scientific reports

Check for updates

OPEN

A method for miRNA diffusion association prediction using machine learning decoding of multi-level heterogeneous graph Transformer encoded representations

SiJian Wen[1,3], YinBo Liu[1,3], Guang Yang[1], WenXi Chen[1], HaiTao Wu[1], XiaoLei Zhu[1,4]✉ & YongMei Wang[1,2,4]✉

MicroRNAs (miRNAs) are a key class of endogenous non-coding RNAs that play a pivotal role in regulating diseases. Accurately predicting the intricate relationships between miRNAs and diseases carries profound implications for disease diagnosis, treatment, and prevention. However, these prediction tasks are highly challenging due to the complexity of the underlying relationships. While numerous effective prediction models exist for validating these associations, they often encounter information distortion due to limitations in efficiently retaining information during the encoding-decoding process. Inspired by Multi-layer Heterogeneous Graph Transformer and Machine Learning XGboost classifier algorithm, this study introduces a novel computational approach based on multi-layer heterogeneous encoder—machine learning decoder structure for miRNA-disease association prediction (MHXGMDA). First, we employ the multi-view similarity matrices as the input coding for MHXGMDA. Subsequently, we utilize the multi-layer heterogeneous encoder to capture the embeddings of miRNAs and diseases, aiming to capture the maximum amount of relevant features. Finally, the information from all layers is concatenated to serve as input to the machine learning classifier, ensuring maximal preservation of encoding details. We conducted a comprehensive comparison of seven different classifier models and ultimately selected the XGBoost algorithm as the decoder. This algorithm leverages miRNA embedding features and disease embedding features to decode and predict the association scores between miRNAs and diseases. We applied MHXGMDA to predict human miRNA-disease associations on two benchmark datasets. Experimental findings demonstrate that our approach surpasses several leading methods in terms of both the area under the receiver operating characteristic curve and the area under the precision-recall curve.

Keywords MiRNA-disease association prediction, Multi-view similarity networks, Multi-layer heterogeneous encoder, XGBoost decoder

MicroRNAs (miRNAs) are small regulatory non-coding RNAs produced by a variety of cells in the body and consist of approximately 22 nucleotides[1]. Studies have shown that miRNAs can regulate gene expression[2] and influence a wide range of biological processes. In the past decades, many diseases have been found to be associated with miRNAs[3–6]. For example, Hollams et al.[7] showed that overexpression of c-myc mRNA is associated with tumours, and IGF II mRNA binding to p62 protein is significant to the pathogenesis of hepatocellular carcinoma. Yuko Seko et al.[8] elucidated that prolongation of IL-2 mRNA half-life is closely related to the development of

[1]School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei 230036, China. [2]Anhui Provincial Engineering Laboratory for Beidou Precision Agriculture Information, Hefei 230036, China. [3]These authors contributed equally: SiJian Wen and YinBo Liu. [4]These authors jointly supervised this work: XiaoLei Zhu and YongMei Wang. ✉email: xlzhu_mdl@hotmail.com; wym0152@foxmail.com

