scientific reports

Check for updates

OPEN

Prediction of miRNA-disease association based on multisource inductive matrix completion

YaWei Wang & ZhiXiang Yin✉

MicroRNAs (miRNAs) are endogenous non-coding RNAs approximately 23 nucleotides in length, playing significant roles in various cellular processes. Numerous studies have shown that miRNAs are involved in the regulation of many human diseases. Accurate prediction of miRNA-disease associations is crucial for early diagnosis, treatment, and prognosis assessment of diseases. In this paper, we propose the Autoencoder Inductive Matrix Completion (AEIMC) model to identify potential miRNA-disease associations. The model captures interaction features from multiple similarity networks, including miRNA functional similarity, miRNA sequence similarity, disease semantic similarity, disease ontology similarity, and Gaussian interaction kernel similarity between miRNAs and diseases. Autoencoders are used to extract more complex and abstract data representations, which are then input into the inductive matrix completion model for association prediction. The effectiveness of the model is validated through cross-validation, stratified threshold evaluation, and case studies, while ablation experiments further confirm the necessity of introducing sequence and ontology similarities for the first time.

Keywords miRNA-disease association, Multi-source information, Autoencoder, Inductive matrix completion, Ablation experiment, Optimization algorithm

MicroRNAs (miRNAs) are small, single-stranded, non-coding RNA molecules widely found in eukaryotes, typically containing 21 to 23 nucleotides. According to existing studies, miRNAs can regulate nearly one third of human genes[1], and then participate in the regulation of various physiological and pathological processes such as inflammation, cell proliferation and apoptosis. The relationship between miRNAs and diseases is mainly reflected in their regulation of gene expression. More and more evidence shows that changes in the expression of miRNAs may lead to over-expression or under-expression of disease-related genes, which may be related to the occurrence of cancer, cardiovascular disease, metabolic disease and other diseases[2].Based on the function of miRNAs in cancer, miRNAs can be used as biomarkers to diagnose tumors or provide new therapeutic targets for cancer therapy[3–7]. Therefore, it will be of great significance to accurately predict and identify relevant information between miRNAs and human diseases. However, in the actual scientific research process, some traditional experimental methods usually require a lot of money and time investment, and are inefficient. According to the relevant knowledge of bioinformatics, the prediction of miRNA-disease correlation by computational biology method is not only efficient and low cost, but also very accurate. Therefore, the study of miRNA-disease correlation has important theoretical value and application significance in the field of human disease research and treatment. In this paper, we propose a model called Autoencoder Inductive Matrix Completion (AEIMC) for identifying potential miRNA-disease associations. Finally, to evaluate the performance of the model, we compared it with several existing matrix-based methods. We performed 5-fold cross-validation on experimentally validated miRNA-disease associations and conducted case studies on three complex human diseases. The results show that the AUC values obtained by this model are higher than those achieved by other methods, indicating that this model outperforms existing models and can be applied in practice.

Related research

The existing methods of solving miRNA-disease association prediction mainly include network method , traditional machine learning method , matrix method and deep learning method.

School of Mathematics, Physics and Statistics, Institute for Frontier Medical Technology, Center of Intelligent Computing and Applied Statistics, Shanghai University of Engineering Science, Shanghai 201620, China. ✉email: Zxyin66@163.com

Network-based approach

Complex network-based identification methods have a fundamental assumption: functionally similar miRNAs are more likely to be associated with phenotypically similar diseases. Jiang et al.[8] proposed the first computational model to predict miRNA-disease associations. Later, network-based methods were proposed successively. By integrating Gaussian interaction spectral kernel similarity, Zeng et al.[9] designed a two-layer network to predict miRNA disease associations in the two-layer network using structural perturbation methods. In addition, due to the limited understanding of miRNA's mechanism of action and disease, unknown factors affecting their similarity calculation remain to be revealed.

Traditional machine learning methods

In recent years, researchers have developed a large number of machine learning methods to predict miRNA disease associations. Based on the fact that miRNAs with higher similarity tend to be associated with similar diseases, a variety of machine learning models have been proposed for miRNA-disease association prediction. Biffon et al.[10] integrated neighborhood information and proposed user-based collaborative filtering to determine the association scores of new miRNA-disease pairs. Finally, by comparing various machine learning algorithms, they concluded that the support vector machine SVN had the best performance. Wang et al.[11] also considered introducing miRNA sequence information, using natural language processing technology to extract features and merge them with other similarity information, and then using random forest classifier and logistic tree regression respectively to predict the association between miRNA-disease. Traditional machine learning requires manual feature selection, which makes it difficult for common methods to reveal complex nonlinear relationships in biological data. It is known that miRNA-disease association is limited, and there are too many potential non-associations, resulting in unbalanced samples, and random selection of negative samples may introduce noise.

Matrix-based approach

Matrix-based methods utilize matrix completion or matrix decomposition to predict unknown miRNA-disease associations. Matrix completion-based approaches inspired by recommendation systems, the problem of predicting gene-disease associations can be thought of as similar to designing a recommendation system whose goal is to predict a user's (gene's) "preference" for an item (disease). Chen et al. [12] proposed an inductive matrix completion model considering that the classical matrix completion model cannot be used to predict new diseases without any known associated mirnas and new miRNAs without any known associated diseases. However, Inductive Matrix Completion (IMC) predicts the level of association through the inner product of miRNA and disease features projected into the potential space. This bilinear model only linearly combines the multiplication of potential features and is insufficient to capture the complex and subtle interactions between miRNA and disease. Therefore, Li et al.[13] proposed neural induction matrix completion, and the low-rank feature projection matrix was replaced by a non-linear fully connected layer. Matrix Factorization (MF) has also been applied to miRNA-disease association prediction problems. Chen et al.[14] developed an efficient model for matrix factorization and heterogeneous graph inference. However, matrix factorization is a cold start problem, and MF may be difficult to deal with for novel miRNAs or diseases without any known association.

Some scholars also combine matrix decomposition and matrix completion. Zheng et al.[15] integrated non-negative matrix decomposition, matrix completion algorithm and graph regularization constraints to build a non-negative matrix decomposition model based on matrix completion, which is more robust. Although the matrix approach presents some challenges, for example, it may be difficult to accurately predict associations with other entities due to lack of sufficient known association information. In addition, matrix computations are often computation-expensive, but matrix-based approaches remain a powerful tool for exploring and predicting miRNA-disease associations.

Deep learning methods

Deep learning methods can extract effective feature representations of data, which in turn show great advantages in classification problems. Li et al[16]. proposed the GGAECDA model, which is based on a graph autoencoder (GAE) and combines a graph attention network (GAT) with random walk with restart (RWR). GAT is used to learn low-order neighbor information from the circRNA and disease similarity networks, while RWR is employed to capture high-order neighbor information. The model jointly trains two GAEs to integrate the newly combined feature representations from the circRNA and disease spaces and calculates potential association scores.Guo[17] et al. designed a new graph autoencoder to extract multilevel representations based on variational mechanisms from heterogeneous networks, and then used a gate-based association predictor to predict the miRNA-disease association. Ding[18] deployed two variational graph autoencoders on the miRNA-based network and the disease-based network, and used graph convolution to absorb node features from the graph structure. This model can mitigate the effect of noise generated by randomly selecting negative samples. Deep learning can capture complex patterns, but the models are opaque and difficult to interpret. In addition, the uncertainty of miRNA-disease associations increases the challenge of training robust models.

Considering the difficulties and limitations of the above methods, this paper proposes a new miRNA-disease association prediction model based on multi-source induction matrix completion. The specific ideas are as follows. First, the interaction features of miRNA-disease association were captured based on multi-source similarity networks, including miRNA functional similarity, miRNA sequence similarity, disease semantic similarity, disease ontology similarity and Gauss interaction spectral kernel similarity between disease and miRNA. Secondly, autoencoders, a nonlinear feature learning method, are used to capture more complex and abstract data representations of miRNA and disease. Finally, the learned high-quality features are used as the input of the induction matrix completion model to obtain the miRNA-disease association prediction matrix. The process framework is shown in Fig. 1:

Results

Cross-validation

Cross-validation is a widely recognized method for evaluting and validating the generalization power of miRNA-disease association prediction models. Following the steps of many cutting-edge studies, the strategy of 5-fold cross-validation was adopted in this study[19]. Specifically, we randomly divide the sample set into 5 mutually exclusive subsets, ensuring that each subset is equally used as the test set once during the entire cross-validation process, while the other four subsets are used as the training set. With this approach, the model is trained and tested on different subsets of data, with each iteration producing an independent evaluation of the model's performance.

