NCMD: Node2vec-Based Neural Collaborative Filtering for Predicting MiRNA-Disease Association

Jihwan Ha and Sanghyun Park

Abstract—Numerous studies have reported that micro RNAs (miRNAs) play pivotal roles in disease pathogenesis based on the deregulation of the expressions of target messenger RNAs. Therefore, the identification of disease-related miRNAs is of great significance in understanding human complex diseases, which can also provide insight into the design of novel prognostic markers and disease therapies. Considering the time and cost involved in wet experiments, most recent works have focused on the effective and feasible modeling of computational frameworks to uncover miRNA-disease associations. In this study, we propose a novel framework called node2vec-based neural collaborative filtering for predicting miRNA-disease association (NCMD) based on deep neural networks. Initially, NCMD exploits Node2vec to learn low-dimensional vector representations of miRNAs and diseases. Next, it utilizes a deep learning framework that combines the linear ability of generalized matrix factorization and nonlinear ability of a multilayer perceptron. Experimental results clearly demonstrate the comparable performance of NCMD relative to the state-of-the-art methods according to statistical measures. In addition, case studies on breast cancer, lung cancer and pancreatic cancer validate the effectiveness of NCMD. Extensive experiments demonstrate the benefits of modeling a neural collaborative-filtering-based approach for discovering novel miRNA-disease associations.

Index Terms—MiRNA-disease association, deep neural network, node2vec, Gaussian interaction profile kernel, collaborative filtering

1 Introduction

MICRO RNAs (miRNAs) are a class of small, single-stranded, non-coding RNAs (approximately 22 nucleotides) that suppress the expression of target messenger RNAs at the post-transcriptional level by binding to 3 untranslated regions [1], [2], [3]. Increasing evidence has shown that miRNAs can also function as positive regulators affecting the development of diseases [4], [5]. Since the first discovery of two miRNAs (lin-4 and let-7) in *Caenorhabditis elegans* [6], numerous studies have reported the pivotal roles of miRNAs in multiple biological processes, including aging [7], apoptosis [8], development [9], differentiation [8], proliferation [10], and viral infection [8]. Additionally, increasing evidence has suggested that miRNAs are involved in multiple cancer-related processes [11]. For example, miR-335 and miR-31 have been found to play a crucial role in inhibiting breast cancer [12]–[14]. Additionally, further experiments validated that miR-101 inhibits breast cancer by targeting *Stathmin1* [15] and miR-122 suppresses the tumorigenesis of breast cancer and cell proliferation by targeting ICF1R [16]. Therefore, the identification of miRNA-disease associations can shed new light on disease pathogenesis from a genetic perspective. With the development of high-throughput techniques, numerous miRNAs have been detected. However, due to the laborious task in clinical identification methods require significant human and material resources, an increasing number of computational methods have been developed to predict potential miRNA-disease associations [17]. Existing computational models can largely be categorized into two categories: similarity- and machine-learning-based models.

1.1 Related Works

Similarity-based models have made significant progress in terms of designing miRNA-disease prediction models based on the well-known biological assumption that functionally similar miRNAs tend to associate with phenotypically similar diseases and vice versa [18], [19], [20]. Jiang *et al*. proposed a novel computational framework that integrates a miRNA functional similarity network, disease similarity network, and phenome-miRNAome network [21]. However, this method is highly dependent on neighborhood information, which leaves room for enhancing its performance by exploring global networks. Mørk *et al*. developed a novel framework called miRPD that explicitly takes advantage of verified miRNA-protein links and text-mined results regarding protein-miRNA associations [22]. In their work, a scoring function was calculated by multiplying the scores of miRNA-protein interactions and protein-disease

- *Jihwan Ha is with the Major of Big Data Convergence, Division of Data Information Science, Pukyoung National University, Busan 48513, South Korea. E-mail: jhha@pknu.ac.kr.*
- *Sanghyun Park is with the Department of Computer Science, Yonsei University, Seoul 03722, South Korea. E-mail: sanghyun@yonsei.ac.kr.*

Manuscript received 12 December 2020; revised 16 May 2022; accepted 8 July 2022. Date of publication 18 July 2022; date of current version 3 April 2023. This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP), funded by the Korean Government, MSIT, under Grant IITP-2017-0-00477, (SW starlab) Research and development of the high performance in-memory distributed DBMS based on flash memory storage in an IoT environment.
(Corresponding author: Sanghyun Park.)
Digital Object Identifier no. 10.1109/TCBB.2022.3191972

associations to indirectly infer miRNA-disease associations. In other words, miRNA-disease associations without common shared protein links are not applicable, which limits the improvement of this method. Chen *et al.* proposed a prediction model called RWRMDA, which stands for "random walk with restart for miRNA-disease association" [23]. This method implements a random walk of the miRNA-miRNA functional similarity network. However, despite its impressive accuracy, this method cannot be applied to miRNAs with no verified disease associations. Shi *et al.* utilized the protein-protein interaction (PPI) network, miRNA-target interactions, and disease-gene associations to implement a modified random walk algorithm [24]. Xuan *et al.* developed a model called HDMP by exploring the k-most-similar neighbors of each node in a network [25]. However, due to its strong dependency on neighborhood information, this model is not suitable for miRNAs with no verified disease associations. Chen *et al.* proposed a general framework called HGIMDA, which stands for "heterogeneous graph inference for miRNA-disease association prediction" [26]. Various studies have reported that environmental factors (EFs) play critical roles in miRNAs. Ha *et al.* proposed a network framework based on the hypothesis that phenotypically similar miRNAs tend to share a large number of EFs [27]. By applying a propagation algorithm to a similarity-based network, their model efficiently predicts potential miRNA-disease associations. However, this model leaves significant room for improvement by leveraging the chemical structures of EFs. In summary, similarity-based models are biased toward known miRNA-disease associations, which limits their improvement.

Many machine-learning-based models have successfully enhanced prediction accuracy by adjusting model parameters and making use of model expandability for diverse biological data. Chen *et al.* proposed the model of a restricted Boltzmann machine for multiple types of miRNA-disease association prediction (RBMMMDA) [28]. The RBMMMDA achieved success not only in terms of enhancing performance, but also in terms of discriminating corresponding types of miRNA-disease associations. Chen *et al.* also developed the "hybrid approach to miRNA-disease association" model to reveal miRNA-disease associations by combining disease semantic similarity, miRNA functional similarity, and GIP kernel similarity [29]. However, this model only takes advantage of neighboring nodes in the same layer, rather than using the topological characteristics of subgraphs in heterogeneous networks. Chen *et al.* proposed matrix decomposition and heterogeneous graph inference (MDHGI) to reveal potential miRNA-disease associations based on a matrix decomposition algorithm [30].

In this era of explosive information growth, recommender systems have achieved immense success not only in terms of aiding users in their decision making, but also in terms of increasing profits for companies that sell products to users. For modeling the crucial factors of discovering miRNA-disease associations, the application of additional biological data is necessary to enhance performance [31], [32]. Therefore, numerous prediction models have adopted a machine learning technique that is widely used in recommender systems. Li *et al.* proposed a novel model called matrix completion for miRNA-disease association (MCMDA) that infers disease-related miRNAs based on a matrix completion algorithm [33]. MCMDA implements a binary adjacency matrix based on known miRNA-disease associations and a singular value threshold algorithm is used to extract novel disease-related miRNAs. ifying an optimal combination of parameters is a critical iss Xio *et al.* proposed a novel framework called graph-regularized nonnegative MF (GRNMF) [34]. GRNMF utilizes a weighted gene network and the semantic associations among diseases to calculate interaction profiles for miRNAs and diseases. Ha *et al.* proposed an MF-based model called prediction of microRNA-disease associations utilizing a matrix completion approach (PMAMCA) to predict novel miRNA-disease associations [35]. PMAMCA applies miRNA expression values as implicit feedback for its objective function to enhance prediction accuracy. Ha *et al.* developed an MF-based framework called inferring miRNA-disease interactions using probabilistic MF (IMIPMF) to uncover disease-related miRNAs [36]. In the domain of disease-related miRNA prediction, there are infrequent miRNAs with few known associations with diseases. Such undesirable anomalies are handled by incorporating a constrained model. IMIPMF is a probabilistic factor-based model that performs well on imbalanced datasets while enhancing prediction accuracy. However, this model still has room for further improvement by incorporating additional biological datasets. Chen *et al.* developed a model called inductive matrix completion for miRNA-disease association prediction (IMCMDA) by applying inductive matrix completion with a heterogeneous graph [37]. The main advantage of IMCMDA is that it not only predicts known miRNA-disease associations, but also measures comprehensive similarities among miRNAs and diseases. Ha *et al.* proposed a miRNA-disease prediction model called improved prediction of miRNA-disease associations based on matrix completion with network regularization [38]. Their model uses an MF technique with miRNA similarity network data as regularization terms to represent miRNA latent vectors more precisely and achieves significant performance improvements by considering direct neighbors. Chen *et al.* presented a computational model of ensemble of decision tree-based miRNA-disease association (EDTMDA), which exploited principal component analysis (PCA) for dimensionality reduction and applied ensemble learning to infer miRNA-disease association [39]. Chen *et al.* also developed a computational model of LRSSLMDA, which integrated two feature profiles based on feature extraction. This model utilized sparse subspace learning with Laplacian regularization and L1-norm to improve the prediction accuracy [40]. Chen *et al.* further presented a prediction model of bipartite network projection for miRNA-disease association prediction (BNPMDA). Based on bias ratings, BNPMDA exploited bipartite network recommendation algorithm to predict potential disease-related miRNAs [41]. Chen *et al.* developed another computational model of NCMCMDA that utilize matrix completion algorithm with neighborhood constraint. This model utilized miRNA functional similarity, disease semantic similarity and Gaussian interaction profile kernel similarity to integrate comprehensive similarity information to infer potential miRNA-disease associations [42]. Chen *et al.* also proposed a computational model of deep-belief network for miRNA-disease association (DBNMDA). DBNMDA utilized pre-train restricted Boltzmann machines

