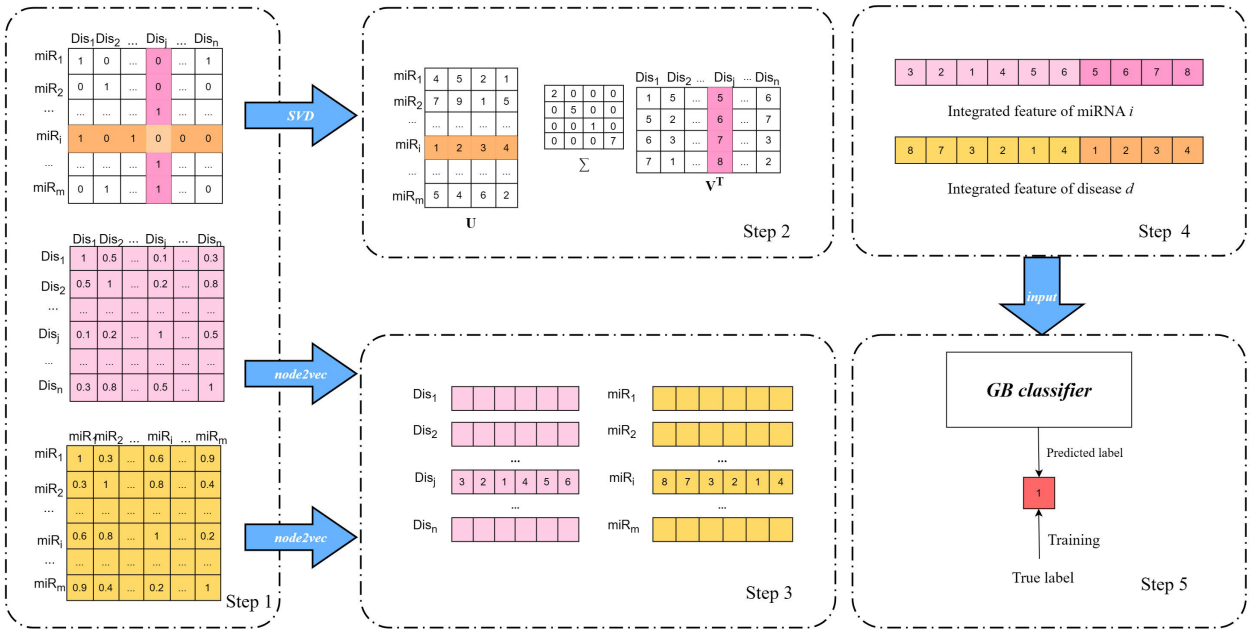


FIGURE 1. Overview of SNMDA (miR stands for miRNA and Dis stands for disease).

and V_j^T represent linear features of miRNA i and disease j, respectively.

2) NON-LINEAR REPRESENTATIONS BY NODE2VEC

Graph representation learning, also known as graph embedding, aims to represent a graph as a specific vector by mapping nodes or the entire graph into a low-dimensional space [37]. The goal is to optimize this mapping so that the vector relationships in the embedding space capture the original structure of the graph. The learned embedding vectors can be used as input feature values for various machine learning tasks [38].

In our study, we applied the node2vec algorithm to extract features from miRNA similarity network and disease similarity network [39]. Node2vec is a random walk-based graph embedding algorithm based on word2vec [40]. In the process of learning the network topology, node2vec integrates two neighborhood sampling strategies, Breadth First Search (BFS) and Depth First Search (DFS).

Let the previous node be t. Consider a random walk that just traversed edge (t, v) and is currently located at node v. The transition probabilities π_{vx} on edge (t, v) leading from v should be evaluated. The transition probability, without normalization, is $\pi_{vx} = \alpha_{pq} \cdot (W_{vx})$, where

$$\alpha_{pq}(t,x) = \begin{cases} \frac{1}{p} \text{ if } d_{tx} = 0 \\ 1 \text{ if } d_{tx} = 1 \\ \frac{1}{q} \text{ if } d_{tx} = 2 \end{cases} \tag{20}$$

Here d_{tx} represents the shortest path distance between nodes t and x. W_{vx} represents the edge weight between node v and node x, and in this study, it signifies the degree of similarity between two diseases or the degree of similarity between two miRNAs.

As shown in Figure 3, the parameter p controls the likelihood of immediately revisiting a node during the walk. A higher value of p means that the random walk is more inclined to retrace along short paths in the graph, that is, it is more likely to return to nodes that have been visited before. This setting helps the algorithm to deeply explore local areas within the graph rather than broadly exploring the entire network. A lower value of p means that the random walk retraces less, that is, it returns less frequently to previously visited nodes. This setup encourages the random walk to explore new nodes and paths, which helps to uncover the global structure and long-distance relationships in the graph. A low p value contributes to increasing the breadth of the walk, but it may lead to a walk that is too random, lacking in-depth exploration of local structures. The parameter q controls the likelihood of searching "local" or "global" nodes during the random walk. If q>1, the random walk has a greater probability to sample nodes around the node v, which is a BFS based sampling. In contrast, if q<1, the random walk has a greater probability to sample nodes far away from v, which can get more global features information. According to the recommendations in the node2vec paper [39], the parameters p and q usually take empirical values within a certain range, such as $\{0.25, 0.50, 1, 2, 4\}$. However, in theory, parameters p and q can take any positive real value, as long as they satisfy the requirements of specific application scenarios.