autoimmune diseases. Later, more and more researches have found that changes in miRNA expression lead to abnormalities in gene expression and cellular function[9–12], and play a non-negligible regulatory role in pathology. Therefore, in-depth exploration of miRNA expression profiles is of great significance for dissecting pathogenesis and developing new diagnostic and therapeutic options. The majority of miRNAs have been identified through gene chips[13] and high-throughput sequencing technologies[14], but very few have been determined to be associated with diseases. Traditional clinical trials are time-consuming and uncertain, and in order to systematically reveal the effect on miRNA expression of disease, a large number of studies have been designed in recent years to predict the relationships between the two. Depending on the type of data, existing models are divided into two main categories: prediction methods based on a single data type and prediction methods based on multi-view data. Based on the methodology, existing prediction models are mainly classified into four categories: machine learning-based models, scoring function-based models, network topology-based models, and deep learning-based models. Jiang et al.[15] calculated the probability of miRNA involvement in disease using a scoring function via the miRNA functional similarity network and the human phenome-microRNAome network, and optimised the results. Wang et al.[16] constructed acyclic graphs by miRNA functional similarity and disease semantic similarity, and Chen et al.[17] calculated spectral kernel by miRNA and disease Gaussian similarity interaction similarity, however, such models ignore the dynamic information of the network and the heterogeneity of the data. Zeng et al.[18] proposed a structural perturbation method called SPM for bilayer networks, which integrates information from multiple sources and comprehensively optimises the unknown predictive associations between miRNAs and diseases based on the network topology. Then Zhao et al.[19–21] proposed miRNA-disease association prediction models on account of a series of machine learning methods such as XGBoost, but wrong selection of negative samples in such methods can adversely affect the prediction results[22,23]. It bears mentioning that the above methods are only based on a single view and the miRNA-disease associations are incomplete. So as to improve the reliability of relationship prediction by extracting information from different types of similarity views as features of miRNAs and diseases[24], Yin et al.[25] proposed NCPLP, which innovatively integrates disease semantic similarity and microbial functional similarity, utilizing network consistency projection and label propagation techniques to evaluate microbial similarity from different perspectives for precise prediction. Ji et al.[26] adopted an end-to-end approach, combining representation learning and deep autoencoders, to predict association information. SVAEMDA[27] integrates multi-source similarities and utilizes variational autoencoders to train predictors, assessing miRNA-disease associations through reconstruction probabilities, addressing the problem of semi-supervised learning. Liu et al.[28] proposed GCNPCA by exploiting the superior performance of graph neural networks to capture deep topological information from heterogeneous networks. GATMDA[29] exploits the mechanism of graph attention to assign different weights to miRNA neighbour nodes during aggregation to obtain higher-order neighbour information from multivariate associations. Zhou et al.[30] introduced DAEMKL, an innovative approach that harnesses the power of multi-kernel learning to construct intricate miRNA similarity networks and disease similarity networks. Subsequently, features meticulously extracted from the regression models serve as the cornerstone inputs for a sophisticated deep autoencoder to identify the associations. Jin et al.[31] proposed MAMFGAT, which obtains the fusion embedding of miRNAs and diseases through multimodal adaptive fusion, effectively combining the complementary information of the two modalities. Liu et al.[32] proposed TWMHGT to obtain miRNA-disease embeddings by extracting information from heterogeneous graph neural networks and decoding them by matrix multiplication. Yang et al.[33] used multiple heterogeneous networks applied to a graph convolutional network to obtain embedding information from different perspectives, which were fed into a random forest (MGCNRF) to predict potential associations. Recently, HGSMDA[34] integrates HyperGCN, constructs a miRNA-disease heterogeneous hypergraph, trains GCN for information aggregation, and evaluates prediction similarity with Sørensen-Dice loss. Jiao et al.[35] proposed MGADAE, which employs a multi-kernel learning algorithm to construct a similarity heterogeneous network, and then predicts association scores by a graph-convolution encoder, a bilinear decoder. Although the above methods improved the performance of miRNA-disease association prediction to a certain extent, different types of associations are still not fully detected and the embedding features are not completely preserved, resulting in insufficient local information being incorporated into the network models. In this study, we propose a computational method based on multi-layer heterogeneous encoder - machine learning decoder structure for miRNA-disease association prediction, called MHXGMDA. To be more specific, MHXGMDA integrates three similarity knowledge networks firstly, including miRNA-miRNA, miRNA-disease, and disease-disease networks to construct biological feature vectors from the miRNA and disease semantic similarity matrices and Gaussian similarity matrices, respectively, for multi-view encoding to construct biological feature vectors. The embedding features of miRNAs and diseases are fully extracted using the multi-layer heterogeneous encoder, and all layers are spliced to maximise the degree of information retention, which are used as inputs to the XGBoost classifier in the decoding stage to complete the association prediction task. We validated the effectiveness of MHXGMDA on two benchmark datasets using five-fold cross-validation. Experimental results show that MHXGMDA outperforms several state-of-the-art models on several independent metrics. The main contributions of this paper are summarised below:

- Consider biological meta-pathway information. Combining a multi-layer heterogeneous encoder to capture different types of associations provides rich contextual information for encoding complex associative relationships between miRNA-disease and enhances the reliability in the prediction of unknown relationships.
- Embedding information deep fusion. Given XGBoost's significant advantage over most machine learning algorithms in handling embedded features, we utilise XGBoost as a decoder for miRNA-disease feature splicing matrices to further enhance the accuracy and stability of prediction.
- Experimental Validation. We undertake comprehensive experiments across two benchmark datasets to ensure the validity of MHXGMDA and provide constructive comments through case studies with model predictions.