(RBM) to extract the features of all miRNA-disease pairs to fine-turn deep-belief network to calculate predicted scores [43]. Chen *et al.* further presented a novel framework for detecting miRNA-disease associations called SAEMDA. In this work, the authors a Stacked Autoencoder (SAE) to pre-train the model in an unsupervised manner. Then, the same number of each positive and negative samples were used to train SAE efficiently, which yielded superior performance [44]. Liu *et al.* [45] presented a stacked auto-encoder framework to predict miNRA-disease associations. This model utilized miRNA functional similarity, disease semantic similarity, miRNA latent feature, and disease latent feature to obtain miRNA-disease feature vector. Then, they applied XGBoost to predict novel miRNA-disease associations.

In recent years, matrix factorization (MF), which has yielded incredible success in recommender systems, has received significant attention for predicting miRNA-disease associations [46]. MF-based prediction models use the inner products of latent feature vectors to predict miRNA-disease associations. In this regard, MF can be thought of linear model of latent features that violates the triangle inequality. The triangle inequality states that given three entities, the sum of the distances between any two pairwise entities should be greater than or equal to the remaining pairwise distance. When the triangle inequality is violated, prediction models fail to capture the precise features of miRNAs (or diseases), leading to suboptimal prediction performance.

1.2 State-of-the-Art Methods

To evaluate the performance of NCMD, we compared it to the following state-of-the-art methods: MDHGI [30], PMAMCA [35], HGIMDA [26], and MCMDA [33]. Chen *et al.* proposed matrix decomposition and heterogeneous graph inference (MDHGI) to reveal potential miRNA-disease associations based on a matrix decomposition algorithm [30]. Based on matrix decomposition, they could avoid noise in an original adjacency matrix to enhance prediction accuracy. Ha *et al.* proposed an MF-based model called prediction of microRNA-disease associations utilizing a matrix completion approach (PMAMCA) to predict novel miRNA-disease associations [35]. PMAMCA applies miRNA expression values as weights for its objective function to enhance prediction accuracy by considering the biological mechanisms of miRNAs. PMAMCA exhibits competitive performance, even though this model only considers known miRNA-disease associations and miRNA expression data. Chen *et al.* proposed a general framework called HGIMDA, which stands for "heterogeneous graph inference for miRNA-disease association prediction" [26]. HGIMDA integrates miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile (GIP) kernel similarity. It analyzes three-length paths in a constructed hetero graph to infer novel disease-related miRNAs. Li *et al.* proposed a novel model called matrix completion for miRNA-disease association (MCMDA) that infers disease-related miRNAs based on a matrix completion algorithm [33]. MCMDA implements a binary adjacency matrix based on known miRNA-disease associations and a singular value threshold algorithm is used to extract novel disease-related miRNAs. However, identifying an optimal combination of parameters is a critical issue.

1.3 Method Overview

In this work, we strive to address the aforementioned problems by designing a neural architecture called NCMD, which stands node2vec-based neural collaborative filtering for predicting miRNA-disease association. Because neural networks have shown excellent capabilities for approximating continuous functions and deep neural networks (DNNs) have achieved excellent performance in various areas ranging from speech recognition and computer vision to various biological domains, it is inevitable to apply DNNs to the task of disease-related miRNA detection. The crucial factors for predicting miRNA associations can be defined as 1) accurately defining miRNA and disease similarities based on Gaussian profile interaction kernel and modeling, and 2) developing a DNN-based architecture that combines the linear functionality of MF with the nonlinear capabilities of a multilayer perceptron (MLP). NCMD achieves outstanding performance compared to previous methods with areas under the receiver operating characteristic (ROC) curve (AUCs) of 0.924 and 0.845 in the frameworks of global and local leave-one-out cross validation (LOOCV), respectively. Additionally, we conducted various experiments to qualitatively ascertain the benefits of combining the linear function abilities of GMF with the nonlinear abilities of MLP. Experimental results clearly demonstrate the comparable performance of NCMD relative to the state-of-the-art methods according to statistical measures.

2 Materials and Methods

2.1 Method Overview

In this section, we first introduce the datasets that we used in this study and then formalize our novel framework for the prioritization of disease-related miRNAs, namely NCMD. First, we construct a miRNA functional similarity network and disease semantic network based on a public dataset called "misim." We also construct a Gaussian profile interaction kernel. Second, we learn low-dimensional network representations of miRNAs and diseases by employing the node2vec method while preserving network structures and properties. Next, we construct a deep learning framework that fuses the linear function abilities of GMF with the nonlinear abilities of MLP [47]. Finally, we prioritize potential candidates based on the scores assigned by NCMD. The overall workflow for NCMD is illustrated in Fig. 1.

2.2 Human Microrna-Disease Association Data

The data on known human miRNA-disease associations that were used in this study were obtained from the HMDD V3.2 public database. HMDD V3.2 contains 35547 experimentally validated human miRNA-disease associations with 1206 miRNAs and 893 diseases [48]. We can construct a miRNA-disease interaction matrix $Y \in R^{U \times I}$ from the HMDD dataset (U = 1206, I = 893), where U and I represent the numbers of miRNAs and diseases, respectively. We assigned the entry of the miRNA-disease matrix y_{ui} as one if there exists known miRNA-disease association. The miRNA-disease interaction matrix can be defined as:

$$y_{ui} = \begin{pmatrix} 1, miRNA\ u\ and\ disease\ i\ is\ related \\ 0,\ otherwise \end{pmatrix} \quad (1)$$

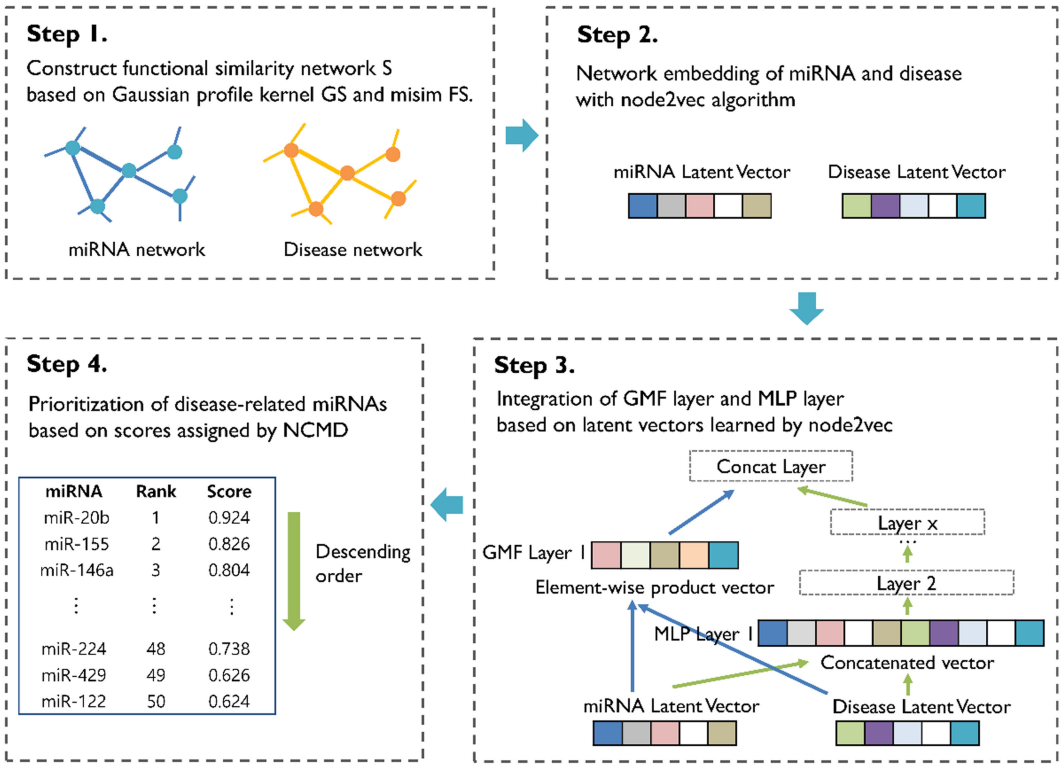

Fig. 1. Overall workflow for NCMD.