3) FEATURE COMBINATION

So far, we have obtained the linear feature matrixes U, V^T based on the decomposition of $A_{M\times N}$, and the non-linear feature representations of miRNA and disease nodes using

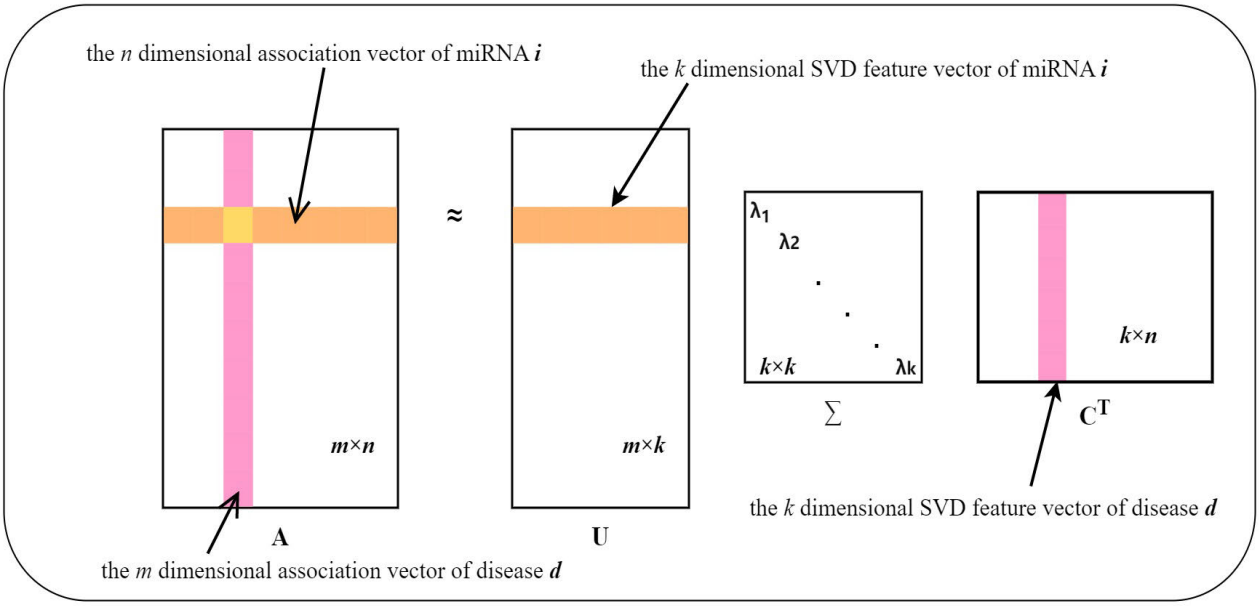


FIGURE 2. The illustration of applying SVD on miRNA-disease association matrix.

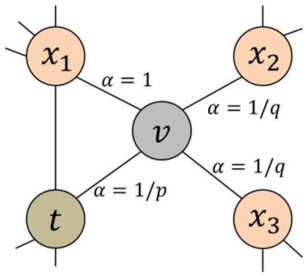


FIGURE 3. Illustration of the random walk procedure in node2vec. After transitioning from node *t* to node *v*, the random walk is currently assessing its next step from node *v*. The edge labels denote search biases represented by α.

Network Representation Learning (NRL) method node2vec. The next question is how can we combine them into a computational framework in the prediction of miRNAs and diseases. For each miRNA i and disease j, the feature integration rules are as follows: The linear features corresponding to miRNA i is the ith row of U, which is noted as $LM_i(U_i)$ after being converted into a column vector. Similarly, the linear features corresponding to disease j is the jth column of V^T, represented as $LD_j(V_j^T)$. The non-linear features corresponding to miRNA i are noted as NM_i as well as the non-linear features corresponding to disease j are noted as ND_j. The final integrated features of i and j are expressed as:

$$miRNA_{new} = concatenating(LM_i, NM_i) \quad (21)$$

$$Dis_{new} = concatenating(LD_j, ND_j) \quad (22)$$

where *concatenating*() represents the concatenation operation. Then, concatenate the two vectors to get a new vector for model prediction.

$$Vec_{new} = concatenating(miRNA_{new}, Dis_{new}) \quad (23)$$

4) CLASSIFIER PREDICTION

Based on the experimental outcomes from classifier selection, we utilize the GB classifier for association prediction. Our research objective is to predict the unknown miRNA-disease associations in the adjacency matrix A. The GB classifier is adept at addressing binary classification issues. It employs an ensemble learning approach, progressively assembling a strong learner by integrating multiple weak learners. Within the gradient boosting algorithm, each weak learner is trained to address residuals, implying that each must compensate for the predictive errors made by all preceding weak learners. Consequently, as training progresses, the influence of each weak learner incrementally intensifies, culminating in enhanced predictive accuracy. In Python, the *GradientBoostingClassifier* class encapsulates the gradient boosting algorithm and offers tunable parameters such as the learning rate and the quantity of trees, enabling control over the model's complexity and efficacy. Typically, the gradient boosting algorithm boasts high predictive precision and robustness, which has led to its broad adoption in practical applications. In the GB classifier, we set the number of weak learners (*n_estimators*) to 100, and the contribution of each weak learner (*learning_rate*) to 0.1. All other parameters are set to their default values.

5) CASE STUDY

We selected all known miRNA-disease associations and randomly chose an equal number of unknown miRNA-disease associations as the training set to train the SNMDA model on the HMDD v2.0 database. Then, the well-trained model was used to identify all unknown associations. After obtaining all predicted scores for each pair of unknown miRNA-disease associations, for a specific disease, we ranked the predicted

association scores for all related miRNAs. MiRNAs with higher rankings are likely to be associated with that specific disease. Besides, we performed a validation process for the top 50 candidate miRNAs predicted results one by one according to dbDEMC [28] and mir2disease [27] databases, respectively.