Materials

To validate the generalisability of our model, we downloaded two miRNA-disease datasets for benchmarking from references[36,37], both of which are derived from the Human MicroRNA Disease Database (HMDD) v3.2. The first dataset (VG-data) comes from the work of VGAMF[36], and after de-duplication of relationships includes 8968 relationships between 788 miRNAs and 374 diseases, and the second dataset (DA-data) comes from the work of DAmiRLocGNet[37] and includes 15,547 relationships between 1041 miRNAs and 640 diseases. We found that negative association samples are much more than the positive samples in these two datasets, and there is a large amount of noisy data in the unknown associations. In order to reduce the adverse impact on the noise of the prediction results and to ensure the rationality of the selection of negative samples, as well as the balancing of the dataset, we labelled the positive samples of all the confirmed miRNA-disease associations as 1, and sampled the same number of negative samples at random as the number of positive samples in the remaining negative samples labelled as 0.

Methods

MHXGMDA framework

MiRNA-disease data is often heterogeneous, including different types of entities and complex relationships among them. In order to fully consider the associations between multiple biological entities while effectively retaining the information of the encoding-decoding process, and considering the excellent performance of HGT[38] in heterogeneous data processing, we propose a computational method based on multi-layer heterogeneous encoder - machine learning decoder structure for miRNA-disease association prediction (MHXGMDA). As shown in Fig. 1, MHXGMDA mainly includes three distinct stages:

- Multi-view similarity feature extraction. We constructed homogeneous similarity matrices for miRNAs and diseases separately as inputs.
- Construction of multi-layer heterogeneous graph Transformer. We consider miRNAs and diseases as nodes, and traverse meta-paths in HGT to integrate multiple high-level coding information.
- Splice matrix classification. We apply the direct splicing method to fully fuse all the output features of the multi-layer heterogeneous encoder and decode them with the XGBoost classifier to derive the ultimate prediction outcomes.

Multi-view similarity feature extraction

Based on previous methods[39], we apply Gaussian kernel function to the association network of topology between bioinformatic nodes, so as to obtain miRNA semantic similarity matrix Gaussian interaction profile kernel similarity. Similarly, the disease similarity matrix is obtained by applying Gaussian kernel according to the disease semantic similarity matrix. The multi-view similarity matrices are fused to extract miRNA–miRNA and disease–disease similarity features.