2.3 MicroRNA Expression Data

As vast amounts of biological data have become available with the explosive growth of high-throughput techniques, injecting various biological data into models can improve performance as well as decipher the meaningful biological functions, such as disease etiology and pathogenesis. Thanks to high-throughput techniques in biologyical sciences, various omics molecular data at different levels have become readily available, including genomes, transcriptomes, proteomes, and metabolomes. Therefore, integrating data-driven approaches has become an inevitable option. We conclude that various complementary and distinct data should be considered because each type of data may contain unique information regarding the target organism. To this end, utilizing additional biological data could be a promising alternative for enhancing prediction accuracy while reflecting biological mechanisms in a model. Aiming at training NCMD more accurately and precisely, we utilize miRNA expression data to compensate for lacking miRNA-disease associations. We downloaded miRNA expression data from the Cancer Genome Atlas, which provides comprehensive multimodal genomics and proteomics data [49]. In the case where y_{ui} is zero, it does not necessarily mean that miRNA u is unrelated to disease i. There could be a potential association that has not been verified yet. In the light of this reason, we assign miRNA expression values for cases where there are no verified associations between miRNAs and diseases. In manner, we aim to train our deep learning model more precisely by incorporating biological meanings.

2.4 MicroRNA and Disease Network

Networks reflect natural paradigms by representing diverse relationships or interactions among entities in real-world relational datasets, such as social networks, PPI networks, and citation networks. Therefore, we injected miRNA and disease network information into a model to expand the understanding of miRNA/disease functions and their corresponding regulatory activities. By utilizing the nature of miRNAs and diseases in the constructed similarity network, we can capture precise network embeddings of miRNAs and diseases. In the miRNA functional similarity network, nodes represent miRNAs and edges represent the functional similarity values among miRNAs. In the disease functional similarity network, nodes indicate diseases and edges stand for the functional similarity values among diseases. Therefore, it has become a vital issue to construct accurate networks that reflect biological mechanisms accurately.

2.4.1 MicroRNA Functional Similarity

Many important tasks in network analysis involve predictions regarding nodes and edges. To account for network structures and information propagation, constructing an accurate network based on precise edge information is crucial. In general, the physical meanings of edges in the miRNA network can be interpreted as functional similarity scores among miRNAs. We downloaded a miRNA functional similarity dataset called "misim" from http://www.lirmed.com/misim/ and assigned the corresponding edge information to construct a U $\times$ U miRNA functional similarity matrix FS [50]. The similarity score between miRNA(i) and miRNA(j) is expressed as FS(i,j).

2.4.2 Disease Semantic Similarity 1

A directed acyclic graph (DAG) is a common format for representing the relationships among different diseases. Disease

D can be defined as *DAG(D) = (D, T(D), E(D))*. *T(D)*denotes the ancestor nodes of *D* and *E(D)* indicates all of the direct edges pointing from parent nodes to child nodes. A disease mesh descriptor was obtained from the National Library of Medicine (http://www.nlm.nih.gov) [51]. The semantic disease similarity D can be expressed as follows:

$$DV(D) = \sum\nolimits_{t \in T(D)} D_D(t) \tag{2}$$

$$\begin{cases} D_D(d) = 1 \\ D_D(d) = \max\{\Delta_* D_D(d') | d' \in children\ of\ d\}\ if\ d\ \neq\ D \end{cases} \tag{3}$$

Here, Δ denotes the semantic contribution factor. The contribution score for disease *d* decreases as the distance between *D* and *d* increases. It is known that semantically similar diseases tend to share larger portions of DAGs. We define *SS* as the disease semantic similarity matrix. The disease semantic score between *d(i)* and *d(j)* can be expressed as follows:

$$SS_1(d(i), d(j)) = \frac{\sum_{t \in T(i) \cap T(j)} (D_i(t) + D_j(t))}{DV(i) + DV(j)} \tag{4}$$

2.4.3 Disease Semantic Similarity 2

To integrate precise and accurate miRNA similarity, we adopted the methodology of similarity calculation model [25]. In disease semantic simailrity model 1(SS_1), different diseases in the same layer have same contribution to the semantic value. However, the contribution of each disease should be different according to their frequency in the DAGs. In other words, for the diseases that appear less in the DAGs should have higher contribution value of disease t. Therefore, we define new contribution of sematic value of disease *D* as follows:

$$D2_D(t) = -\log\left(\frac{num(DAGs(t))}{num(disease)}\right) \tag{5}$$

here, $num(DAGs(s))$ represents the number of DAGs that contains disease t, and $num(disease)$stands for the number of all diseases. The disease sematic similarity 2 can be calculated in a similar manner to disease sematic similarity 1 as follows:

$$DV2(D) = \sum\nolimits_{t \in T(D)} D2_D(t) \tag{6}$$

$$SS_2(d(i), d(j)) = \frac{\sum_{t \in T(i) \cap T(j)} (D2_i(t) + D2_j(t))}{DV2(i) + DV2(j)} \tag{7}$$

2.4.4 Gaussian Interaction Profile Kernel Similarity for Mirnas and Diseases

Modeling the interactions among entities is a challenging task that should be performed in a sophisticated manner with aid from various computational approaches. The GIP kernel has attracted significant attention for its effective performance in prediction problems in similar areas. Based on the well-known biological assumption that phenotypically similar diseases tend to associate with functionally similar miRNAs and vice versa [18], [19], [20], we computed the GIP similarities of miRNAs and diseases according to known human miRNA-disease associations. IP(m(i)) denotes the interaction profile of miRNA m(i) when observing whether there is a known association between miRNA m(i) and each disease d(i). The GIP similarity between m(i) and m(j) can be computed as

$$GS(m(i), m(j)) = \exp(-r_m ||IP(m(i)) - IP(m(j))||^2) \tag{8}$$

where GS is the GIP kernel, r_m is the hyper parameter for the bandwidth of the kernel, and the value of $r_m^{'}$ is set to one based on empirical experimental results [50]. Disease similarity can be calculated using GS in the same manner.

$$r_m = \frac{r'_m}{\frac{1}{n_m}\sum_{i=1}^{n_m} ||IP(m(i)||^2} \tag{9}$$

2.4.5 Integrated miRNA Similarity

To integrate comprehensive edge information to construct a miRNA functional similarity network, we combined the miRNA functional similarity FS and miRNA Gaussian interaction kernel similarity GS. The integrated edge information in the miRNA similarity network S_m can be expressed as follows:

$$S_m(m(i), m(j)) = \begin{cases} FS(m(i), m(j)),\ if\ m(i)\ and\ m(j)\ has\ functional\ similarity \\ GS_m(m(i), m(j),\ otherwise \end{cases} \tag{10}$$

Similarly, the integrated disease similarity matrix between disease $d(i)$ and disease $d(j)$ can be defined as follows:

$$S_d(d(i), d(j)) = \begin{cases} \frac{SS_1(d(i),d(j)) + SS_2(d(i),d(j))}{2},\ if\ \mathrm{d(i)}\ and\ \mathrm{d(j)}\ has\ semantic\ similarity \\ GS_d(d(i), d(j),\ otherwise \end{cases} \tag{11}$$

2.5 Node2vec Algorithm

Aiming at training deep models more accurately and realistically, we applied an accurate and sophisticated network embedding method called node2vec to capture the vector representations of miRNAs and diseases in a low-dimensional space. Node2vec generates a mapping of nodes to low-dimensional space features while maximizing the likelihood of preserving network properties [53]. In other words, node2vec is a semi-supervised method for representing feature embeddings for the nodes in a network. A common underlying assumption in network embedding is that a node vector should preserve the neighboring network structures. This algorithm explores neighborhoods using both breadth-first sampling and depth-first sampling by tuning two parameters p and q, as illustrated in Fig. 2. NCMD applies the node2vec algorithm to learn rich feature representations for miRNAs and diseases by exploring a high-quality similarity network.