D. EVALUATION METRICS

To ensure impartial comparisons, we adopt the 5-fold cross-validation to evaluate the performance of SNMDA. More precisely, we randomly divided all the samples (all known miRNA-disease associations as positive samples and an equal count of unknown miRNA-disease associations as negative samples) into five equal parts, each part was used as a test set and the other four parts in turn as a training set. The receiver-operating characteristics (ROC) curves were plotted based on the results of 5-fold cross-validation. The x axis and y axis of the ROC curves represent false positives rate (FPR) and true positives rate (TPR), respectively. FPR and TPR can be calculated by the following formulas:

$$FPR = \frac{FP}{TN + FP} \tag{24}$$

$$TPR = \frac{TP}{TP + FN} \tag{25}$$

where TP and TN are the numbers of miRNA-disease association pairs and non-association pairs which are correctly identified, respectively; FP and FN are the numbers of miRNA-disease association pairs and non-association pairs which are incorrectly identified, respectively. The AUC value is the area under the ROC curve, and its value is between 0 and 1. In general, the higher the AUC value, the better the performance of the model. In our study, we mainly use the area under the curve (AUC) and the area under precision-recall curve (AUPR) to evaluate the overall performance of our models. In classification problems, AUC is an important method to evaluate the model's overall performance, and for unbalanced data sets, AUPR is more suitable for evaluation models than AUC. Moreover, for a more comprehensive evaluation of model performance, we also used four common performance measures to measure the performance of SNMDA, such as $Accuracy$, $Precision$, $Recall$, and $F1 - score$. Four metrics are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{26}$$

$$Precision = \frac{TP}{TP + FP} \tag{27}$$

$$Recall = \frac{TP}{TP + FN} \tag{28}$$

$$F1 - score = \frac{2 \times Pre \times Rec}{Pre + Rec} \tag{29}$$

where TP, TN, FP, FN denote true positive, true negative, false positive and false negative, respectively.

III. RESULTS

In this section, we present all the research results from our experiments, including parameter selection, model performance, comparison with five state-of-the-art methods, and case studies. These five state-of-the-art models are GAEMDA, SMALF, HGANMDA, GBDT-LR, ABMDA, and introductions to these models are shown in the related work section. Additionally, we also display the results of the ablation study: the results based solely on the SVD method and the results based solely on the Node2vec method.

A. CLASSIFIER SELECTION AND PARAMETER TUNING

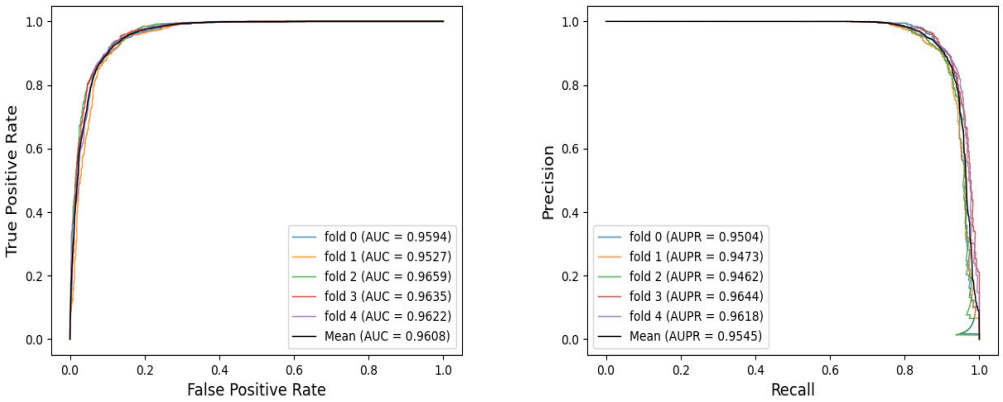
In our experiments, we implemented the SNMDA model based on the Numpy library, node2vec library and the scikit-learn framework. After gaining the linear feature matrixes U and V^T based on SVD, we found a huge decay gap from 10^{-2} to 10^{-15} between the 291rd and the 292rd dimensions of the importance matrix Σ. According to the SVD principle, the linear characteristics of entities are mainly concentrated in the first 291 dimensions. Therefore, the linear feature vectors of miRNA and disease were fixed to 291 dimensions. In the non-linear feature vectors section, we used the same parameters as the node2vec paper [39]. We set the dimensions of feature vector as 16, 32, 64 and 128, respectively. The optimal dimension is found as 128 in the following experiment. Since p is small and q is large, a random walk is formed in the BFS method to obtain an embedding value containing local rather than global information. Since miRNA similarity network and disease similarity network in our study are small and dense networks, according to previous study [41], in small and dense networks, it is recommended to use a higher parameter q, in which case BFS is superior to DFS. According to the open-source paper on node2vec [39], the range of parameters p and q is $\{0.25, 0.50, 1, 2, 4\}$. In the training process, the parameters p and q were set to 0.25 and 4 for efficient clustering.

In the selection process of machine learning classifiers, XGBoost (XGB), Logistic regression (LR), Naive Bayes (NB), Random forest (RF), AdaBoost (ADB), Gradient Boosting(GB), and Multilayer perception (MLP) were tested based on every integrated features, respectively. The results of AUC values of all classifiers are shown in Table 1. The "SVD" column represents the features extracted using only the SVD method. Similarly, the "N2V16" column represents the 16-dimensional features extracted using only the node2vec method. "SN2V16" represents the integrated features that combine SVD features with 16-dimensional node2vec vectors, and so on. All of the above classifiers are imported from the scikit-learn library and implemented in Python with all internal classifier parameters set to default values.