$$A_m = mean\{A_{ms}[S_{ms}], A_{ms}[S_{ms}]\} \tag{1}$$

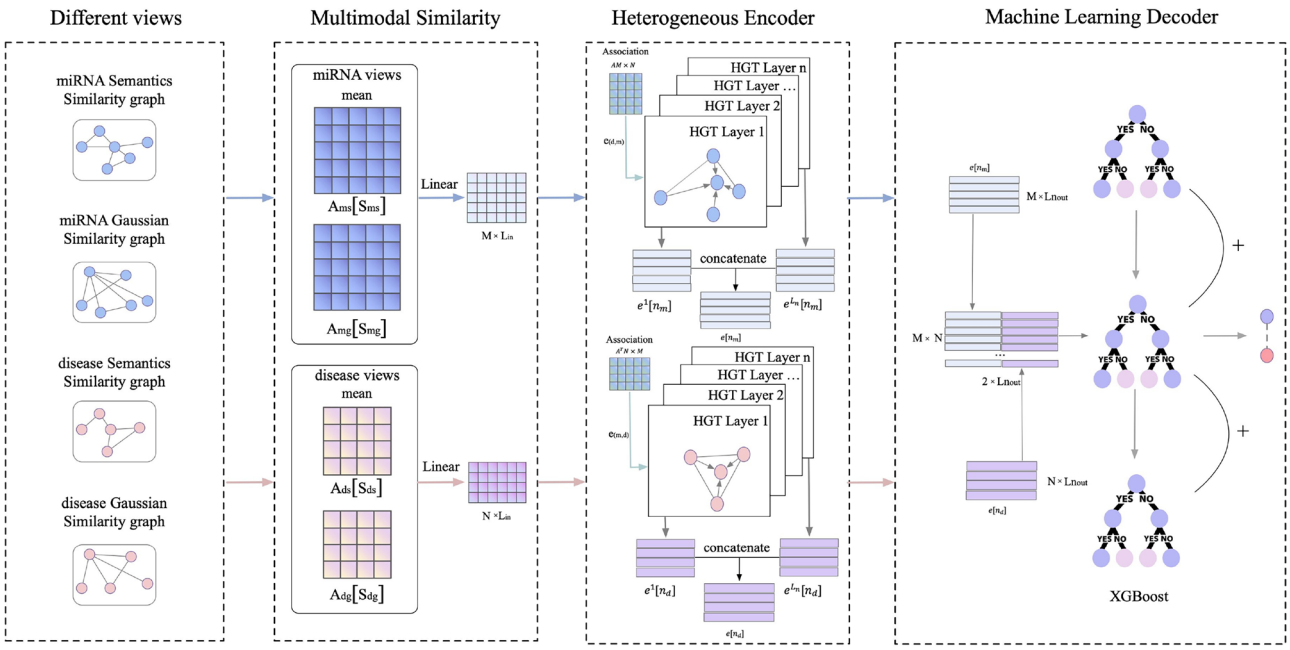

Figure 1. The overall architecture of the MHXGMDA for predicting miRNA-disease association.

$$A_d = mean\{A_{ds}[S_{ds}], A_{ds}[S_{ds}]\} \tag{2}$$

where S_{ms} represents miRNA semantic similarity, its matrix expression $A_{ms} \in R^{M\times M}$, S_{mg} represents miRNA Gaussian similarity, its matrix expression $A_{mg} \in R^{M\times M}$; Similarly, S_{ds} represents disease semantic similarity, its matrix expression $A_{ds} \in R^{D\times D}$, S_{dg} represents disease Gaussian similarity, its matrix expression $A_{dg} \in R^{D\times D}$, the specific calculation method is detailed in the Supplementary Information. *mean* represents the average of the two, which is used to fuse multi-view similarity as the final miRNA similarity matrix A_m and disease similarity matrix A_d.

Construction of multi-layer heterogeneous graph Transformer

Most of the previous methods failed to capture the dynamic property information of heterogeneous graphs, and the design of HGT in heterogeneous graph data processing makes it a powerful tool for dealing with complex relationships and structures, so we use HGT to learn node representations to capture potential features between miRNAs and diseases. This can be divided into three steps: Firstly, Heterogeneous Mutual Attention, the attention weights of the target node miRNA with respect to the disease of each neighbouring source node, are calculated. Specifically:

$$Attention_{HGT}\left(n_d, e_{d,m}, n_m\right) = \underset{\forall d\in N(d)}{Softmax}\left(\underset{i\in[1,h_{th}]}{||} ATT\text{-}head^i\left(n_d, e_{d,m}, n_m\right)\right) \tag{3}$$

$$e^{(l)}[n_d] \leftarrow \underset{\forall d\in N(d),\forall e\in E(m,d)}{Aggregate}\left(Attention\left(n_m, n_d \cdot Message(n_m)\right)\right) \tag{4}$$

$$e^{(l)}[n_m] \leftarrow \underset{\forall d\in N(m),\forall e\in E(d,m)}{Aggregate}\left(Attention\left(n_d, n_m \cdot Message(n_d)\right)\right) \tag{5}$$

$$\sum_{\forall d\in N(m)} Attention_{HGT}\left(n_d, e_{d,m}, n_m\right) = 1_{th\times 1} \tag{6}$$

where *Attention* is used to evaluate the significance of the source node, *Message* extracts information based on the source node, and *Aggregate* aggregates the neighbourhood information through attention weights. v_d denotes the coding of disease, v_m denotes the coding of miRNA, $e_{d,m}$ denotes the edge from the source node (disease) to the target node (miRNA).