Consider a random walk that is currently at node v that traverses a preceding node t. This walk selects its next step according to the probabilities π_{vx} of the edges (v, x) leaving from v. Random walks decides the next node based on the static edge weights w_{vx} *i.e.*, $\pi_{vx} = w_{vx}$ (In case of unweighted

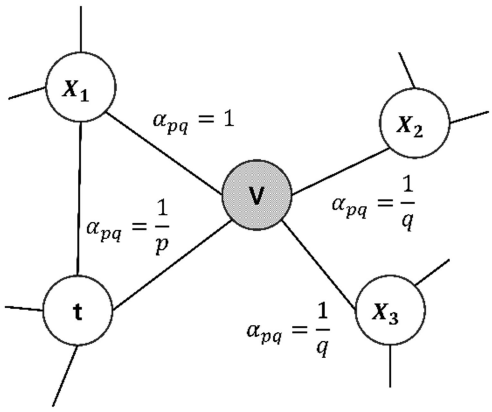

Fig. 2. Illustration of the random walk procedure in node2vec.

graphs $w_{vx}. = 1$). We define the un-normalized transition probability as π_{vx}=$\alpha_{pq}(t,\ x)\cdot w_{vx}$, where the parameter α_{pq} is defined as follows:

$$a_{pq}\begin{cases}\frac{1}{p}\ if\ d_{tx}=0\\ 1\ if\ d_{tx}=1\\ \frac{1}{q}\ if\ d_{tx}=2\end{cases} \tag{12}$$

where d_{tx} represents the shortest path between nodes t and x is an integer ranging from zero to two. The parameter p determines the possibility of revisiting the same node in a random walk. A high value of p indicates a low chance of revisiting nodes that were already visited. This scheme allows the random walk to moderate exploration and avoid two-hop redundancy during sampling. The parameter q allows the walk to differentiate between local and global nodes. If q $>$ 1, then the random walk has a high probability of sampling nodes close to node v. In contrast, the walk searches farther away from node v to obtain global feature information when q is small. In this study, by utilizing the node2vec algorithm, we obtained five-dimensional dense vectors for each miRNA and disease.

2.6 Generalized Matrix Factorization

We propose NCMD, which takes advantage of the linear properties of MF and nonlinear properties of an MLP. Because MF maps miRNAs and diseases to the same latent space, the relationships between objects can be calculated by adopting the inner product as a similarity measure. We demonstrate that MF can be regarded as a specialization of neural collaborative filtering. First, aiming at recovering the linear properties of MF, GMF can be expressed as follows

$$f_1(m_u, d_i) = m_u \odot d_i \tag{13}$$

m_k and d_k represent node2vec-based network embeddings for miRNAs and diseases with dimensions of k. $\odot$ represents the element-wise product of vectors. We then project the result into the output layer.

$$y'_{ui} = a_{out}(h^T(m_u \odot d_i))d_i \tag{14}$$

where a_{out} and h represent the activation function and weight of the output layer, respectively. If we consider the activation function as an identity function and let h be a vector of ones, this is the same as the MF model. In this work, we utilize the sigmoid function $(x) = 1/(1 + e^{-x})$ for a_{out} and update h according log loss values from the data.

2.7 Multi-Layer Perceptron

In multimodal deep learning studies [54], [55], a design for concatenating features from two pathways is frequently adopted. However, simple vector concatenation cannot fully represent the interactions between miRNAs and diseases. Therefore, we adopted multiple hidden layers using a standard MLP. We concatenate network embeddings of miRNAs and diseases and feed them into the MLP. In this manner, we can not only improve the flexibility of the model, but also leverage nonlinearity to learn relationships between m_u and d_i. We attribute the improvement in terms of strong nonlinearities introduced by NCMD to the stacking of multiple nonlinear layers. The MLP under the framework of NCMD is defined as

$$\begin{aligned} z_1 &= f_1(m_u, d_i) = \begin{bmatrix} m_u \\ d_i \end{bmatrix},\\ f_2(z_1) &= a_2(W_2^T z_1 + b_2),\\ &\cdots\cdots\\ f_L(z_{L-1}) &= a_L(W_L^T z_{L-1} + b_L),\\ y'_{dm} &= \sigma(h^T f_L(z_{L-1})), \end{aligned} \tag{15}$$

where, a_x, b_x , and w_x denote the activation function, bias vector, and weight matrix in the x-th layer's perceptron, respectively. For the activation function in each MLP layer, we adopt a rectified linear unit (ReLU), which has been proven to be non-saturated. Empirical results demonstrated that the ReLU performs slightly better than the tanh function in our model, which in turn performs slightly better than the sigmoid function. Therefore, we select ReLU for activation function among the various activation functions including tanh and sigmoid.

2.8 Combination of Generalized Matrix Factorization and Multi-Layer Perceptron

Recall that the main objectives of our proposed model are 1) to compensate for the limitations of MF-based models, which utilize inner products as scoring functions, thereby violating the triangle inequality, and 2) to enhance prediction accuracy by designing a hybrid model that captures both linear and nonlinear latent feature vectors. To this end, we fuse GMF, which uses a linear kernel to capture latent feature interactions, and an MLP, which learns interaction functions from data. We share the same input latent vectors learned by node2vec for GMF and the MLP. The fused model is called NCMD and can be formulated as follows:

$$\begin{aligned} f^{GMF} &= m_u^G \cdot d_i^G,\\ f^{MLP} &= a_L W_L^T(a_{L-1}(\ldots\ a_2(W_2^T\begin{bmatrix} m_u^M \\ d_i^M \end{bmatrix} + b_2)\ \ldots)) + b_L),\\ y'_{ui} &= \left(h^T\begin{pmatrix} f^{GMF} \\ f^{MLP} \end{pmatrix}\right), \end{aligned} \tag{16}$$

where m_u^G and m_u^M represent the network embeddings for GMF and MLP learned by node2vec, respectively. Similarly,

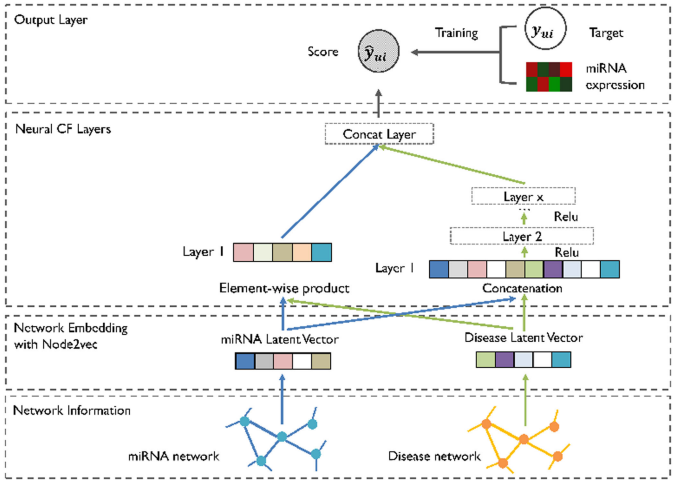

Fig. 3. Overall framework of NCMD.

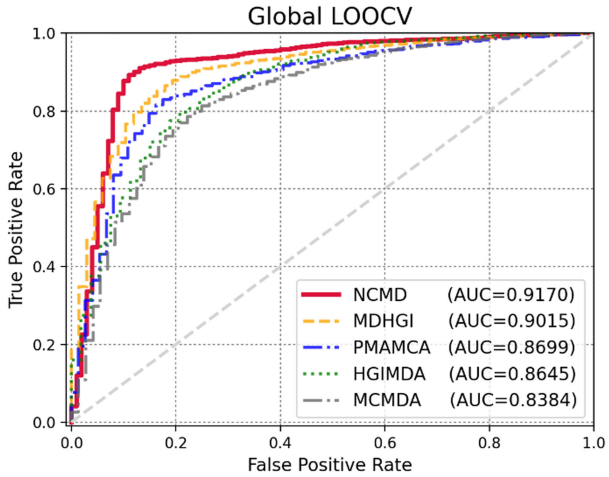

Fig. 4. Performance evaluation in the framework of global LOOCV. NCMD achieves an AUC score of 0.9170, outperforming the previous models.

d_i^G and d_i^M represent the network embeddings learned by node2vec for GMF and MLP, respectively. A ReLU is used for the activation function in the MLP layers and a pairwise loss function is adopted to train the objective function. The NCMD framework is illustrated in Fig. 3.

3 Results

In this section, we enumerate the results of various experiments to demonstrate the superiority of NCMD qualitatively.

3.1 Evaluation Metrics

LOOCV was adopted to evaluate the performance of NCMD. LOOCV can be considered as a special type of n-fold cross validation, where each known disease-related miRNA is considered as a test data and the remaining samples are regarded as training data. This procecure was repeated such that every data sample was used as a test data at least once. Generally, LOOCV can be divided into two types: global LOOCV and local LOOCV. Global LOOCV considers all diseases simultaneously, whereas local LOOCV considers only miRNA candidates for certain diseases. ROC curves were plotted based on the results of LOOCV. The x axis and y axis of the ROC curves represent the true positive rate (sensitivity) and false positive rate (1 − specificity), respectively, where specificity and sensitivity can be calculated as belows:

$$Specificity = \frac{TP}{TN + FP} \tag{17}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{18}$$

Sensitivity denotes the proportion of disease-related miRNA candidates that are ranked higher than the given threshold and specificity stands for the proportion of candidates that are ranked below the threshold. TP and TN refer to the numbers of true positive and negative samples, respectively, and FP and FN refer to the numbers of false positive and negative samples, respectively. AUC values were calculated to validate the superiority of NCMD. In general, an AUC value of one represents perfect prediction and a value of 0.5 is equivalent to the results of random choices. Therefore, as AUC values are close to one, we can conclude that our model can accurately predict miRNA-disease associations.