According to the data presented in Table 1, the prediction outcomes based on the integrated feature are superior to the single linear feature prediction results and the single non-linear feature prediction results in most classifiers. And,

TABLE 1. The AUC results of different features on classifiers.

	SVD	N2V16	SN2V16	N2V32	SN2V32	N2V64	SN2V64	N2V128	SN2V128
XGB	0.9382	0.9454	**0.9543**	0.9408	0.9501	0.9463	0.9527	0.9504	0.9521
LR	0.8773	0.9457	0.9526	0.9356	0.9492	0.9432	0.9519	0.9530	**0.9551**
GaussianNB	0.8214	**0.9273**	0.8397	0.9143	0.8441	0.9115	0.8599	0.9078	0.8862
RF	0.9290	0.9400	0.9464	0.9317	0.9429	0.9354	0.9433	0.9430	**0.9469**
AdaBoost	0.9273	0.9431	0.9458	0.9355	0.9408	0.9389	0.9435	0.9441	**0.9477**
GB	0.9506	0.9516	0.9607	0.9478	0.9593	0.9519	0.9591	0.9547	**0.9608**
MLP	0.9158	**0.9489**	0.9450	0.9432	0.9442	0.9416	0.9446	0.9425	0.9447



(a) ROC curve (b) PR curve

FIGURE 4. The 5-fold cross-validated ROC curve and PR curve of SNMDA model with AUC of 96.08% and AUPR of 95.45%.

TABLE 2. 5-fold cross-validation results performed by SNMDA.

Test set	Accuracy	Recall	F1-score	Precision
1	0.8959	0.9088	0.8973	0.8860
2	0.8950	0.9024	0.8958	0.8893
3	0.9015	0.9006	0.9014	0.9022
4	0.9052	0.9245	0.9070	0.8901
5	0.9070	0.9245	0.9086	0.8932
Mean	0.9009	0.9122	0.9020	0.8922

the combination of linear features and 128-dimensional node2vec features obtained the optimal classification results in the GB classifier. In addition, the best results for each classifier are shown in bold. Therefore, in the subsequent experiments, we choose to combine "SN2V128" with the GB classifier to perform the experiments.

B. PREDICTION MIRNA-DISEASE ASSOCIATIONS BY SNMDA

To obtain reliable experimental results of the model, we use five-fold cross-validation based on "SN2V128" to evaluate the performance of SNMDA. From Figure 4, we can see that AUCs of SNMDA are 0.9594, 0.9527, 0.9659, 0.9635, 0.9622, respectively. The average AUC value is 0.9608. From Figure 4(b), we can see that AUPRs of SNMDA are 0.9504, 0.9473, 0.9462, 0.9644, 0.9618, respectively. The average AUPR value is 0.9545. The results show that SNMDA exhibits strong performance in revealing unknown miRNA-disease associations.

Table 2 presents a detailed overview of the average results of various evaluation metrics for our model, obtained through 5-fold cross-validation. We observe that the average accuracy, recall, f1-score, and precision of SNMDA at 5-fold cross-validation are 0.9009, 0.9122, 0.9020, 0.8922. It is further proved that the SNMDA model is effective for association prediction.

C. PERFORMANCE COMPARISON

To further demonstrate the predictive ability of SNMDA, we compared SNMDA with seven state-of-the-art existing computational methods, which are GAEMDA [22], SMALF [17], HGANMDA [23], GBDT-LR [16], ABMDA [19], SVD-MDA and NODE2VEC-MDA. The parameters involved in all comparison models are consistent with those in the source paper. All the algorithms are implemented and ensure to run in the same computer environment. To ensure an impartial assessment, all the aforementioned models were evaluated using five-fold cross-validation on the HMDDv2.0 database. The contrast outcomes were condensed and presented in Table 3. We can see that SNMDA achieves the highest values on all evaluation metrics in between these seven models. In addition, in terms of AUC values, which can provide a more comprehensive evaluation of the model's performance, SNMDA yields superior outcomes than the other models, and 1.05% higher than the second highest SMALF model. The experimental

TABLE 3. Comparison table of each evaluation metric for different models.

methods	AUC mean	AUPR mean	Accuracy	Precision	Recall	F1-score
ABMDA	0.8871	0.8608	0.8277	0.8237	0.8301	0.8267
GBDT-LR	0.9302	0.9262	0.8571	0.8507	0.8665	0.8585
HGANMDA	0.9346	0.9310	0.8579	0.8751	0.8354	0.8545
GAEMDA	0.9350	0.9328	0.8516	0.8172	0.9080	0.8596
SMALF	0.9503	0.9472	0.8860	0.8809	0.8932	0.8869
SVD-MDA	0.9506	0.9468	0.8829	0.8731	0.8960	0.8844
NODE2VEC-MDA	0.9547	0.9502	0.8933	0.8814	0.9090	0.8949
SNMDA	**0.9608**	**0.9545**	**0.9009**	**0.8922**	**0.9122**	**0.9020**

TABLE 4. The top 50 predicted miRNAs which may be associated with gastric cancer.