For the i-th attention head $ATT\text{-}head^i\left(n_d, e_{d,m}, n_m\right)$ we project the source node d of type $\tau(n_d)$ to generate the i-th key vector $K^i(n_d)$. This linear projection process uses the $K-Linear^i_{\tau(n_d)} : R^{dim} \rightarrow R^{dim}_{h_{th}}$ function,where h_{th} represents the number of attention heads and dim denotes the vector dimension of each head. More specifically, in an effort to cope with different meta relationships,we prepare different mapping matrices, and $K-Linear^i_{\tau(n_d)}$ is indexed according to the type $\tau(n_d)$ of the source node d, which aims to maximally preserve the unique features of various relationships and accurately reflect the relationships between different node types. Similarly, the target node m is linearly projected into $Q-Linear^i_{\tau(n_m)}$ and the i-th query vector is generated,which is designed to help capture the associations between source and target nodes more precisely. The specific calculation formulas are as follows:

$$ATT\text{-}head^i\left(n_d, e_{d,m}, n_m\right) = \left(K^i(n_m) W^{ATT}_{\phi(e_d,m)} Q^i(n_d)^T\right) \cdot \frac{\mu_{\langle\tau(n_d),\phi(e_{d,m}),\tau(n_m)\rangle}}{\sqrt{dim}} \tag{7}$$

$$K^i(n_m) = K\text{-}Linear^i_{\tau(n_d)}\left(h^{(l-1)}[n_d]\right) \tag{8}$$

$$Q^i(n_d) = Q\text{-}Linear^i_{\tau(n_m)}\left(h^{(l-1)}[n_m]\right) \tag{9}$$

Since different meta-paths contribute to the target node to different degrees,for each meta-path triad we set a prior importance weight $\mu_{\langle\tau(n_d),\phi(e_{d,m}),\tau(n_m)\rangle}$, which serves as an adjustment factor for attention. In order to integrate enough information from different source nodes, for each target node m, we collect all the attention vectors from its neighbours $N(m)$.

The next part is Heterogeneous Message Passing, which calculates the information contribution of each source node to the target node. The idea of multi-head information merging is adopted to splice the information of h heads to get the final representation. The specific calculation method is as follows:

$$Message_{HGT}\left(n_d, e_{d,m}, n_m\right) = \underset{i\in[1,e_{th}]}{||} MSG\text{-}head^i(n_d, e_{d,m}, n_m) \tag{10}$$

$$MSG\text{-}head^i(n_d, e_{d,m}, n_m) = \text{M-Linear}^i_{\tau(n_d)}(e^{(l-1)}[n_d]) W^{MSG}_{\phi(e_{d,m})} \tag{11}$$

Considering the heterogeneity of different edge types in information propagation, we invoke a mapping matrix $M\text{-}Linear^{i}_{\tau(n_d)}$ based on a specific edge type, denoting the i-th key vector obtained by linearly projecting the type $\tau(n_d)$ of the source node n_d. $W^{MSG}_{\phi(e_{d,m})}$ is the weight matrix associated with the edge $e_{d,m}$. The information contributions from source nodes to target nodes are obtained from the view of multiple-heads (multi-heads) and they are combined into a final representation.

The final part is Target-Specific Aggregation. Considering that the result of each single-head attention is softmax operated, which means that the sum of the attention weights of all the source nodes is 1, it is straightforward to use the attention as the weight and perform a weighted summation of the message of all the source nodes to obtain the update vector of the target node.

$$\widetilde{e}^{(l)}[n_m] = \bigoplus_{\forall d \in N(m)} (Attention_{HGT}(n_d, e_{d,m}, n_m) \cdot Message_{HGT}(n_d, e_{d,m}, n_m)) \tag{12}$$

Similarly, to ensure the heterogeneity of the propagated information, the model incorporates corresponding linear matrices in the residual network, representing the embeddings of miRNAs and diseases.