3.2 Performance Comparisons With Other Methods

To evaluate the performance of NCMD, we compared it to the following state-of-the-art methods: MDHGI [30], PMAMCA [35], HGIMDA [26], and MCMDA [33]. In the framework of global LOOCV, NCMD achieves superior performance compared to the state-of-the-art methods, as shown in Fig. 4. NCMD obtains an AUC value of 0.9170, which is superior to the values of MDHGI (0.9015), PMAMCA (0.8699), HGIMDA (0.8645), and MCMDA (0.8384). Also, based on the various additional evaluation metrics, NCMD showed best performance (Table 1). Next, we performed local LOOCV to evaluate prediction accuracy. As illustrated in Fig. 5, compared to the other methods (i.e., PMAMCA (0.8324), MDHGI (0.8246), HGIMDA (0.7938), and MCMDA (0.7785)) NCMD achieves a superior AUC value of 0.8724. Also, NCMD demonstrated its superior performance based on the various additional evaluation metrics (Table 2). Furthermore, we implemented 5-fold cross validation to clearly demonstrate the performance of NCMD. As shown in Fig. 6, NCMD achieves a superior AUC value of 0.8812 compared to MDHGI (0.8713), PMAMCA (0.8393), MCMDA (0.8362), and HGIMDA (0.8327), which validates its reliable performance. Also, to demonstrate the effectiveness of NCMD, we implemented further experiements based on the various additional evaluation metrics (Table 3). To sum up, experimental results on various evaluation metrics clearly demonstrated the comparable performance of NCMD on predicting miRNA-disease associations.

3.3 Fusion of Generalized Matrix Factorization and Multi-Layer Perceptron

We also performed additional evaluations to determine the benefits of combining GMF and MLP. The key concept of NCMD is to fuse the linear functionality of GMF and nonlinear functionality of MLP to reinforce each method while enhancing overall performance. To this end, we investigated the performance of 1) NCMD, 2) NCMD with GMF alone, and 3) NCMD with MLP alone. As shown in Fig. 7, the experimental results demonstrate that the combination

TABLE 1
Performance Comparison Based on Global LOOCV

Methods	AUC	AUPR	F1	ACC	MCC
NCMD	**0.9170**	**0.8627**	**0.8894**	**0.8612**	**71.26**
MDHGI	0.9015	0.8417	0.8537	0.8481	66.71
PMAMCA	0.8699	0.8328	0.8641	0.8349	68.93
HGIMDA	0.8645	0.8117	0.8406	0.8194	63.41
MCMDA	0.8384	0.7943	0.8219	0.8276	64.58

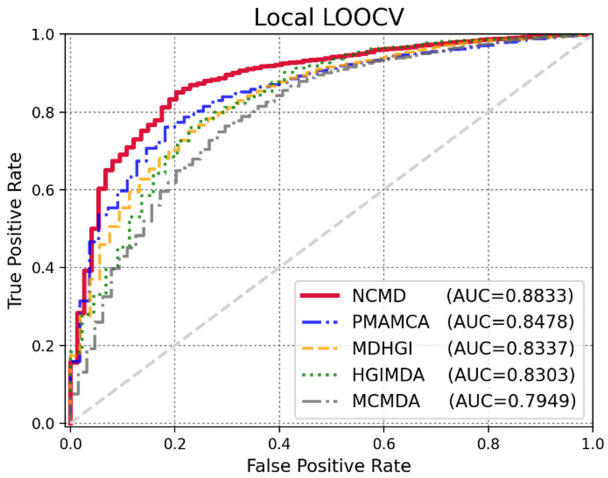

Fig. 5. Performance evaluation in the framework of local LOOCV. NCMD achieves an AUC score of 0.8833, outperforming the previous models.

TABLE 2
Performance Comparison Based on Local LOOCV

Methods	AUC	AUPR	F1	ACC	MCC
NCMD	**0.8833**	**0.8451**	**0.8494**	**0.8309**	**70.63**
PMAMCA	0.8478	0.8302	0.8244	0.8208	68.22
MDHGI	0.8337	0.8042	0.8319	0.8176	64.65
HGIMDA	0.8303	0.8271	0.8168	0.8085	63.82
MCMDA	0.7949	0.7596	0.7885	0.7933	60.14

of GMF and MLP yields significant performance gains. This result clearly demonstrates that the ensemble of linear and nonlinear kernels under the NCMD framework accurately captures known miRNA-disease associations.

3.4 Analysis of Nonlinearity

Neural networks have shown excellent ability in terms of approximating continuous functions. Thus far, DNNs have shown excellent performance in various areas ranging from speech recognition to computer vision and various biological domains. With the goal of developing a more precise prediction model, we endowed our model with nonlinearities by stacking additional layers in the MLP. Generally, stacking more layers is not only beneficial in terms of performance, but also in terms of handling nonlinearities. However, beyond the certain point, performance degrades while the computational cost rises. Therefore, we analyzed experimental results with different numbers of hidden layers in the MLP. MLP-0 denotes no hidden layers and MLP-4 denotes four hidden layers. As

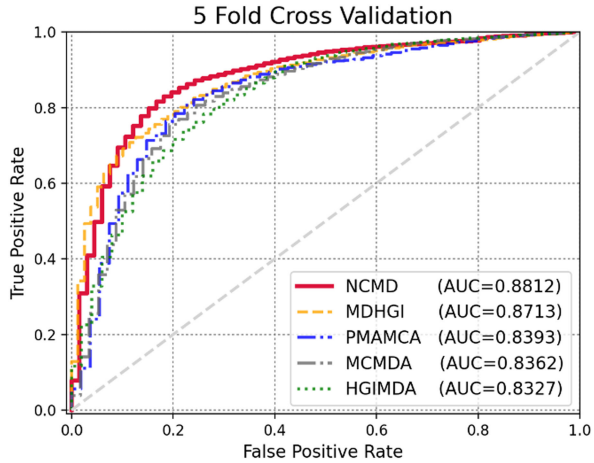

Fig. 6. Performance evaluation in the framework of 5-fold cross validation. NCMD achieves an AUC score of 0.8812, outperforming the previous models.

TABLE 3
Performance Comparison Based on 5-Fold Cross Validation

Methods	AUC	AUPR	F1	ACC	MCC
NCMD	**0.8812**	**0.8427**	**0.8341**	**0.8265**	**69.17**
MDHGI	0.8713	0.8291	0.8165	0.8194	68.09
PMAMCA	0.8393	0.8247	0.8264	0.8106	64.16
MCMDA	0.8362	0.8169	0.8078	0.8173	64.45
HGIMDA	0.8327	0.8044	0.8145	0.8164	63.72

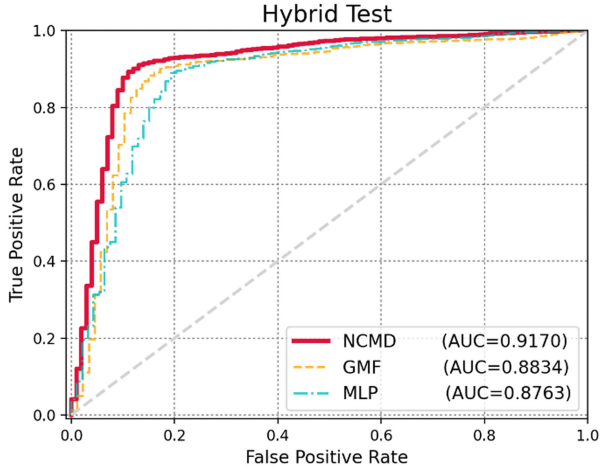

Fig. 7. Performance evaluation of GMF and MLP. NCMD with both GMF and MLP achieves the best performance with a reliable AUC value of 0.9170.

shown in Table 4, our model achieves the best performance with MLP-4. Therefore, we adopted MLP-4 for our other experiments.

3.5 Case Studies (Breast Cancer, Lung Cancer, and Pancreatic Cancer)

To assess the prediction capabilities of NCMD further, we performed case studies on three important human cancers. Validation was implemented based on the two public miRNA-disease association databases: miR2Disease [56] and dbDEMC [57].