Rank	miRNA	Evidence	Rank	miRNA	Evidence
1	hsa-mir-21	dbDEMC, mir2disease	26	hsa-mir-31	dbDEMC
2	hsa-mir-126	dbDEMC	27	hsa-mir-132	dbDEMC
3	hsa-mir-155	dbDEMC	28	hsa-mir-210	dbDEMC
4	hsa-mir-133a	dbDEMC	29	hsa-mir-146b	unconfirmed
5	hsa-mir-34a	dbDEMC	30	hsa-mir-24	dbDEMC
6	hsa-mir-146a	dbDEMC	31	hsa-mir-106b	dbDEMC, mir2disease
7	hsa-mir-1	dbDEMC	32	hsa-mir-221	dbDEMC, mir2disease
8	hsa-mir-16	dbDEMC, mir2disease	33	hsa-mir-182	dbDEMC
9	hsa-mir-29a	unconfirmed	34	hsa-mir-92a	dbDEMC
10	hsa-mir-143	dbDEMC, mir2disease	35	hsa-mir-195	dbDEMC
11	hsa-mir-29b	dbDEMC	36	hsa-let-7i	dbDEMC
12	hsa-mir-214	dbDEMC, mir2disease	37	hsa-let-7d	dbDEMC
13	hsa-mir-125b	dbDEMC	38	hsa-let-7b	dbDEMC
14	hsa-let-7c	dbDEMC	39	hsa-mir-145	dbDEMC, mir2disease
15	hsa-mir-223	dbDEMC, mir2disease	40	hsa-mir-9	unconfirmed
16	hsa-mir-20a	dbDEMC	41	hsa-mir-15a	dbDEMC
17	hsa-mir-17	dbDEMC	42	hsa-mir-141	dbDEMC, mir2disease
18	hsa-mir-19b	dbDEMC	43	hsa-mir-30a	dbDEMC
19	hsa-mir-222	dbDEMC, mir2disease	44	hsa-mir-125a	dbDEMC
20	hsa-mir-150	dbDEMC, mir2disease	45	hsa-mir-192	dbDEMC
21	hsa-mir-196a	dbDEMC	46	hsa-mir-200b	dbDEMC
22	hsa-let-7e	dbDEMC	47	hsa-let-7f	dbDEMC
23	hsa-mir-122	dbDEMC	48	hsa-mir-199a	dbDEMC
24	hsa-mir-26a	dbDEMC	49	hsa-mir-200a	dbDEMC
25	hsa-mir-18a	dbDEMC	50	hsa-mir-29c	dbDEMC

TABLE 5. The top 50 predicted miRNAs which may be associated with esophageal neoplasms.

Rank	miRNA	Evidence	Rank	miRNA	Evidence
1	hsa-mir-16	dbDEMC	26	hsa-mir-125a	dbDEMC
2	hsa-mir-1	dbDEMC	27	hsa-mir-107	dbDEMC, mir2disease
3	hsa-mir-9	dbDEMC	28	hsa-mir-212	dbDEMC
4	hsa-let-7i	dbDEMC	29	hsa-mir-24	dbDEMC
5	hsa-mir-18a	dbDEMC	30	hsa-mir-122	dbDEMC
6	hsa-mir-19b	dbDEMC	31	hsa-mir-30a	dbDEMC
7	hsa-mir-29a	dbDEMC	32	hsa-mir-23b	dbDEMC
8	hsa-let-7e	dbDEMC	33	hsa-mir-224	dbDEMC
9	hsa-mir-96	dbDEMC	34	hsa-mir-30d	dbDEMC
10	hsa-mir-125b	dbDEMC	35	hsa-mir-424	dbDEMC
11	hsa-mir-142	dbDEMC	36	hsa-mir-106b	dbDEMC
12	hsa-mir-199b	dbDEMC	37	hsa-mir-137	dbDEMC
13	hsa-mir-195	dbDEMC	38	hsa-mir-191	dbDEMC
14	hsa-mir-17	dbDEMC	39	hsa-mir-328	dbDEMC
15	hsa-mir-23a	dbDEMC	40	hsa-mir-181a	dbDEMC
16	hsa-mir-146b	dbDEMC	41	hsa-mir-20b	dbDEMC
17	hsa-mir-18b	dbDEMC	42	hsa-mir-181b	dbDEMC
18	hsa-mir-29b	dbDEMC	43	hsa-mir-197	dbDEMC
19	hsa-mir-7	dbDEMC	44	hsa-mir-26b	dbDEMC
20	hsa-mir-373	dbDEMC, mir2disease	45	hsa-mir-219	dbDEMC
21	hsa-mir-206	dbDEMC	46	hsa-mir-335	dbDEMC
22	hsa-mir-200b	dbDEMC	47	hsa-mir-222	dbDEMC
23	hsa-let-7f	dbDEMC	48	hsa-mir-221	dbDEMC
24	hsa-mir-103a	dbDEMC	49	hsa-mir-30c	dbDEMC
25	hsa-let-7g	dbDEMC	50	hsa-mir-106a	dbDEMC

results show that SNMDA model is effective in predicting miRNA-disease association.

D. CASE STUDIES BY SNMDA

To further assess the practicality and effectiveness of SNMDA in real-world scenarios, three case studies on gastric cancer, esophageal neoplasms, and lung neoplasms were provided.

Gastric cancer is the fourth most common cancer worldwide, and it is one of the common malignant tumors. Its incidence ranking is the first among all kinds of tumors in China [42]. So in our initial case study, we devised a methodology to prioritize miRNAs that may have a potential association with gastric cancer. The results are shown in Table 4. Among the top 50 miRNAs associated with gastric cancer, 47 of them are confirmed by dbDEMC or mir2disease. In summary, 98% of the top 50 predicted novel miRNAs associated with gastric cancer were verified, which further demonstrates the effectiveness of SNMDA in predicting miRNA-disease associations.