At each tradeoff, we calculate the area (AUC) value below the receiver Operating characteristic (ROC) curve, a key measure of a categorical model's strength. AUC values range from 0 to 1, with an AUC value of 1 indicating that the model has perfect classification power, while an AUC value of 0.5 indicates that the model's classification power is equivalent to a random guess. In our case, an AUC value of close to or above 0.90 for each fold (as shown in Fig. 4) means that the model exhibits excellent recognition across individual test subsets.

The performance of the final model was determined by calculating the average of the AUC values of the five test subsets, and the resulting composite AUC indicator further validated the robustness and reliability of the model. As shown in Fig. 2, the ROC curve reveals the consistency and accuracy of the model across the various folds, where the overall average AUC value is 0.92, reflecting the strong ability of the model to distinguish between positive and negative samples. This result not only confirms the validity of the model, but also highlights its potential application in predicting the relationship between miRNA and disease.

In particular, we compared AEIMC with several similar prediction methods, including IMCMDA[12], SIMCCDA[20] and NIMCGCN[13], as well as with advanced deep learning models such as PDMDA[21] and LAGCN[22]. All of these methods have demonstrated excellent performance in predicting miRNA-disease association problems, the comprasion results are shown in Table 1. Indicating that our method is more efficient. To further validate the superiority of AEIMC[23], we applied the Bootstrap method to assess the statistical differences between AEIMC and the other five methods. As shown in Fig. 3, the statistical analysis indicates that, at a 0.05 significance level, the differences between AEIMC and all comparison models are significant, since the confidence intervals exclude 0.

Ablation experiment

In the process of constructing heterogeneous similarity network, this study innovatively introduced miRNA sequence simiarity and disease ontology similarity to cover a broader perspective of similarity. The choice of these specific similarity measures is rooted in their distinct biological relevance and their potential to capture different but complementary aspects of miRNA-disease interactions. miRNA sequence similarity leverages the inherent structural features of miRNAs, which often play a key role in their functional behavior and disease associations. Diseases with similar phenotypes or ontological classifications may share underlying biological mechanisms, which makes disease ontology similarity an essential feature for capturing semantic

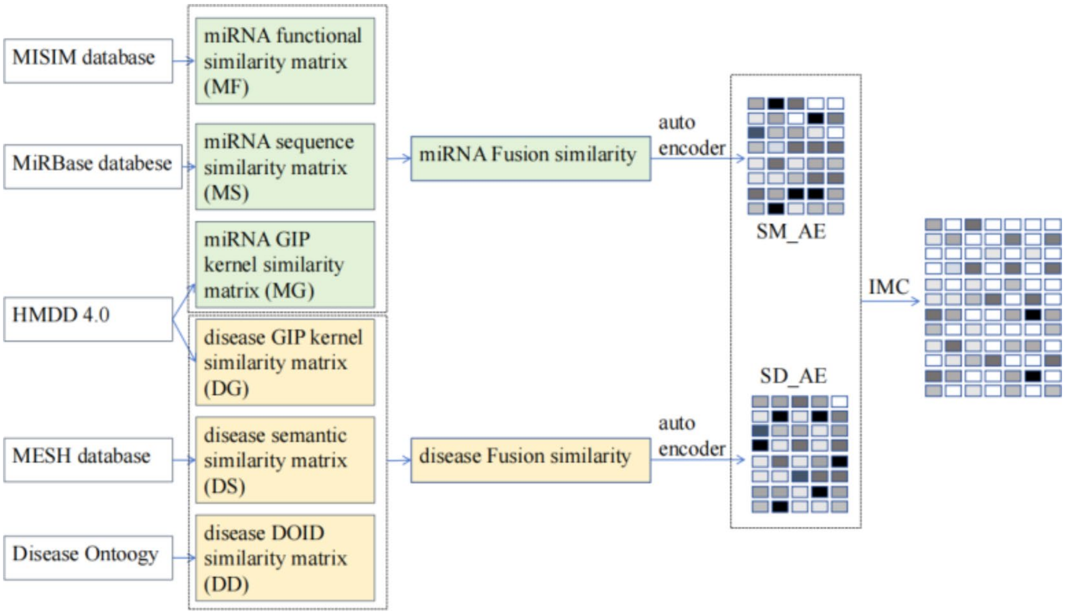

MISIM database
MiRBase datebese
HMDD 4.0
MESH database
Disease Ontoogy
miRNA functional similarity matrix (MF)
miRNA sequence similarity matrix (MS)
miRNA GIP kernel similarity matrix (MG)
disease GIP kernel similarity matrix (DG)
disease semantic similarity matrix (DS)
disease DOID similarity matrix (DD)
miRNA Fusion similarity
disease Fusion similarity
auto encoder
auto encoder
SM_AE
SD_AE
IMC

Fig. 1. Model frame diagram. Firstly, three individual similarity matrices for miRNA and diseases were prepared in the respective databases. Subsequently, a comprehensive similarity matrix for diseases and miRNA was generated. Following this, a low-dimensional representation of the comprehensive similarity matrix was learned through an autoencoder. Finally, the obtained low-dimensional representation was input into an Inductive Matrix Completion (IMC) model to derive the ultimate prediction matrix.

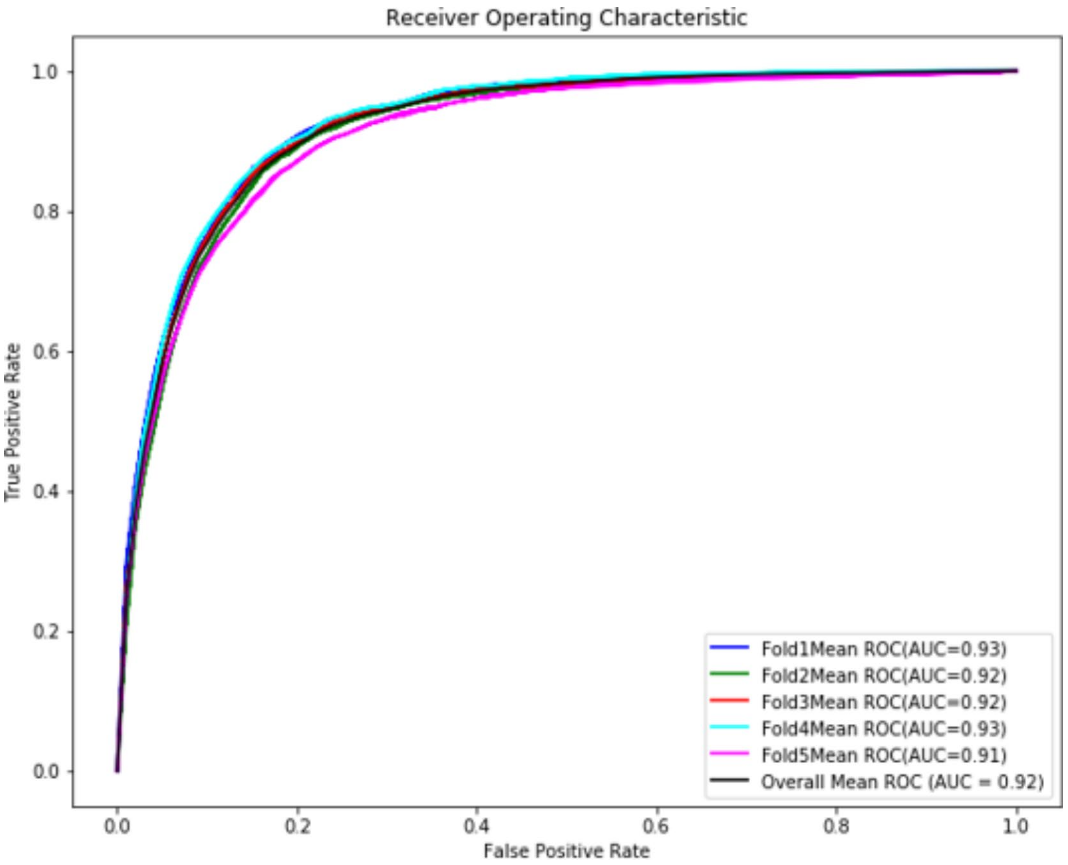


Fig. 2. Roc of AEIMC. The ROC curves and AUC values for each fold based on the AEIMC model, with each fold achieving an AUC value above 0.91, and an average AUC of 0.92.