TABLE 4
Performance of NCF on Multiple Hidden Layers

	MLP-0	MLP-1	MLP-2	MLP-3	MLP-4	MLP-5
AUC	0.8834	0.8847	0.8902	0.8936	**0.9170**	0.9027

3.5.1 Breast Cancer

Breast cancer is regarded the most common female malignant neoplasms, which takes 22% of all female cancers [58]. Accumultated studies confirmed that various miRNAs associate with breast cancer. For example, microarray-based miRNA profiling on whole blood of early stage breast cancer patients verified that two miRNAs (miR-202, miR-718) were differentially expressed, which could be used as biomarker for early stage breast cancer detection [59]. Therefore, we evaluated the top-50 breast-cancer-related miRNAs based on the scores assigned by NCMD. As shown in Table 5, 49 candidates were proven to be true breast-cancer-related miRNAs according to miR2Disease and dbDEMC. Additionally, we conducted literature-based analysis to determine if the remaining miRNA (miR-23b) has any potential relationship with breast cancer incidence. In [60], it was reported that miR-23b is upregulated in MCF-7 breast cancer cells following genistein treatment. Upregulated expression of miR-23b may be a putative biomarker for use in breast cancer therapy. Additionally, miR-23b upregulation may be important in terms of the human responses to genistein. Also, according to [61], miR-23b is found to be a significant regulatory factor in the progression of multiple cancer cell types by downregulating $CCNG_1$ and the expression of the relevant genes. For these reasons, we were able to validate that all of the top-50 candidates are related to breast cancer.

3.5.2 Lung Cancer

Lung cancer is the most common cause of cancer incidence in worldwide [62]. According to previous experimental studies, miRNAs are highly involved in lung cancer incidence. For example, miR-127 is found to play significant role in lung adeno carcinoma. Also, miR-127 associates with the formation of aggressive phenotypes of lung cancer [63]. Therefore, we performed a case study on the identification of lung-cancer-related miRNAs. As shown in Table 6, 50 out of the top-50 candidates were proven to be true lung-cancer-related miRNAs. In summary, we determined that the top-50 candidates are all lung-cancer-related miRNAs.

3.5.3 Pancreatic Cancer

Pancreatic cancer is one of the leading causes of cancer-associated mortality in worldwide which leads to the highest mortality rates and incidence [64]. According to the previous experimental results, miRNAs can play as novel biomarker for detecting pancreatic ductal adenocarcinoma [65]. Therefore, we performed a case study on the identification of pancreatic-cancer-related miRNAs. As shown in Table 7, 50 out of the top-50 candidates were proven to be true pancreatic-cancer-related miRNAs. In other words, we determined that the top-50 candidates are all pancreatic-cancer-related miRNAs. Extensive experimental results on various human cancers verified that our proposed NCMD model is not only appropriate for the identification of disease-related miRNAs, but is also suitable for discovering potential disease biomarkers.

TABLE 5
Top-50 Breast-Cancer-Related Mirnas Predicted By NCMD

Rank	Name	Evidence
1	hsa-mir-199	dbDEMC
2	hsa-let-7	miR2Disease, dbDEMC
3	hsa-mir-195	miR2Disease, dbDEMC
4	hsa-mir-125	miR2Disease, dbDEMC
5	hsa-mir-18	miR2Disease, dbDEMC
6	hsa-mir-146	miR2Disease, dbDEMC
7	hsa-mir-17	miR2Disease, dbDEMC
8	hsa-mir-29	miR2Disease, dbDEMC
9	hsa-mir-155	miR2Disease, dbDEMC
10	hsa-mir-20	miR2Disease, dbDEMC
11	hsa-mir-133	dbDEMC
12	hsa-mir-222	miR2Disease, dbDEMC
13	hsa-mir-106	dbDEMC
14	hsa-mir-221	miR2Disease, dbDEMC
15	hsa-mir-145	miR2Disease, dbDEMC
16	hsa-mir-200	miR2Disease, dbDEMC
17	hsa-mir-210	miR2Disease, dbDEMC
18	hsa-mir-193	miR2Disease, dbDEMC
19	hsa-mir-122	miR2Disease, dbDEMC
20	hsa-mir-143	miR2Disease, dbDEMC
21	hsa-mir-19	dbDEMC
22	hsa-mir-127	miR2Disease, dbDEMC
23	hsa-mir-26	miR2Disease, dbDEMC
24	hsa-mir-21	miR2Disease, dbDEMC
25	hsa-mir-16	dbDEMC
26	hsa-mir-22	miR2Disease, dbDEMC
27	hsa-mir-1306	dbDEMC
28	hsa-mir-181	miR2Disease, dbDEMC
29	hsa-mir-101	miR2Disease, dbDEMC
30	hsa-mir-10	miR2Disease, dbDEMC
31	hsa-mir-126	miR2Disease, dbDEMC
32	hsa-mir-141	miR2Disease, dbDEMC
33	hsa-mir-103	dbDEMC
34	hsa-mir-15	dbDEMC
35	hsa-mir-196	miR2Disease, dbDEMC
36	hsa-mir-183	dbDEMC
37	hsa-mir-182	miR2Disease, dbDEMC
38	hsa-mir-149	miR2Disease, dbDEMC
39	hsa-mir-135	dbDEMC
40	hsa-mir-223	dbDEMC
41	hsa-mir-153	dbDEMC
42	hsa-mir-134	dbDEMC
43	hsa-mir-107	dbDEMC
44	hsa-mir-204	miR2Disease, dbDEMC
45	hsa-mir-136	miR2Disease, dbDEMC
46	hsa-mir-100	dbDEMC
47	hsa-mir-1244	dbDEMC
48	hsa-mir-23b	Literature [60], [61]
49	hsa-mir-150	dbDEMC
50	hsa-mir-214	dbDEMC

Validation was performed using public databases (miR2disease and dbDEMC) and literature-based analysis. All 50 miRNAs were found to be associated with breast cancer.

3.6 Functional Analysis

We further conducted functional enrichment analysis for the identified miRNA candidates to determine whether the

TABLE 6
Top-50 Lung-Cancer-Related Mirnas Predicted by NCMD

Rank	Name	Evidence
1	hsa-mir-125	miR2Disease, dbDEMC
2	hsa-mir-199	miR2Disease, dbDEMC
3	hsa-mir-155	miR2Disease, dbDEMC
4	hsa-mir-146	miR2Disease, dbDEMC
5	hsa-mir-29	miR2Disease, dbDEMC
6	hsa-let-7	miR2Disease, dbDEMC
7	hsa-mir-18	miR2Disease, dbDEMC
8	hsa-mir-106	miR2Disease, dbDEMC
9	hsa-mir-221	dbDEMC
10	hsa-mir-17	miR2Disease, dbDEMC
11	hsa-mir-195	miR2Disease, dbDEMC
12	hsa-mir-20	miR2Disease, dbDEMC
13	hsa-mir-145	miR2Disease, dbDEMC
14	hsa-mir-21	miR2Disease, dbDEMC
15	hsa-mir-182	miR2Disease, dbDEMC
16	hsa-mir-15	dbDEMC
17	hsa-mir-200	miR2Disease, dbDEMC
18	hsa-mir-19	miR2Disease, dbDEMC
19	hsa-mir-16	miR2Disease, dbDEMC
20	hsa-mir-101	miR2Disease, dbDEMC
21	hsa-mir-126	miR2Disease, dbDEMC
22	hsa-mir-181	miR2Disease, dbDEMC
23	hsa-mir-133	miR2Disease, dbDEMC
24	hsa-mir-198	miR2Disease, dbDEMC
25	hsa-mir-26	miR2Disease, dbDEMC
26	hsa-mir-222	dbDEMC
27	hsa-mir-214	miR2Disease, dbDEMC
28	hsa-mir-191	miR2Disease, dbDEMC
29	hsa-mir-22	miR2Disease, dbDEMC
30	hsa-mir-210	miR2Disease, dbDEMC
31	hsa-mir-141	miR2Disease, dbDEMC
32	hsa-mir-134	dbDEMC
33	hsa-mir-143	miR2Disease, dbDEMC
34	hsa-mir-1	miR2Disease, dbDEMC
35	hsa-mir-27	miR2Disease, dbDEMC
36	hsa-mir-204	miR2Disease, dbDEMC
37	hsa-mir-223	dbDEMC
38	hsa-mir-135	dbDEMC
39	hsa-mir-100	dbDEMC
40	hsa-mir-23	dbDEMC
41	hsa-mir-194	dbDEMC
42	hsa-mir-148	dbDEMC
43	hsa-mir-10	dbDEMC
44	hsa-mir-193	dbDEMC
45	hsa-mir-130	miR2Disease, dbDEMC
46	hsa-mir-107	dbDEMC
47	hsa-mir-206	dbDEMC
48	hsa-mir-183	miR2Disease, dbDEMC
49	hsa-mir-122	dbDEMC
50	hsa-mir-154	dbDEMC

Validation was performed using public databases (miR2disease and dbDEMC). All 50 miRNAs were found to be associated with lung cancer.