Table 5 lists the top 50 esophageal neoplasms-associated miRNAs. The esophageal tumor is a malignant growth that develops in the tissues of the esophagus, and its etiology is associated with chronic exposure to nitrosamines, inflammation, and the levels of trace elements found in common food sources [43]. Here, we chose esophageal neoplasms as our second case study. All 50 of the predicted miRNAs associated with esophageal neoplasms are confirmed by dbDEMC or mir2disease, with a prediction accuracy of 100%. This indicates that the SNMDA model presented in this paper can predict the potential associations between unknown diseases and miRNAs without any known miRNA associations.

Lung cancer is the leading cause of cancer occurrence globally [44]. It is a malignant tumor originating from the bronchial mucosa or glands of the lung, which is one of the malignant tumors with the fastest growing morbidity and mortality and poses the greatest threat to human health

TABLE 6. The top 50 predicted miRNAs which may be associated with lung cancer.

Rank	miRNA	Evidence	Rank	miRNA	Evidence
1	hsa-mir-195	dbDEMC, mir2disease	26	hsa-mir-184	dbDEMC
2	hsa-mir-16	dbDEMC, mir2disease	27	hsa-mir-451a	dbDEMC
3	hsa-mir-342	dbDEMC	28	hsa-mir-320a	dbDEMC
4	hsa-mir-373	dbDEMC	29	hsa-mir-483	dbDEMC
5	hsa-mir-130b	dbDEMC	30	hsa-mir-204	dbDEMC, mir2disease
6	hsa-mir-23b	dbDEMC	31	hsa-mir-302b	dbDEMC
7	hsa-mir-106b	dbDEMC	32	hsa-mir-193b	dbDEMC
8	hsa-mir-122	dbDEMC	33	hsa-mir-129	dbDEMC
9	hsa-mir-15a	dbDEMC	34	hsa-mir-10a	dbDEMC
10	hsa-mir-99a	dbDEMC, mir2disease	35	hsa-mir-297	dbDEMC
11	hsa-mir-20b	dbDEMC	36	hsa-mir-491	dbDEMC
12	hsa-mir-328	dbDEMC	37	hsa-mir-299	unconfirmed
13	hsa-mir-372	dbDEMC	38	hsa-mir-152	dbDEMC
14	hsa-mir-148b	dbDEMC	39	hsa-mir-187	dbDEMC
15	hsa-mir-28	dbDEMC	40	hsa-mir-508	dbDEMC
16	hsa-mir-139	dbDEMC, mir2disease	41	hsa-mir-378a	dbDEMC
17	hsa-mir-141	dbDEMC, mir2disease	42	hsa-mir-409	unconfirmed
18	hsa-mir-130a	dbDEMC, mir2disease	43	hsa-mir-625	dbDEMC
19	hsa-mir-429	dbDEMC, mir2disease	44	hsa-mir-370	dbDEMC
20	hsa-mir-15b	dbDEMC	45	hsa-mir-302c	dbDEMC
21	hsa-mir-194	dbDEMC	46	hsa-mir-296	dbDEMC
22	hsa-mir-211	dbDEMC	47	hsa-mir-362	unconfirmed
23	hsa-mir-424	dbDEMC	48	hsa-mir-190a	dbDEMC
24	hsa-mir-144	dbDEMC	49	hsa-mir-432	dbDEMC
25	hsa-mir-208b	unconfirmed	50	hsa-mir-340	dbDEMC

and life. Hence, we conducted a case study focusing on the identification of miRNAs associated with lung cancer. Based on the information presented in Table 6, it is evident that our model successfully confirms 46 out of the top 50 miRNAs that are associated with the disease, as validated by dbDEMC or mir2disease. This also demonstrates that the model we proposed has good performance.

IV. CONCLUSION AND DISCUSSIONS

Previous research has demonstrated that dysregulated expression of miRNA is linked to numerous intricate human diseases. Thus, predicting potential miRNA-disease associations can help medical experts better study the pathology of diseases and promote the development of clinical medicine. In this study, we proposes a miRNA-disease association (MDA) prediction model (SNMDA), which addresses the difficulty of obtaining appropriate feature representations for miRNAs and diseases that integrate all information from similarity networks and association matrices. The model combines Singular Value Decomposition (SVD) and Node2vec methods to design a computational framework (SNMDA) for predicting unknown miRNA-disease associations. In the linear feature acquisition stage, the model uses SVD technology to capture key information of miRNA and disease nodes, and in the nonlinear feature acquisition stage, the Node2vec method is used to obtain the feature representations of miRNA and disease nodes. Finally, the features obtained from both stages are merged, with linear and nonlinear features complementing each other and forming the final predictive vector. This merged vector can more comprehensively reflect the features of miRNA and disease nodes, providing richer and more accurate information for subsequent analysis and prediction. Through this feature fusion method, researchers can more effectively identify potential associations between miRNAs and diseases. To illustrate the predictive performance of SNMDA, this study compares five computational methods (GAEMDA, SMALF, HGANMDA, GBDT-LR, ABMDA). The comparison results of five-fold cross-validation show that SNMDA has improved performance in predicting miRNA-disease associations. SNMDA achieved an AUC value of 0.9608, and compared with the baseline methods, the performance of the model in this paper has been enhanced. To further evaluate the performance of SNMDA, three case studies on gastric cancer, esophageal cancer, and lung cancer were conducted, demonstrating that SNMDA can effectively infer unknown miRNA-disease interactions.