Method	AEIMC	IMCMDA	SIMCCDA	NIMCGCN	PDMDA	LAGCN
Average AUC	0.92	0.8034	0.8465	0.8490	0.82	0.90

Table 1. Comprasion of the average AUC of several methods.AEIMC achieved the highest AUC value (AUC = 0.92) compared to advanced models such as IMCMDA, SIMCCDA, NIMCGCN, PDMDA, and LAGCN.

relationships between diseases. These two measures were chosen because they represent both the sequence-level characteristics of miRNAs and the semantic-level relationships between diseases, thus broadening the feature space and allowing for more comprehensive predictive modeling. To quantitatively assess the contribution of this novel integration to overall system performance, we performed ablation experiments. This systematic approach involves strategically removing the similarity measures described above and subsequently evaluating the impact on system performance.

Specifically, in the ablation experiment, we eliminated miRNA sequence similarity and disease ontology similarity, performed 5-fold cross-validation following the same procedure as in the original model, and calculated the area under the receiver Operating Characteristic (ROC) curve (AUC) for each test fold. Through this comparison, we found that the average AUC value per fold decreased when no new similarity dimension was introduced (as shown in Fig. 4), a result that clearly demonstrates the importance of the new similarity dimension. After removing these similarity dimensions, the overall recognition ability of the model decreased, which further validated the role of miRNA sequence similarity and disease ontology similarity in improving the accuracy of miRNA-disease association prediction.

Hierarchical threshold evaluation experiment

Given that precision, recall, and F1 scores are sensitive to class imbalance, we implemented measures to more accurately assess the model's performance. We designed a stratified threshold performance evaluation experiment. Specifically, since the prediction results of the model are influenced by the number of known disease associations, we analyzed the model's performance for diseases with varying numbers of known associations and

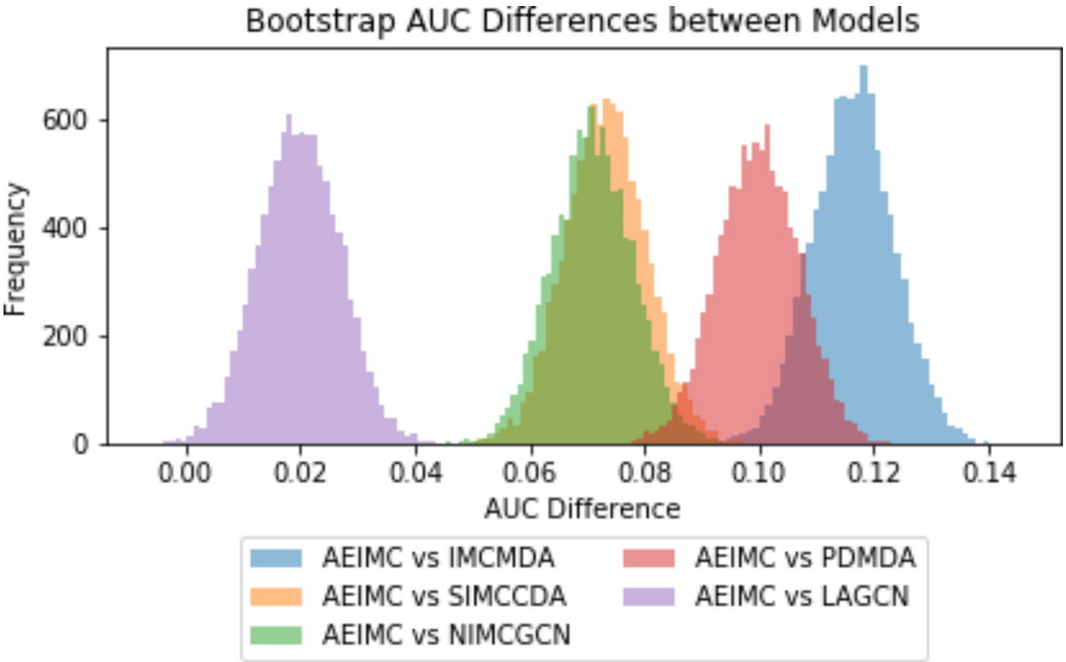


Fig. 3. AUC differences between AEIMC and five comparison models (IMCMDA, SIMCCDA, NIMCGCN, PDMDA, and LAGCN) based on the Bootstrap method.The distributions demonstrate that all AUC differences are significantly positive, as none of the confidence intervals cross zero, indicating that AEIMC performs significantly better than all comparison models at the 0.05 significance level.

further explored the model's precision, recall, and F1 scores under different Top-k thresholds (i.e., the top k% of predicted associations).

For each disease, due to the considerable variation in the number of known associated miRNAs—ranging from fewer than 10 to over 300—we stratified diseases based on the number of known associations and evaluated the model's performance metrics at each level. Simultaneously, we calculated the performance by selecting the top 10%, 30%, 50%, and 80% of predicted miRNAs from the score matrix for each disease. The heatmap (as shown in Fig. 5) clearly visualizes the experimental results, showing that as the number of known associati ons increases, the model's precision, recall, and F1 scores improve significantly. Notably, when the number of miRNAs considered for classification is reduced (i.e., a lower Top-k threshold), the model's performance in all metrics improves. Even for diseases with fewer known associations, the model still demonstrates high performance at the 10% threshold, whereas at higher thresholds, the model's performance depends more on a larger number of known associations.

Case study

Accurate early diagnosis is crucial for disease treatment. To further validate the effectiveness of AEIMC in practical applications, we conducted case studies on breast cancer, lung cancer (non-small cell lung carcinoma), and gastric cancer. These three types of cancer were chosen based on their high incidence rates, mortality rates, and significant impact on specific populations[24]. Breast cancer is the most common cancer among women, lung cancer has the highest mortality rate among men, with non-small cell lung carcinoma being the most deadly subtype, and gastric cancer has a high incidence rate worldwide. Specifically, we used the trained model to calculate the association score matrix between these three diseases and 495 miRNAs, and selected the top 50 miRNAs most strongly associated with each disease. The prediction results were validated using the HMDD v4.0 database. The experimental results shown in Fig. 6, 49 of the top 50 miRNAs associated with breast cancer were validated, 47 for lung cancer (non-small cell lung carcinoma), and 46 for gastric cancer.

Discussion

In this work, we propose a new computational model, called AEIMC, to infer unknown MDA. First, we construct a similar network of miRNA and disease by integrating multi-source similarity features. Then, the autoencoder is used to extract the potential representation of the features and reduce the dimension of the features. Finally, the miRNA-disease association score matrix was obtained by the induction matrix completion model. The effectiveness of the model is demonstrated through cross-validation, stratified threshold performance evaluation experiments, and case studies.In particular, the decrease in the overall mean AUC value after ablation confirmed the effectiveness of the introduced miRNA sequence similarity and disease ontology similarity in capturing complex associations between disease and miRNA, highlighting the need to incorporate multiple biological similarities in the prediction framework.

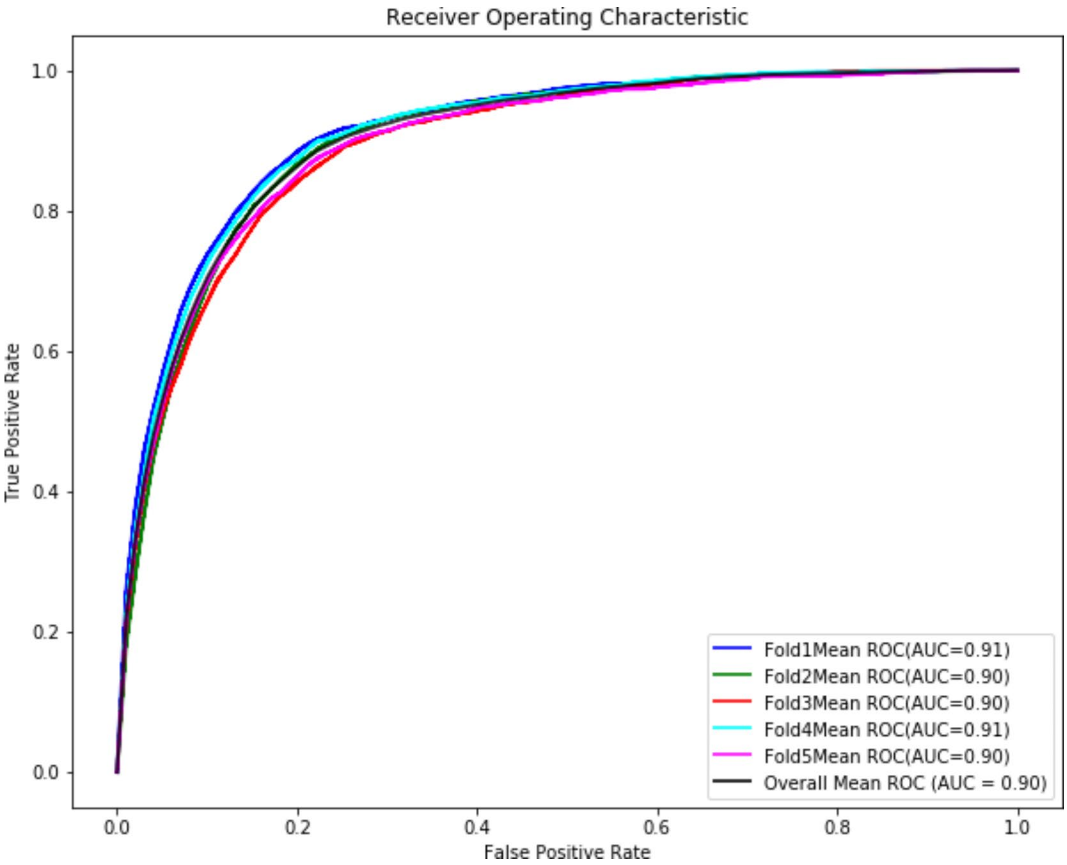


Fig. 4. Roc fusing the two similarity dimensions. Based on the ROC curves and corresponding AUC values for each fold, it is evident that without considering miRNA sequence similarity and disease ontology similarity, there is a noticeable decrease in the accuracy and effectiveness of the model. This validates the rationale for considering additional similarity dimensions.