TABLE 7
Top-50 Pancreatic-Cancer-Related Mirnas Predicted by NCMD

Rank	Name	Evidence
1	hsa-mir-21	miR2Disease, dbDEMC
2	hsa-let-7	miR2Disease, dbDEMC
3	hsa-mir-146	miR2Disease, dbDEMC
4	hsa-mir-29	miR2Disease, dbDEMC
5	hsa-mir-126	dbDEMC
6	hsa-mir-17	miR2Disease, dbDEMC
7	hsa-mir-155	miR2Disease, dbDEMC
8	hsa-mir-125	miR2Disease, dbDEMC
9	hsa-mir-18	dbDEMC
10	hsa-mir-200	dbDEMC
11	hsa-mir-195	dbDEMC
12	hsa-mir-222	miR2Disease, dbDEMC
13	hsa-mir-143	miR2Disease, dbDEMC
14	hsa-mir-106	miR2Disease, dbDEMC
15	hsa-mir-20	miR2Disease, dbDEMC
16	hsa-mir-193	dbDEMC
17	hsa-mir-19	dbDEMC
18	hsa-mir-22	dbDEMC
19	hsa-mir-221	miR2Disease, dbDEMC
20	hsa-mir-133	dbDEMC
21	hsa-mir-199	miR2Disease, dbDEMC
22	hsa-mir-15	miR2Disease, dbDEMC
23	hsa-mir-145	dbDEMC
24	hsa-mir-182	dbDEMC
25	hsa-mir-141	dbDEMC
26	hsa-mir-26	dbDEMC
27	hsa-mir-127	dbDEMC
28	hsa-mir-210	miR2Disease, dbDEMC
29	hsa-mir-196	dbDEMC
30	hsa-mir-181	miR2Disease, dbDEMC
31	hsa-mir-122	dbDEMC
32	hsa-mir-101	dbDEMC
33	hsa-mir-23	miR2Disease, dbDEMC
34	hsa-mir-27	dbDEMC
35	hsa-mir-204	dbDEMC
36	hsa-mir-10	miR2Disease, dbDEMC
37	hsa-mir-24	miR2Disease, dbDEMC
38	hsa-mir-136	dbDEMC
39	hsa-mir-139	miR2Disease, dbDEMC
40	hsa-mir-134	dbDEMC
41	hsa-mir-194	dbDEMC
42	hsa-mir-192	dbDEMC
43	hsa-mir-1	dbDEMC
44	hsa-mir-214	miR2Disease, dbDEMC
45	hsa-mir-150	dbDEMC
46	hsa-mir-16	miR2Disease, dbDEMC
47	hsa-mir-183	dbDEMC
48	hsa-mir-100	miR2Disease, dbDEMC
49	hsa-mir-223	miR2Disease, dbDEMC
50	hsa-mir-1306	dbDEMC

Validation was performed using public databases (miR2disease and dbDEMC). All 50 miRNAs were found to be associated with pancreatic cancer.

extracted candidates were involved in cancer incidence. Ingenuity pathway analysis (IPA) [66] suggests that our top 50 candidates were related to cancer with a p-value range of 4.86E-02 to 2.92E-20. Furthermore, most of the candidates were related to early stage in invasive cervical squamous cell carcinoma, which affects cancer incidence. We have also found that most of the top related functions were associated with disease disorders that directly and indirectly affect cancer pathogenesis. The results of functional analysis of miRNA candidates indicated the excellent prediction performance of NCMD. The related diseases and disorders are described in Table 8. Detailed analysis results are described in the Supplementary file 1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2022.3191972.

4 Discussion and Conclusion

NCMD is a pioneering framework that leverages network embedding information to represent the precise latent

TABLE 8
Related Diseases and Disorders Derived by IPA

Name	p-value range
Cancer	4.86E−02–2.92E−20
Psychological Disorders	2.55E−03–8.15E−28
Neurological Diseases	3.01E−02–4.71E−27
Organismal Injuries and Abnormalities	4.98E−02–4.71E−27
Reproductive System Diseases	4.86E−02–2.92E−20

features of miRNAs and diseases while maximizing model performance. The excellent performance of NCMD can be attributed to several factors. First, we adopted a neural collaborative filtering method that circumvents the triangle inequality, which is a common problem in the area of MF. We combined the linearity of MF and nonlinearity of MLPs by applying a DNN architecture to compensate for the drawbacks of MF and enhance prediction accuracy. Second, we calculated precise similarity scores among miRNAs and diseases using various computational methods, resulting in reliable miRNA and disease networks that contain meaningful information. Networks are simple data structures in which nodes represent various components and edges represent paired biological concepts, allowing meaningful relationships to be discovered using computational and statistical methods. Based on the network structures, we learned a mapping of nodes to a dense low-dimensional space using the node2vec algorithm to capture the semantics underlying pairwise relationships. Additionally, we incorporated miRNA expression data into our model to generate rich input data while preserving the biological aspects of miRNAs. In summary, our work represents a simple architectural modification to make use of the linearity of MF and nonlinearity of MLPs, which yields excellent performance for discovering potential miRNA-disease associations. However, NCMD still has room for further improvement by incorporating additional biological data, such as target genes and RNA sequence data. Even a simple deep learning architecture based on node2vec performed remarkably well, leaving the opportunity to enhance performance further by employing more sophisticated machine learning methods in the future. We anticipate that NCMD will serve as an essential tool for discovering potential disease biomarkers and predicting miRNA-disease associations.

REFERENCES

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

[2] D. P. Bartel, "MicroRNAs: Genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

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

[4] C. L. Jopling, "Modulation of hepatitis c virus RNA abundance by a liver-specific MicroRNA," *Science*, vol. 309, no. 5740, pp. 1577–1581, Sep. 2005.

[5] S. Vasudevan, Y. Tong, and J. A. Steitz, "Switching from repression to activation: Micrornas can up-regulate translation," *Sci.*, vol. 318, no. 5858, pp. 1931–1934, Dec. 2007.

[6] B. J. Reinhart *et al.*, "The 21-nucleotide let-7 RNA regulates developmental timing in caenorhabditis elegans," *Nature*, vol. 403, no. 6772, pp. 901–906, Feb. 2000.

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

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

[9] X. Karp, "Developmental biology: Enhanced: Encountering microrna in cell fate signaling," *Science*, vol. 310, no. 5752, pp. 1288–1289, Nov. 2005.

[10] 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.

[11] Z. Yu *et al.*, "Aberrant allele frequencies of the SNPs located in microRNA target sites are potentially associated with human cancers," *Nucleic Acids Res.*, vol. 35, no. 13, pp. 4535–4541, Jun. 2007.

[12] K. J. Png *et al.*, "MicroRNA-335 inhibits tumor reinitiation and is silenced through genetic and epigenetic mechanisms in human breast cancer," *Genes Develop.*, vol. 25, no. 3, pp. 226–231, Feb. 2011.

[13] S. F. Tavazoie *et al.*, "Endogenous human microRNAs that suppress breast cancer metastasis," *Nature*, vol. 451, no. 7175, pp. 147–152, Jan. 2008.

[14] S. Valastyan *et al.*, "RETRACTED: A pleiotropically acting MicroRNA, miR-31, inhibits breast cancer metastasis," *Cell*, vol. 137, no. 6, pp. 1032–1046, Jun. 2009.

[15] R. Wang *et al.*, "MiR-101 is involved in human breast carcinogenesis by targeting stathmin1," *PLoS ONE*, vol. 7, no. 10, Oct. 2012, Art. no. e46173.

[16] B. Wang, H. Wang, and Z. Yang, "MiR-122 inhibits cell proliferation and tumorigenesis of breast cancer by targeting IGF1R," *PLoS ONE*, vol. 7, no. 10, Oct. 2012, Art. no. e47053.

[17] X. Chen, D. Xie, Q. Zhao, and Z. H. You, "MicroRNAs and complex diseases: From experimental results to computational models," *Brief. Bioinf.*, vol. 20, no. 2, pp. 515–539, 2019.

[18] W. Zhang, Z. Li, W. Guo, W. Yang, and F. Huang, "A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 2, pp. 405–415, 2019.

[19] W. Lan, J. Wang, M. Li, J. Liu, F.-X. Wu, and Y. Pan, "Predicting microRNA-disease associations based on improved microrna and disease similarities," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 6, pp. 1774–1782, 2016.

[20] J. Luo, P. Ding, C. Liang, B. Cao, and X. Chen, "Collective prediction of disease-associated miRNAs based on transduction learning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 6, pp. 1468–1475, 2017.

[21] Q. Jiang *et al.*, "Prioritization of disease microRNAs through a human phenome-microRNAome network," *BMC Syst. Biol.*, vol. 4, no. 1, May 2010, Art. no. S2.

[22] S. Mork, S. Pletscher-Frankild, A. Palleja Caro, J. Gorodkin, and L. J. Jensen, "Protein-driven inference of miRNA-disease associations," *Bioinformatics*, vol. 30, no. 3, pp. 392–397, Nov. 2013.

[23] X. Chen, M.-X. Liu, and G.-Y. Yan, "RWRMDA: Predicting novel human microRNA–disease associations," *Mol. Biosyst.*, vol. 8, no. 10, pp. 2792–2798, 2012.

[24] H. Shi *et al.*, "Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes," *BMC Syst. Biol.*, vol. 7, no. 1, 2013, Art. no. 101.

[25] P. Xuan *et al.*, "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.