However, SNMDA also has some limitations, which require further investigation. Due to the lack of negative samples, we chose unknown miRNA-disease associations as negative samples. There may be false negatives in these negative samples, which may also affect the results of the experiment. Therefore, finding a reliable negative sample will help to further enhance the model's performance. Meanwhile, the use of the associated data alone hardly fully reflects the complex interactions between miRNAs and other biomolecules. In future studies, it is necessary to improve the extension of experimental data by combining supplementary biological data, such as target gene information and RNA sequence data. While the node2vec simple computing framework performs well, we can further improve performance by adopting other novel machine learning approaches in the future. Since the node2vec method generates a fixed feature vector for each node, it cannot capture the dynamic changes of the graph. If the nodes of the graph are changed, such as deleting or adding nodes, the model needs to be retrained to update the feature vector of each node. On the other hand, the acquisition of node feature vectors is affected by the parameters p and q of the node2vec model, and appropriate parameters need to be selected after several experiments.

REFERENCES

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

[2] G. Meister and T. Tuschl, "Mechanisms of gene silencing by double-stranded RNA," *Nature*, vol. 431, no. 7006, pp. 343–349, Sep. 2004.

[3] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The C. Elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14," *Cell*, vol. 75, no. 5, pp. 843–854, Dec. 1993.

[4] A. M. Cheng, "Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis," *Nucleic Acids Res.*, vol. 33, no. 4, pp. 1290–1297, Feb. 2005.

[5] X. Karp and V. Ambros, "Encountering microRNAs in cell fate signaling," *Science*, vol. 310, no. 5752, pp. 1288–1289, Nov. 2005.

[6] E. A. Miska, "How microRNAs control cell division, differentiation and death," *Current Opinion Genet. Develop.*, vol. 15, no. 5, pp. 563–568, Oct. 2005.

[7] D. P. Bartel, "MicroRNAs: Target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215–233, Jan. 2009.

[8] N. Lynam-Lennon, S. G. Maher, and J. V. Reynolds, "The roles of microRNA in cancer and apoptosis," *Biol. Rev.*, vol. 84, no. 1, pp. 55–71, Feb. 2009.

[9] D. Sayed and M. Abdellatif, "MicroRNAs in development and disease," *Physiological Rev.*, vol. 91, no. 3, pp. 827–887, Jul. 2011.

[10] G. A. Calin, C. D. Dumitru, M. Shimizu, R. Bichi, S. Zupo, E. Noch, H. Aldler, S. Rattan, M. Keating, K. Rai, L. Rassenti, T. Kipps, M. Negrini, F. Bullrich, and C. M. Croce, "Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13Q14 in chroniclymphocytic Leukemia," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 24, pp. 15524–15529, 2002.

[11] J. Yu, T. Moriyama, K. Ohuchida, L. Cui, N. Sato, M. Nakamura, S. Takahata, E. Nagai, K. Mizumoto, and M. Tanaka, "430 micro RNA (miR-17-5p) is overexpressed in pancreatic cancer, and upregulation of miR-17-5p enhanced cancer cell proliferation and invasion in vitro," *Gastroenterology*, vol. 134, no. 4, p. 62, Apr. 2008.

[12] K.-I. Kozaki, I. Imoto, S. Mogi, K. Omura, and J. Inazawa, "Exploration of tumor-suppressive MicroRNAs silenced by DNA hypermethylation in oral cancer," *Cancer Res.*, vol. 68, no. 7, pp. 2094–2105, Apr. 2008.

[13] Y. Chen, J. A. Gelfond, L. M. Mcmanus, and P. K. Shireman, "Reproducibility of quantitative RT-PCR array in miRNA expression profiling and comparison with microarray analysis," *BMC Genomics*, vol. 10, no. 1, p. 407, 2009.

[14] C. Liang, S. Yu, and J. Luo, "Adaptive multi-view multi-label learning for identifying disease-associated candidate miRNAs," *PLOS Comput. Biol.*, vol. 15, no. 4, Apr. 2019, Art. no. e1006931.

[15] S. Wang, B. Lin, Y. Zhang, S. Qiao, F. Wang, W. Wu, and C. Ren, "SGAEMDA: Predicting miRNA-disease associations based on stacked graph autoencoder," *Cells*, vol. 11, no. 24, p. 3984, Dec. 2022.

[16] 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," *Comput. Biol. Chem.*, vol. 85, Apr. 2020, Art. no. 107200.

[17] D. Liu, Y. Huang, W. Nie, J. Zhang, and L. Deng, "SMALF: MiRNA-disease associations prediction based on stacked autoencoder and XGBoost," *BMC Bioinf.*, vol. 22, no. 1, pp. 1–18, Dec. 2021.