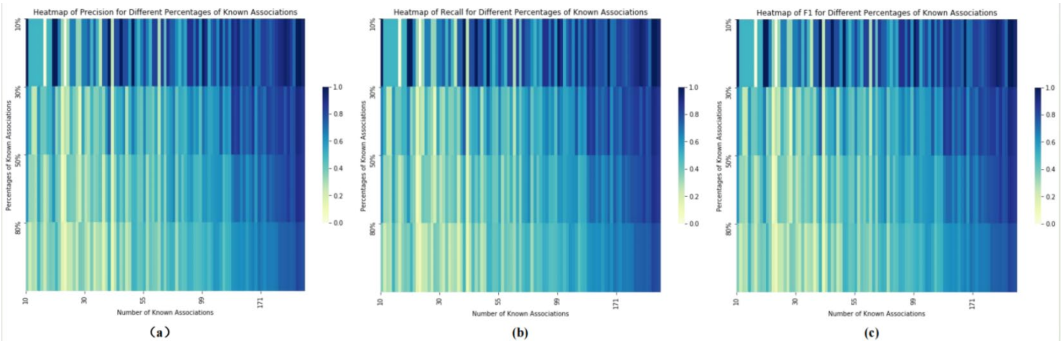


Fig. 5. Heatmaps of Model Performance under different numbers of known associations and Top-k thresholds. (**a**) Precision for different percentages of known associations, (**b**) Recall for different percentages of known associations, and (**c**) F1 for different percentages of known associations. The color variations in the heatmaps show that as the number of known associations increases, the Precision, Recall, and F1 scores gradually improve. Additionally, lowering the Top-k threshold further enhances model performance.

Compared to traditional wet-lab biological experiments, this computational model provides a faster and more cost-effective alternative. Traditional experiments often require expensive materials and time-consuming procedures, whereas computational models can process large datasets in a short amount of time, rapidly inferring potential miRNA-disease associations. Moreover, computational approaches enable the exploration of a vast number of potential biomarkers and therapeutic targets, which is often impractical under conventional experimental conditions. Thus, the AEIMC model not only enhances research efficiency but also expands the

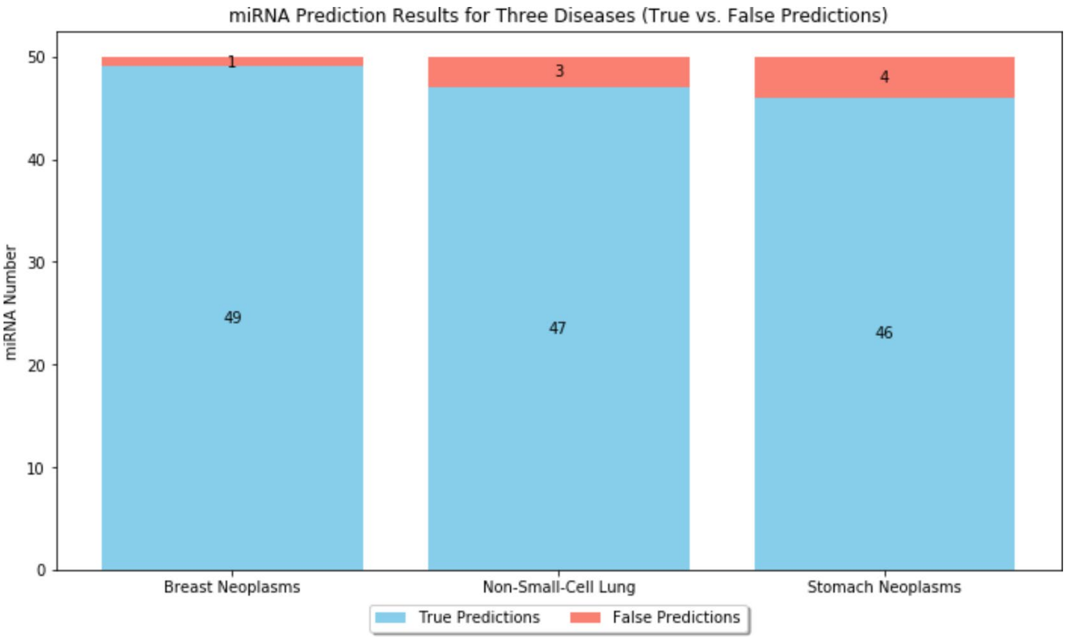


Fig. 6. Validation of prediction results for three types of cancer. This figure shows the validation of the top 50 miRNA prediction results for breast cancer, lung cancer (non-small cell lung carcinoma), and gastric cancer. The blue sections represent correct predictions validated by the HMDD v4.0 database, while the red sections represent incorrect predictions that were not validated.

scope and depth of investigations, facilitating an understanding and prediction of interactions between diseases and miRNAs within complex disease networks.

However, despite the significant advantages of the AEIMC model, there are still some limitations that need to be addressed. One major challenge is the issue of class imbalance, where the number of negative samples far exceeds the number of positive samples. Although the model takes this into account to some extent, the imbalance remains a prominent issue that could bias the results. Therefore, more robust data preprocessing techniques should be considered to mitigate this effect and improve model performance. Additionally, the choice of certain parameters, such as similarity thresholds and model hyperparameters, may also affect the model's results. Careful parameter tuning and validation are required to reduce the risk of overfitting. Future work should focus on further optimizing the balance between positive and negative samples and conducting more comprehensive evaluations to ensure the model's generalizability and robustness.

Methods

This section describes methods for constructing mirNA-disease heterogeneous networks based on different sources of information.

Human miRNA-disease associations

The known human miRNA-disease association data used in this paper were derived from the HMDD V4.0 database[25]. After downloading and collating the database, 12,905 experimentally verified miRNA-disease pairs were finally obtained, including 495 mirnas and 383 diseases.Define a sparse correlation matrix $A \in R^{n\times m}$ to describe the known associations between miRNAs and diseases, where variables n and m represent the number of miRNAs and diseases in the known association dataset, respectively. The known association matrix A can be defined as:

$$\begin{cases} A(i,j) = 1 & if\ miRNA\, i\ has\ association\ with\ disease\ j \\ A(i,j) = 0 & otherwise \end{cases} \tag{1}$$

miRNA functional similarity

The calculation of miRNA similarity is based on the assumption that functionally similar miRNAs tend to be associated with phenotypically similar diseases. Wang et al. [26] proposed a method for calculating functional similarity of mirnas in literature, which is based on the fact that functionally similar miRNAs are often associated with similar diseases, while functionally different miRNAs are often not associated with similar diseases[27,28]. Therefore, we can directly use the miRNA function similarity data, the data can be downloaded from http://www.cuilab.cn/files/images/cuilab/misim.zip.Using these data, a matrix $MF \in R^{n\times n}$ was constructed to represent the functional similarity of miRNA, where n is the number of miRNAs. The element $MF(i,j)$ represents the functional similarity between the i-th and j-th miRNAs.

miRNA sequences similarity

miRNA sequences contain a wealth of information, and if two miRNA sequences are similar, they may regulate a similar set of genes or participate in similar biological processes. MiRBase is a database dedicated to storing miRNA sequences and annotations[29]. The nucleotide sequences of 495 miRNAs were collected from this database. In this paper, the similarity between two miRNA sequences will be calculated using Levenshtein distance[30]. Levenshtein distance is a measure of distance between two string sequences. Specifically, the Levenshtein distance between two sequences is the minimum number of single character edits for one sequence to become another sequence, so the Levenshtein distance is also called the edits distance. The formula for calculating the sequence similarity between two mirnas is as follows:

$$MSim(i,j) = 1 - \frac{dis(i,j)}{len(i) + len(j)} \tag{2}$$

where, $dis(i,j)$ represents the number of edits required to convert miRNA i sequence to miRNA j sequence, and $len()$ represents the sequence length of miRNA.