[26] X. Chen, C. C. Yan, X. Zhang, Z.-H. You, Y.-A. Huang, and G.-Y. Yan, "HGIMDA: Heterogeneous graph inference for miRNA-disease association prediction," *Oncotarget*, vol. 7, no. 40, pp. 65 257–65 269, Aug. 2016.

[27] J. Ha, H. Kim, Y. Yoon, and S. Park, "A method of extracting disease-related microRNAs through the propagation algorithm using the environmental factor based global miRNA network," *Bio-Med. Mater. Eng.*, vol. 26, no. s1, pp. S1763–S1772, Aug. 2015.

[28] X. Chen *et al.*, "RBMMMDA: Predicting multiple types of disease-microRNA associations," *Sci. Rep.s*, vol. 5, Sep. 2015, Art. no. e13877.

[29] X. Chen, Y.-W. Niu, G.-H. Wang, and G.-Y. Yan, "HAMDA: Hybrid approach for mirna-disease association prediction," *J. Biomed. Inform.*, vol. 76, pp. 50–58, Dec. 2017.

[30] X. Chen, J. Yin, J. Qu, and L. Huang, "MDHGI: Matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction," *PLOS Comput. Biol.*, vol. 14, no. 8, Aug. 2018, Art. no. e1006418.

[31] M. M. Yin, Z. Cui, M. M. Gao, J. X. Liu, and Y.-L. Gao, "LWPCMF: Logistic weighted profile-based collaborative matrix factorization for predicting mirna-disease associations," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 3, pp. 1122–1129, 2019.

[32] J. Ha and C. Park, "MLMD: Metric learning for predicting mirna-disease associations," *IEEE Access*, vol. 9, pp. 78 847–78 858, 2021.
[33] 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, pp. 21187–21199, Feb. 2017.
[34] Q. Xiao, J. Luo, C. Liang, J. Cai, and P. Ding, "A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations," *Bioinformatics*, vol. 34, no. 2, pp. 239–248, Sep. 2017.
[35] J. Ha, C. Park, and S. Park, "PMAMCA: Prediction of microRNA-disease association utilizing a matrix completion approach," *BMC Syst. Biol.*, vol. 13, no. 1, Mar. 2019, Art. no. 33.
[36] J. Ha, C. Park, C. Park, and S. Park, "IMIPMF: Inferring miRNA-disease interactions using probabilistic matrix factorization," *J. Biomed. Inform.*, vol. 102, Feb. 2020, Art. no. 103358.
[37] 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, Jun. 2018.
[38] J. Ha, C. Park, C. Park, and S. Park, "Improved prediction of miRNA-disease associations based on matrix completion with network regularization," *Cells*, vol. 9, no. 4, Apr. 2020, Art. no. 881.
[39] X. Chen, C. C. Zhu, and J. Yin, "Ensemble of decision tree reveals potential miRNA-disease associations," *PLoS Comput. Biol.*, vol. 15, no. 7, 2019, Art. no. e1007209.
[40] X. Chen and L. Huang, "LRSSLMDA: Laplacian regularized sparse subspace learning for mirna-disease association prediction," *PLOS Comput. Biol.*, vol. 13, no. 12, 2017, Art. no. e1005912.
[41] X. Chen, D. Xie, L. Wang, Q. Zhao, Z. H. You, and H. Liu Chen, "BNPMDA: Bipartite network projection for mirna–disease association prediction," *Bioinformatics*, vol. 34, no. 18, pp. 3178–3186, 2018.
[42] X. Chen, L. G. Sun, and Y. Zhao, "NCMCMDA: MiRNA–disease association prediction through neighborhood constraint matrix completion," *Brief. Bioinf.*, vol. 22, no. 1, pp. 485–496, 2021.
[43] X. Chen, T. H. Li, Y. Zhao, C. C. Wang, and C. C. Zhu, "DeepBelief network for predicting potential miRNA-disease associations," *Brief. Bioinf.*, vol. 22, no. 3, 2021, Art. no. bbaa186.
[44] C. C. Wang, T. H. Li, L. Huang, and X. Chen, "Prediction of potential miRNA–disease associations based on stacked autoencoder," *Brief. Bioinf.*, vol. 23, 2022, Art. no. bbac021.
[45] 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, 2021, Art. no. 219.
[46] S. H. Wang, C. C. Wang, L. Huang, L. Y. Miao, and X. Chen, "Dual-network collaborative matrix factorization for predicting small molecule-miRNA associations," *Brief. Bioinf.*," vol. 23, no. 1, 2022, Art. no. bbab500.
[47] H. Xiangnan, L. Lizi, Z. Hanwang, N. Liqiang, H. Xia, and C. Tat-Seng, "Neural collaborative filtering," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 173–182.
[48] Z. Huang *et al.*, "HMDD v3.0: A database for experimentally supported human microRNA–disease associations," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1013–D1017, Oct. 2018.
[49] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The cancer genome atlas (TCGA): An immeasurable source of knowledge," *Współczesna Onkologia*, vol. 1A, pp. 68–77, 2015.
[50] 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, May 2010.
[51] C. E. Lipscomb, "Medical subject headings (MeSH)," *Bull. Med. Library Assoc.*, vol. 88, pp. 265–266, 2000.
[52] X. Chen and G.-Y. Yan, "Novel human lncRNA–disease association inference based on lncRNA expression profiles," *Bioinformatics*, vol. 29, no. 20, pp. 2617–2624, Sep. 2013.
[53] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 855–864.
[54] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *J. Mach. Learn. Res.*, vol. 15, pp. 2949–2980, 2014.
[55] H. Zhang, Y. Yang, H. Luan, S. Yang, and T.-S. Chua, "Start from scratch: Towards automatically identifying, modeling, and naming visual attributes," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 187–196.
[56] Q. Jiang *et al.*, "miR2Disease: A manually curated database for microRNA deregulation in human disease," *Nucleic Acids Res.*, vol. 37, pp. 98–104, Jan. 2009.
[57] Z. Yang *et al.*, "dbDEMC: A database of differentially expressed miRNAs in human cancers," *BMC Genomic.*, vol. 11, 2010, Art. no. S5.
[58] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics," *CA: A Cancer J. Clinicians*, vol. 69, 2019, Art. no. 7.
[59] M. G. Schrauder *et al.*, "Circulating micro-RNAs as potential blood-based markers for early stage breast cancer detection," *PLoS ONE*, vol. 7, no. 1, 2012, Art. no. e29770.
[60] C. B. Avci *et al.*, "Genistein-induced mir-23b expression inhibits the growth of breast cancer cells," *Współczesna Onkologia*, vol. 1, pp. 32–35, 2015.
[61] J. Yan, J. Jiang, X. N. Meng, Y. L. Xiu, and Z. H. Zong, "MiR-23b targets cyclin G1 and suppresses ovarian cancer tumorigenesis and progression," *J. Exp. Clin. Cancer Res.*, vol. 35, no. 1, pp. 1–10, 2016.
[62] W. D. Travis, L. B. Travis, and S. S. Devesa, "Lung cancer," *Cancer*, vol. 75, no. 1 Suppl, pp. 191–202, 1995.
[63] L. Shi *et al.*, "miR-127 promotes EMT and stem-like traits in lung cancer through a feed-forward regulatory loop," *Oncogene*, vol. 36, no. 12, pp. 1631–1643, 2017.
[64] T. Kamisawa, L. D. Wood, T. Itoi, and K. Takaori, "Pancreatic cancer," *Lancet*, vol. 388, pp. 73–85, 2016.
[65] J. B. Munding *et al.*, "Global microRNA expression profiling of microdissected tissues identifies miR-135b as a novel biomarker for pancreatic ductal adenocarcinoma," *Int. J. Cancer*, vol. 131, no. 2, pp. E86–E95, 2012.
[66] A. Krämer, J. Green, J. Pollard, and S. Tugendreich, "Causal analysis approaches in ingenuity pathway analysis," *Bioinformatics*, vol. 30, no. 4, pp. 523–530, Dec. 2013.

Jihwan Ha received the BS degree in bio-electronics engineering from Pusan National University, in 2013, and the MS and PhD degrees in computer science from Yonsei University, in 2015 and 2020, respectively, under the supervision of Prof. Sanghyun Park. His thesis title is A machine learning approach to predict miRNA-disease associations. His research interests include the areas of machine learning, data mining, bioinformatics and medical/health informatics. He is currently working as assistant professor with the Pukyong National University in major of big data convergence, division of data information science.

Sanghyun Park received the BS and MS degrees in computer engineering from Seoul National University, in 1989 and 1991, respectively, under the supervision of Prof. Sukho Lee, and the PhD degree in computer science from the University of California at Los Angeles (UCLA), in 2001 under the supervision of Prof. Wesley W. Chu. His thesis title is Indexing Techniques for Similarity Searches in Sequence Databases. His research interests includes database, data mining, machine learning, bioinformatics and health informatics, and he is currently supervising the Data Engineering Lab in Yonsei University.