[18] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[19] Y. Zhao, X. Chen, and J. Yin, "Adaptive boosting-based computational model for predicting potential miRNA-disease associations," *Bioinformatics*, vol. 36, no. 1, p. 330, Jan. 2020.
[20] 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 Comput. Biol.*, vol. 13, no. 3, Mar. 2017, Art. no. e1005455.
[21] Z.-W. Li, Q.-K. Wang, C.-A. Yuan, P.-Y. Han, Z.-H. You, and L. Wang, "Predicting MiRNA-disease associations by graph representation learning based on jumping knowledge networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 1, no. 1, pp. 1–10, Aug. 2022.
[22] Z. Li, J. Li, R. Nie, Z.-H. You, and W. Bao, "A graph auto-encoder model for miRNA-disease associations prediction," *Briefings Bioinf.*, vol. 22, no. 4, pp. 1–20, Jul. 2021.
[23] Z. Li, T. Zhong, D. Huang, Z.-H. You, and R. Nie, "Hierarchical graph attention network for miRNA-disease association prediction," *Mol. Therapy*, vol. 30, no. 4, pp. 1775–1786, Apr. 2022.
[24] Y. Zhong, P. Xuan, X. Wang, T. Zhang, J. Li, Y. Liu, and W. Zhang, "A non-negative matrix factorization based method for predicting disease-associated miRNAs in miRNA-disease bilayer network," *Bioinformatics*, vol. 34, no. 2, pp. 267–277, Jan. 2018.
[25] Z. Cui, J.-X. Liu, Y.-L. Gao, C.-H. Zheng, and J. Wang, "RCMF: A robust collaborative matrix factorization method to predict miRNA-disease associations," *BMC Bioinf.*, vol. 20, no. S25, pp. 1–10, Dec. 2019.
[26] Y. Ding, X. Lei, B. Liao, and F.-X. Wu, "Predicting miRNA-disease associations based on multi-view variational graph auto-encoder with matrix factorization," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 1, pp. 446–457, Jan. 2022.
[27] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang, and Y. Liu, "MiR2Disease: A manually curated database for microRNA deregulation in human disease," *Nucleic Acids Res.*, vol. 37, no. 1, pp. D98–D104, Jan. 2009.
[28] Z. Yang, F. Ren, C. Liu, S. He, G. Sun, Q. Gao, L. Yao, Y. Zhang, R. Miao, Y. Cao, Y. Zhao, Y. Zhong, and H. Zhao, "dbDEMC: A database of differentially expressed miRNAs in human cancers," in *BMC Genomics*, vol. 11. Springer, 2010, pp. 1–8.
[29] C. Cui, B. Zhong, R. Fan, and Q. Cui, "HMDD v4.0: A database for experimentally supported human microRNA-disease associations," *Nucleic Acids Res.*, vol. 52, no. D1, pp. D1327–D1332, Jan. 2024.
[30] H. Khan, M. Ullah, F. Al-Machot, F. Alaya Cheikh, and M. Sajjad, "Deep learning based speech emotion recognition for Parkinson patient," *Electron. Imag.*, vol. 35, no. 9, pp. 1–298, Jan. 2023.
[31] H. Khan, T. Hussain, S. Ullah Khan, Z. Ahmad Khan, and S. W. Baik, "Deep multi-scale pyramidal features network for supervised video summarization," *Expert Syst. Appl.*, vol. 237, Mar. 2024, Art. no. 121288.
[32] P. Xuan, K. Han, M. Guo, Y. Guo, J. Li, J. Ding, Y. Liu, Q. Dai, J. Li, Z. Teng, and Y. Huang, "Prediction of microRNAs associated with human diseases based on weighted K most similar neighbors," *PLoS One*, vol. 8, no. 8, Aug. 2013, Art. no. e70204.
[33] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, Jul. 2010.
[34] C. Pasquier and J. Gardès, "Prediction of miRNA-disease associations with a vector space model," *Sci. Rep.*, vol. 6, no. 1, pp. 1–10, Jun. 2016.
[35] X. Chen, C. C. Yan, X. Zhang, Z.-H. You, L. Deng, Y. Liu, Y. Zhang, and Q. Dai, "WBSMDA: Within and between score for MiRNA-disease association prediction," *Sci. Rep.*, vol. 6, no. 1, pp. 1–9, Feb. 2016.
[36] D. Billsus and M. J. Pazzani, "Learning collaborative information filters," in *Proc. ICML*, vol. 98, 1998, pp. 46–54.
[37] H. Cai, V. W. Zheng, and K. C. Chang, "A comprehensive survey of graph embedding: Problems, techniques, and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1616–1637, Sep. 2018.
[38] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," 2017, *arXiv:1709.05584*.
[39] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 855–864.
[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[41] S. De Winter, T. Decuypere, S. Mitrovic, B. Baesens, and J. De Weerdt, "Combining temporal aspects of dynamic networks with node2vec for a more efficient dynamic link prediction," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2018, pp. 1234–1241.
[42] D. Pavithra, K. Sabitha, and T. Rajkumar, "Identification of small molecule inhibitors for differentially expressed miRNAs in gastric cancer," *Comput. Biol. Chem.*, vol. 77, pp. 442–454, Dec. 2018.
[43] H. Kollarova, L. Machova, D. Horakova, G. Janoutova, and V. Janout, "Epidemiology of esophageal cancer–an overview article," *Biomed. Papers*, vol. 151, no. 1, pp. 17–28, Jun. 2007.
[44] W. D. Travis, L. B. Travis, and S. S. Devesa, "Lung cancer," *Cancer*, vol. 75, no. 1, pp. 191–202, 1995.

YUNXIA LIU received the B.Sc. degree from Inner Mongolia University of Technology, Hohhot, China, in 2021. She is currently pursuing the master's degree with the College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China.

JIAZHEN LIN is currently pursuing the B.Eng. degree in biomedical engineering with Northeastern University, Shenyang, China. His research interests include artificial intelligence, biomedical data science, and machine learning.

PIN LIANG received the B.Sc. degree in biomedical engineering from Northeastern University, Shenyang, China, in 2022, where he is currently pursuing the master's degree with the College of Medicine and Biological Information Engineering.

YAYU TIAN received the B.Sc. degree in intelligent medical engineering from Northeastern University, Shenyang, China, in 2023, where she is currently pursuing the master's degree with the College of Medicine and Biological Information Engineering.

XUAN HE received the B.Sc. degree in computer science from Northeast Normal University, Changchun, China, in 2009, and the M.Sc. and Ph.D. degrees in computer science from Northeastern University, Shenyang, China, in 2011 and 2016, respectively. In 2016, she joined the College of Medicine and Biological Information Engineering, Northeastern University, as an Associate Professor.