A matrix $MS \in R^{n \times n}$ is constructed to represent the sequence similarity of miRNA, where n is the number of miRNAs. The element $MS(i,j)$ represents the sequence similarity between the i-th and j-th miRNAs.

Disease semantic similarity

This study from MeSH database (http://www.ncbi.nlm.nih.gov/) has been a hierarchy directed acyclic graph (DAG), to represent the different types of disease[31], for example for the DAG of breast neoplasms (Fig. 7). In the DAG, each node corresponds to a specific disease, while the directed edge represents the direction from a more general disease class to a more specific disease class. This paper calculates the semantic similarity score between different diseases based on the disease DAG.

Suppose i∈D represents a disease. dag(i) represents a set of nodes, including the disease node i and its ancestor nodes in the DAG. Then, the first semantic contribution of disease t∈D to disease i∈D is $S_1(i,t)$, defined as follows [12] :

$$\begin{cases} S_1(i,t) = 1 \qquad if\ t = i \\ S_1(i,t) = \max\{\gamma\ S_1(i,t')|t' \in children\ of\ t\} \quad if\ t \neq i \end{cases} \tag{3}$$

where γ represents a semantic contribution attenuation factor, which means that as the distance between disease t and its ancestor disease increases, its semantic value contribution to a particular disease d will decrease accordingly. According to wang et al.[32], the attenuation factor γ is set at 0.5.

According to the first semantic contribution, the first semantic similarity score $DS_1(i,j)$ between disease i and disease j is defined, expressed as follows:

$$DS_1(i,j) = \frac{\sum_{t \in dag(i) \cap dag(j)} (S_1(i,t) + S_1(j,t))}{\sum_{t \in dag(i)} S_1(i,t) + \sum_{t \in dag(j)} S_1(j,t)} \tag{4}$$

It can be seen that when a large part of the DAG is shared by i and j, it means that i and j have many children in common, which will contribute to i and j in the computation of S_1. The contribution is greatest when $t = i$, as this represents an exact match of the disease to itself; When $t \neq i$, the maximum contribution value of all the children of i needs to be considered. Therefore, if most of the child nodes are shared by i and j, they will contribute larger values to the calculation of S_1, resulting in $DS_1(i,j)$ being higher.

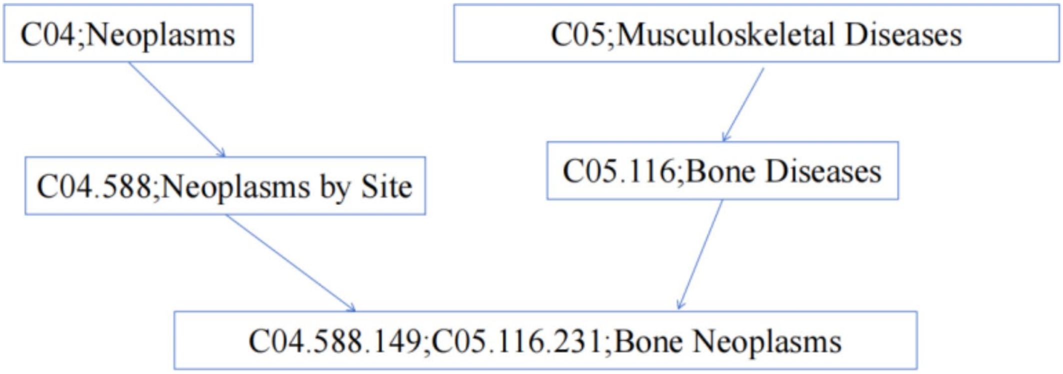

C04;Neoplasms
C05;Musculoskeletal Diseases
C04.588;Neoplasms by Site
C05.116;Bone Diseases
C04.588.149;C05.116.231;Bone Neoplasms

Fig. 7. The DAG of bone neoplasms. The DAG illustrates two distinct paths in Bone Neoplasms, one originating from C04 and the other from C05. The first path involves successive nodes C04.588 and C04.588.149, while the second path follows nodes C05, C05.116, and C05.116.231.

On the other hand, since S_1 only considers the maximum contribution value, this may overlook the importance of different disease contributions. For example, if a particular child node t is only associated with disease i (i.e., only appears in dag(i)), and another child node q appears in both dag(i) and DAG for other diseases, then in the calculation of $DS_1(i,j)$, although the contribution of t may be more unique to i, it will mask the importance of t if the contribution value of q is larger, Because S_1 only takes the maximum value. Such calculations may overlook the unique semantic contributions of certain diseases, especially those that may be unique and important to a particular disease. Therefore, a second semantic contribution is proposed[33]. The second semantic contribution of disease $t \in$ D to disease $i \in$ D is $S_2(i,t)$, which is defined as follows:

$$S_2(i,t) = -\log\frac{(the\ number\ of\ dags\ including\ t)}{the\ number\ of\ disease} \tag{5}$$

According to the second semantic contribution, define the second semantic similarity score $DS_2(i,j)$ between disease i and disease j, expressed as follows:

$$DS_2(i,j) = \frac{\sum_{t\in dag(i)\cap dag(j)}(S_2(i,t)+S_2(j,t))}{\sum_{t\in dag(i)}S_2(i,t)+\sum_{t\in dag(j)}S_2(j,t)} \tag{6}$$

Construct a matrix $DS \in R^{m\times m}$ to represent the semantic similarity of diseases, element $DS(i,j)$ represents the semantic similarity of the i and j diseases, calculated as follows :

$$DS(i,j) = \begin{cases} \dfrac{DS_1(i,j)+DS_2(i,j)}{2} & if\ i\ and\ j\ has\ semantic\ similarity\ score \\ 0 & otherwise \end{cases} \tag{7}$$

Disease ontology similarity

Different from Disease semantic similarity, disease Ontology similarity focuses on analyzing disease similarity from the perspective of Disease Ontology Identifier (DOID). Ontological similarity focuses more on the biological and genetic characteristics of disease. Firstly, Disease DOID information is collected from the database Disease Ontoogy[31]. Then, based on Wang's method[34], the DOSim[35] function was used to calculate the disease semantic similarity. Due to the presence of R-packet DOSE, we can obtain the semantic similarity of disease by inputting disease DOID. Wang's method makes full use of the path information of the two terms on the ontology, i.e. the topological information between the ontology nodes, based on the following formula:

$$DD(i,j) = \frac{\sum_{t\in T_{d_i}\cap T_{d_j}}(S_{d_i}(t)+S_{d_j}(t))}{\sum_{t\in T_{d_i}}S_{d_i}(t)+\sum_{t\in T_{d_j}}S_{d_j}(t)} \tag{8}$$

where, T_{d_i} and T_{d_j} represents the disease and all ancestor nodes of the disease in the disease directed acyclic graph, and $S_{d_i}(t)$ represents the contribution of all nodes in the set $S_{d_i}(t)$ to the disease d_i.The specific formula is as follows:

$$\begin{cases} S_{d_i}(d_i) = 1 \\ S_{d_i}(t) = \max\{\eta * S_{d_i}(t')|t' \in children\ of\ t\} \end{cases} \tag{9}$$

Therefore, a matrix $DD \in R^{m\times m}$ is constructed to represent the ontology similarity of diseases, and elements $DD(i,j)$ represent the ontology similarity of diseases i and j.

Similarity between miRNA and Gaussian interaction spectra of disease.

Gaussian interaction spectral kernel similarity is another algorithm based on known correlation matrix to measure disease similarity and miRNA similarity[36]. In the association matrix, the i-th row $RA(i)$ represents the associations of the i-th miRNA with all diseases. The j-th column $CA(j)$ represents the associations of the j-th disease with all miRNAs. Based on the assumption that similar miRNAs are more likely to be associated with similar diseases, and vice versa, the formula for calculating the spectral similarity of Gauss interactions between miRNAs and diseases is as follows:

$$KD(m_i,m_j) = \exp(-\beta_m||RA(i)-RA(j)||^2) \tag{10}$$

$$KD(d_i,d_j) = \exp(-\beta_d||CA(i)-CA(j)||^2) \tag{11}$$

where,β_m,β_d is the nuclear bandwidth parameter, calculated as follows:

$$\beta_m = \beta_m\prime/(\frac{1}{nd}\sum_{i=1}^{n}RA(i)) \tag{12}$$

$$\beta_d = \beta_d\prime/(\frac{1}{nd}\sum_{i=1}^{n}CA(i)) \tag{13}$$

where,$\beta_m\prime$ and $\beta_d\prime$ is the original bandwidth, and according to previous research, we assign the initial bandwidth to 1[37]. Finally, we can obtain two Gaussian interaction spectral kernel similarity matrices, disease kernel similarity matrix $DG \in R^{m\times m}$, and miRNA kernel similarity matrix $MG \in R^{n\times n}$.

The fusion of disease and miRNA is similar

The combined miRNA-miRNA similarity and disease-disease similarity were obtained by fusing three similarity indicators of miRNA and disease respectively[38]. Since these similarity measures are all based on different criteria, improper fusion methods can bring a lot of noise to the model. Previous studies have shown that for the described nonlinear fusion method[39], the average arithmetic similarity integral method and the average geometric similarity combination strategy, the arithmetic average integral method has the best performance. In this paper, the arithmetic mean miRNA fusion similarity matrixis $SM = (MF + MS + MG)/3$, and the disease fusion similarity matrix is $SD = (DS + DD + DG)/3 \in R^{m\times m}$.

Using autoencoders to learn embedding (dimensionality reduction)

High-dimensional features will seriously affect the prediction performance. Therefore, in this paper, autoencoders are used to extract embedding vectors from nodes of the graph, learn potential features, reconstruct original input data, and obtain low-dimensional representations to improve the prediction accuracy of the model[40,41].

As shown in the Fig. 8, the autoencoder is composed of an encoder and a decoder. In the coding stage, the original feature representation is input to the encoder to achieve feature compression and dimension reduction[42]. The calculation formula is as follows:

$$h = f(W_e \cdot x + b_e) \tag{14}$$

where W_e is the weight of the encoder,x is the input of the original high-dimensional feature,b_e is the bias term,$f(\cdot)$ is the activation function, this article is the ReLu function, the function expression is ReLu(x) = max{0,x}, h is the output of the encoder.

The purpose of the decoder is to reconstruct the input using the latent representation.The calculation is as follows:

$$\hat{x} = g(W_d \cdot h + b_d) \tag{15}$$

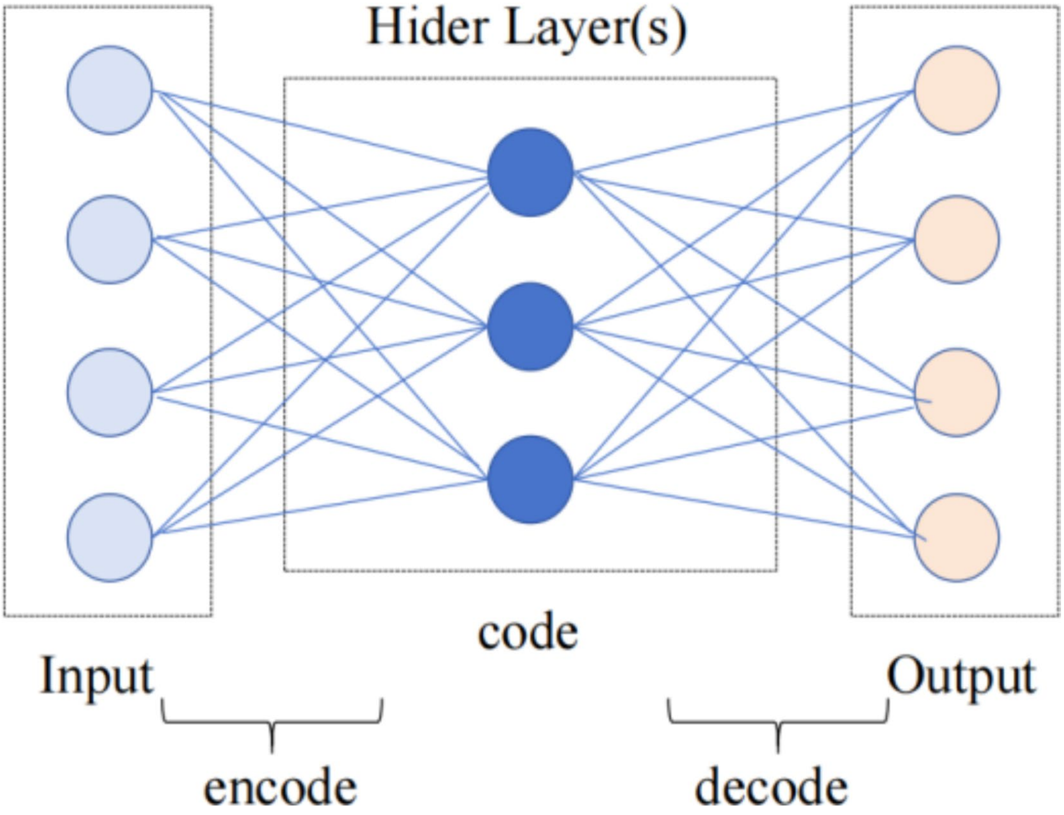


Fig. 8. Schematic diagram of the principle of automatic encoder. The schematic diagram illustrates the encoding and decoding segments, achieving information compression and reconstruction through neurons and interconnections.

where W_d is the weight of the decoder,b_d is the bias term,$g(\cdot)$ is the activation function, this article is the ReLu function, is the input reconstruction feature representation.

Finally, the autoencoder is trained by minimizing the reconstruction error to achieve the purpose of optimizing the loss function. The formula is

$$L = \frac{1}{N}\sum\nolimits_{i=1}^{N}(x_i - \hat{x}_i)^2 \tag{16}$$

where N is the number of sample sets, x_i is the original input data, and $\hat{x}_i$ is the data reconstructed by the automatic encoder.

In this paper, the potential feature output dimension of the autoencoder is set to 64. That is, the final extraction through autoencoders resulted in the reduced feature matrices for miRNA and disease, denoted as $SM\_AE \in R^{n\times 64}$,$SD\_AE \in R^{m\times 64}$。

Complete by inductive matrix

This paper draws on the method of inductive matrix completion for miRNA-disease association prediction. Inductive Matrix Completion (IMC) is a technique that extends the traditional matrix completion problem by introducing auxiliary information (usually the characteristics of matrix rows and columns) to enhance the predictive power of the model. This method is especially suitable for recommender systems and predictive models, where we may not observe all the items, but there is some auxiliary information to assist in the completion of matrix completion[43].

After the above data collection and processing, we obtained human miRNA-disease association matrix A, miRNA similarity matrix SM_AE and disease similarity matrix SD_AE as the feature matrices of miRNA and disease, respectively. Among them, the adjacency matrix A is a very sparse matrix (matrix density 0.068), because between 495 mirnas and 383 diseases, there are only 12,905 experimentally validated human miRNA-disease associations. In addition, the two similar matrices are renamed: SM_AE is named S_m, and SD_AE is named S_d. The goal of IMC is to recover a low-rank matrix $Z \in R^{n\times m}$ using prior knowledge about row and column subspaces when only a few entries are observed. Meanwhile, under the low-rank matrix decomposition idea, Z is of the form $Z = WH^T$, where $W \in R^{n\times r}$, $H \in R^{r\times m}$,r is the desired rank, equal to min(rank(w),rank(H)). The goal is often to more easily integrate row and column features into the process of matrix decomposition, and to allow the model to make predictions on new diseases or new mirnas. The problem of predicting the miRNA-disease association can then be modeled as solving the matrices W and H as solutions to the following optimization problems [12]:

$$\min_{W,H}\ \frac{1}{2}||A - S_m WH^T S_d^T||_F^2 + \frac{\lambda_1}{2}||W||_F^2 + \frac{\lambda_2}{2}||H||_F^2 \tag{17}$$

where,$W \geq 0, H \geq 0, \lambda_1, \lambda_2$ is the regularization parameter, the idea is to make the estimate close to the known correlation matrix A.The final predicted score can be calculated with W and H, and the predicted score between miRNA $m(i)$ and disease $d(j)$ and is calculated as follows:

$$Score(m(i), d(j)) = S_m(i)WH^T S_d^T(j) \tag{18}$$

Optimization algorithm

To solve the optimization problem in Formula 17, we can proceed with the following steps [44] :

(1) Build Lagrangian L.

For the objective function, build Lagrangian function L to introduce the non-negative constraint:

$$L(W, H, \Psi, \Phi) = \Phi(W, H) - tr(\Psi W^T) - tr(\Phi H^T) \tag{19}$$

where,Ψ,Φ is the Lagrange multiplier matrix, ensuring that the elements of W and H are non-negative.

(2) Take the partial derivative

Partial derivative with respect to W:

$$\frac{\partial L}{\partial W} = \frac{\partial \Phi}{\partial W} - \Psi \tag{20}$$

The partial derivative of the objective function with respect to W is obtained:

$$\frac{\partial L}{\partial W} = S_m^T(S_m WH^T S_d - A)S_d H + \lambda_1 W - \Psi \tag{21}$$

Similarly, for partial derivatives of H:

$$\frac{\partial L}{\partial H} = \frac{\partial \Phi}{\partial H} - \Phi \tag{22}$$

Deflecting the objective function with respect to H gives:

$$\frac{\partial L}{\partial H} = S_d^T(S_d H W^T S_m - A^T)S_m W + \lambda_2 H - \Phi \tag{23}$$

(3) Apply KKT conditions

The KKT condition combines the conditions of the Lagrange multiplier and the original variable, for all i, j

There are $\Psi_{i,j} \geq 0,\ \Phi_{i,j} \geq 0,\ W_{i,j} \geq 0,\ H_{i,j} \geq 0$ and complementary relaxation conditions: $\Psi_{ij} W_{ij} = 0$,

$$\Phi_{ij} H_{ij} = 0$$

(4) Derive update rules

According to the KKT conditions, we have:

$$\frac{\partial L}{\partial W} = 0 \Rightarrow S_m^T(S_m W H^T S_d - A)S_d H + \lambda_1 W = \Psi \tag{24}$$

$$\frac{\partial L}{\partial H} = 0 \Rightarrow S_d^T(S_d H W^T S_m - A^T)S_m W + \lambda_2 H = \Phi \tag{25}$$

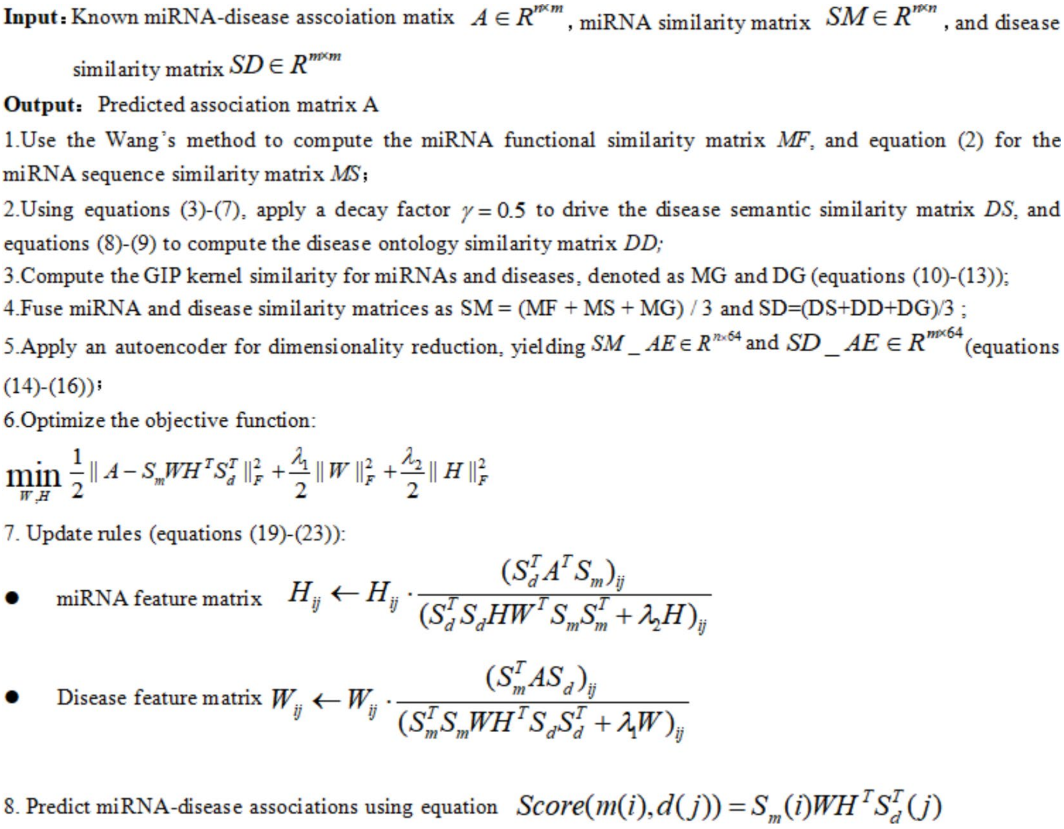
By sum $\Psi_{ij} W_{ij} = 0$ and $\Phi_{ij} H_{ij} = 0$, we can use the multiplication update rule:

$$W_{ij} \leftarrow W_{ij} \cdot \frac{(S_m^T A S_d)_{ij}}{(S_m^T S_m W H^T S_d S_d^T + \lambda_1 W)_{ij}} \tag{26}$$

$$H_{ij} \leftarrow H_{ij} \cdot \frac{(S_d^T A^T S_m)_{ij}}{(S_d^T S_d H W^T S_m S_m^T + \lambda_2 H)_{ij}} \tag{27}$$

These update rules ensure that W and H are kept non-negative when they are updated.

Based on the above equation. We set W and H as random dense matrices, then update the matrices W and H until they converge. Finally, we can predict the Score matrix Y by miRNA-disease association, where $Y_{ij} = score(m(i), d(j))$. To further clarify the implementation process of the proposed method[45], we present the corresponding pseudocode. The training process of AEMC is shown in the Fig. 9.



```
Input: Known miRNA-disease asscoiation matix $A \in R^{n\times m}$, miRNA similarity matrix $SM \in R^{n\times n}$, and disease
    similarity matrix $SD \in R^{m\times m}$
Output: Predicted association matrix A
1.Use the Wang's method to compute the miRNA functional similarity matrix $MF$, and equation (2) for the miRNA sequence similarity matrix $MS$;
2.Using equations (3)-(7), apply a decay factor $\gamma = 0.5$ to drive the disease semantic similarity matrix $DS$, and equations (8)-(9) to compute the disease ontology similarity matrix $DD$;
3.Compute the GIP kernel similarity for miRNAs and diseases, denoted as MG and DG (equations (10)-(13));
4.Fuse miRNA and disease similarity matrices as SM = (MF + MS + MG) / 3 and SD=(DS+DD+DG)/3 ;
5.Apply an autoencoder for dimensionality reduction, yielding $SM\_AE \in R^{n\times 64}$ and $SD\_AE \in R^{m\times 64}$(equations (14)-(16));
6.Optimize the objective function:
$\min_{W,H} \frac{1}{2}\| A - S_m W H^T S_d^T \|_F^2 + \frac{\lambda_1}{2}\| W \|_F^2 + \frac{\lambda_2}{2}\| H \|_F^2$
7. Update rules (equations (19)-(23)):
    ●  miRNA feature matrix $H_{ij} \leftarrow H_{ij} \cdot \frac{(S_d^T A^T S_m)_{ij}}{(S_d^T S_d H W^T S_m S_m^T + \lambda_2 H)_{ij}}$
    ●  Disease feature matrix $W_{ij} \leftarrow W_{ij} \cdot \frac{(S_m^T A S_d)_{ij}}{(S_m^T S_m W H^T S_d S_d^T + \lambda_1 W)_{ij}}$
8. Predict miRNA-disease associations using equation $Score(m(i), d(j)) = S_m(i) W H^T S_d^T(j)$
```

Fig. 9. The pseudocode corresponding to the model training. It demonstrates the key steps in constructing the model and predicting miRNA-disease associations.

Data availability

The code used for the proposed machine learning model and the required datasets can be found at the following location https://github.com/20,001,123/WYW.

Received: 30 June 2024; Accepted: 29 October 2024
Published online: 11 November 2024

References

1. Taguchi, Y.-H. Inference of target gene regulation via miRNAs during cell senescence by using the MiRaGE server, In International Conference on Intelligent Coumputing, Springer, 441–446. https://doi.org/10.1007/978-3-642-31837-5_64(2012).
2. Hua, S., Yun, W., Zhiqiang, Z. & Zou, Q. A discussion of micrornas in cancers. *Curr. Bioinform.* **9**, 453–462. https://doi.org/10.2174/1574893609666140804221135 (2014).
3. Lynam-Lennon, N., Maher, S. G. & Reynolds, J. V. The roles of microRNA in cancer and apoptosis. *Biol. Rev.* **84**, 55–71. https://doi.org/10.1111/j.1469-185X.2008.00061.x (2009).
4. Chen, X. et al. Long non-coding RNAs and complex disease: from experimental results to computational models. *Brief. Bioinform.* **18**, 558–576. https://doi.org/10.1093/bib/bbw060 (2017).
5. Chen, X. et al. NRDTD: a database for clinically or experimentally supported non-coding RNAs and drug targets associations. Database (Oxford), 1–6, https://doi.org/10.1093/database/bax057 (2017).
6. Chen, X. et al. LRSSLMDA: Laplacian regularized sparse subspace learning for MiRNA-disease association prediction. *PLoS Comput. Biol.* **13**, e1005912. https://doi.org/10.1371/journal.pcbi.1005912 (2017).
7. Chen, X. et al. MicroRNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* **20**, 515–539. https://doi.org/10.1093/bib/bbx130 (2019).
8. Jiang, Q. et al. Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst. Biol.* **4**, 1–9. https://doi.org/10.1186/1752-0509-4-S1-S2 (2010).
9. Zeng, X. et al. Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* **34**, 2425–2432. https://doi.org/10.1093/bioinformatics/bty112 (2018).
10. Momanyi, B. M. et al. CFNCM: Collaborative filtering neighborhood-based model for predicting miRNA-disease associations. *Comput. Biol. Med.* https://doi.org/10.1016/j.compbiomed.2023.107165 (2023).
11. Wang, L. et al. LMTRDA: Using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities. *PLoS Comput. Biol.* **15**, e1006865. https://doi.org/10.1371/journal.pcbi.1006865 (2019).
12. Chen, X. et al. Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics* **34**, 4256–4265. https://doi.org/10.1093/bioinformatics/bty503 (2018).
13. Li, J. et al. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. *Bioinformatics* **36**, 2538–2546. https://doi.org/10.1093/bioinformatics/btz965 (2020).
14. Chen, X. et al. MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. *PLoS Comput. Biol.* **14**, e1006418. https://doi.org/10.1371/journal.pcbi.1006418 (2018).

15. Zheng, X., Zhang, C. & Wan, C. MiRNA-Disease association prediction via non-negative matrix factorization based matrix completion. *Signal Process.* **190**, 108312. https://doi.org/10.1016/j.sigpro.2021.108312 (2022).
16. Li, G., Lin, Y., Luo, J., Xiao, Q. & Liang, C. GGAECDA: Predicting circRNA-disease associations using graph autoencoder based on graph representation learning. *Comput Biol Chem.* **99**, 107722. https://doi.org/10.1016/j.compbiolchem.2022.107722 (2022).
17. Guo, Y. et al. Variational gated autoencoder-based feature extraction model for inferring disease-miRNA associations based on multiview features. *Neural Netw.* https://doi.org/10.1016/j.neunet.2023.05.052 (2023).
18. Ding, Y. et al. Variational graph auto-encoders for miRNA-disease association prediction. *Methods* **192**, 25–34. https://doi.org/10.1016/j.ymeth.2020.08.004 (2021).
19. Peng, Y., Zhao, S., Zeng, Z., Hu, X. & Yin, Z. LGBMDF: A cascade forest framework with LightGBM for predicting drug-target interactions. *Front. Microbiol.* **13**, 1092467. https://doi.org/10.3389/fmicb.2022.1092467 (2023).
20. Li, M., Liu, M., Bin, Y. & Xia, J. Prediction of circRNA-disease associations based on inductive matrix completion. *BMC Med. Genomics* **13**, 42. https://doi.org/10.1186/s12920-020-0679-0 (2020).
21. Yan, C. et al. PDMDA: predicting deep-level miRNA–disease associations with graph neural networks and sequence features. *Bioinformatics* **38**, 2226–2234. https://doi.org/10.1093/bioinformatics/btac077 (2022).
22. Han, H. et al. Predicting miRNA-disease associations via layer attention graph convolutional network model. *BMC Med Inform Decis Mak* **22**, 69. https://doi.org/10.1186/s12911-022-01807-8 (2022).
23. Li, G. et al. Predicting miRNA-disease associations based on graph attention network with multi-source information. *BMC Bioinf.* **23**(1), 244. https://doi.org/10.1186/s12859-022-04796-7 (2022).
24. Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **74**, 229–263. https://doi.org/10.3322/caac.21834 (2024).
25. Cui, C. et al. HMDD v4.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkad717 (2023).
26. Wang, D., Wang, J., Lu, M., Song, F. & Cui, Q. Inferring the human microRNA functional similarity and functional network based on MicroRNA-associated diseases. *Bioinformatics* **26**, 1644–1650. https://doi.org/10.1093/bioinformatics/btq241 (2010).
27. Goh, K.-I. et al. The human disease network. *Proc. Natl. Acad. Sci.* **104**, 8685–8690. https://doi.org/10.1073/pnas.0701361104 (2007).
28. Ding, P., Luo, J., Liang, C., Xiao, Q. & Cao, B. Human disease miRNA inference by combining target information based on heterogeneous manifolds. *J. Biomed. Inform.* **80**, 26–36. https://doi.org/10.1016/j.jbi.2018.02.013 (2018).
29. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162. https://doi.org/10.1093/nar/gky1141 (2019).
30. Zuo, Z. L. et al. Double matrix completion for circRNA-disease association prediction. *BMC Bioinf.* **22**, 307. https://doi.org/10.1186/s12859-021-04231-3 (2021).
31. Lipscomb, C. E. Medical subject headings (MeSH). *Bull. Med. Libr. Assoc.* **88**, 265–266 (2000).
32. Wang, D. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* **26**, 1644–1650. https://doi.org/10.1093/bioinformatics/btq241 (2010).
33. Xuan, P. et al. Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PLoS ONE* **8**, e70204. https://doi.org/10.1371/journal.pone.0070204 (2013).
34. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C. F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**, 1274–1281. https://doi.org/10.1093/bioinformatics/btm087 (2007).
35. Li, J. et al. DOSim: An R package for similarity between diseases based on disease ontology. *BMC Bioinf.* **12**, 266. https://doi.org/10.1186/1471-2105-12-266 (2011).
36. Van Laarhoven, T., Nabuurs, S. B. & Marchiori, E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* **27**, 3036–3043. https://doi.org/10.1093/bioinformatics/btr500 (2011).
37. Chen, X. & Yan, G.-Y. Novel human lncRNA–disease association inference based on lncRNA expression profiles. *Bioinformatics* **29**, 2617–2624. https://doi.org/10.1093/bioinformatics/btt426 (2013).
38. Liu, B., Wang, J., Sun, K. & Tsoumakas, G. Fine-grained selective similarity integration for drug–target interaction prediction. *Brief. Bioinform* https://doi.org/10.1093/bib/bbad085 (2023).
39. Ding, Y., Lei, X., Liao, B. & Wu, F. X. MLRDFM: a multi-view Laplacian regularized DeepFM model for predicting miRNA-disease associations. *Brief Bioinform.* **23**, bbac079. https://doi.org/10.1093/bib/bbac079 (2022).
40. Hu, X., Yin, Z., Zeng, Z. & Peng, Y. Prediction of miRNA–disease associations by cascade forest model based on stacked autoencoder. *Molecules* **28**, 5013. https://doi.org/10.3390/molecules28135013 (2023).
41. Ji, C. et al. AEMDA: inferring miRNA-disease associations based on deep autoencoder. *Bioinformatics.* **37**, 66–72. https://doi.org/10.1093/bioinformatics/btaa670 (2021).
42. Nair, V. & Hinton, G.E. Rectified linear units improve restricted Boltzmann machines. In International Conference on Machine Learning, Haifa, Israel, 807–814, https://dl.acm.org/doi/proceedings/https://doi.org/10.5555/3104322 (2010).
43. Natarajan, N. & Dhillon, I. S. Inductive matrix completion for predicting gene-disease associations. *Bioinformatics* **30**, i60–i68. https://doi.org/10.1093/bioinformatics/btu269 (2014).
44. Xiao, Z., Zheng, C., Zhang, C. & Wan, C. MiRNA-Disease association prediction via non-negative matrix factorization based matrix completion. *Signal Processing.* **190**, 108312. https://doi.org/10.1016/j.sigpro.2021.108312 (2022).
45. Li, G., Bai, P., Liang, C. & Luo, J. Node-adaptive graph Transformer with structural encoding for accurate and robust lncRNA-disease association prediction. *BMC Genomics.* **25**(1), 73. https://doi.org/10.1186/s12864-024-09998-2 (2024).

Author contributions

Conceptualization, Y. W. and Z. X.; writing,Y. W.; resources,Z. X.; project administration, Z. X.; funding acquistion, Z. X.

Funding

This work was supported by the National Science Foundation of China (NO:62072296).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.