$$e^{(l)}[n_m] = A\text{-}Linear_{\tau(n_m)}(\sigma\left(\widetilde{e}^{(l)}[n_m]\right)) + e^{(l-1)}[n_m] \tag{13}$$

$$e^{(l)}[n_d] = A\text{-}Linear_{\tau(n_d)}(\sigma\left(\widetilde{e}^{(l)}[n_d]\right)) + e^{(l-1)}[n_d] \tag{14}$$

Then, embedded nodes are connected according to the output of each HGT layer to fuse the information between different layers.The feature extraction effect is shown in Fig. 2.

Splice matrix classification

The miRNAs obtained from the multi-layer heterogeneous encoder were spliced with the disease features to obtain a fusion descriptor for miRNA-disease association, with the following details:

$$Z_{ij} = [F_m(i), F_d\left(j\right)] \tag{15}$$

where $F_m(i)$ refers to the vector representation of the i-th miRNA within feature F_m and $F_d\left(j\right)$ refers to the vector representation of the j-th disease within feature F_d.

XGBoost, an efficient gradient boosting framework, fits the data by iteratively adding new decision trees and uses regularisation to control the complexity of the model and to better understand the data and the model through feature subset selection and feature importance assessment. In each decision tree generation, XGBoost uses a gradient-based splitting criterion to select the best division point to improve the accuracy and generalisation of the model. XGBoost is also widely used in bioinformatics research for classification and regression tasks such as gene expression profiling[40], disease prediction[41], and drug response prediction[42]. In this study, we utilise XGBoost as a classifier for fusion descriptors of miRNA and disease embedding features. From one perspective, the loss of encoding information is minimally avoided to improve the accuracy and reliability of the association prediction. From another perspective, XGBoost can effectively handle large-scale miRNA and disease data, optimise model performance and minimise the likelihood of overfitting through the gradient boosting algorithm and regularisation. Consequently, as shown in Fig. 3, our model demonstrates superior generalization capabilities on novel data, thereby enhancing its robustness.

Experiments and results

To access the performance of MHXGMDA in the aspect of miRNA-disease association prediction, we conducted a comparative analysis with seven state-of-the-art baselines on two benchmark datasets: GATECDA[43], MINIMDA[44], AMHMDA[39], VGAMF[36], CGHCN[45] ,HFHLMDA[46], and MGADAE[35].

GATECDA[43] uses graph attention autoencoder (GATE) to extract the high-dimensional feature information to low dimensions, respectively, and the combination is used as an input to the completely connected layer to predict the associations between RNAs and drugs sensitivity.

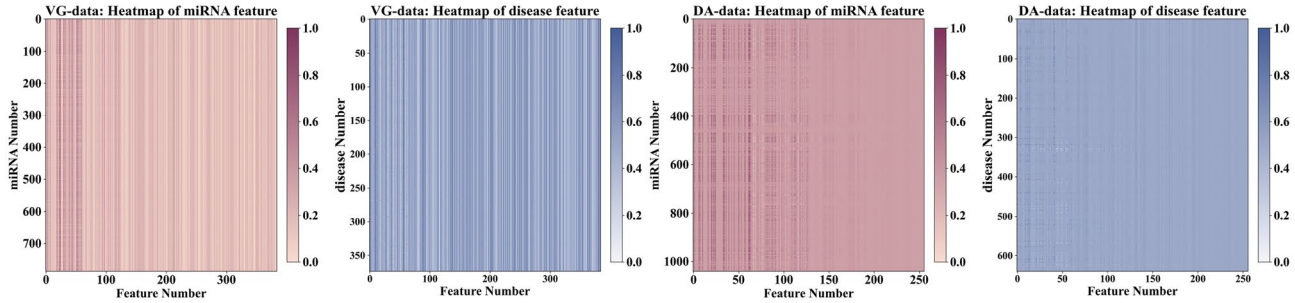

Figure 2. Visualisation of miRNA and disease feature heatmaps. The subplots represent vectors of learned representations of miRNA, disease, with colours indicating the intensity of the individual feature components.

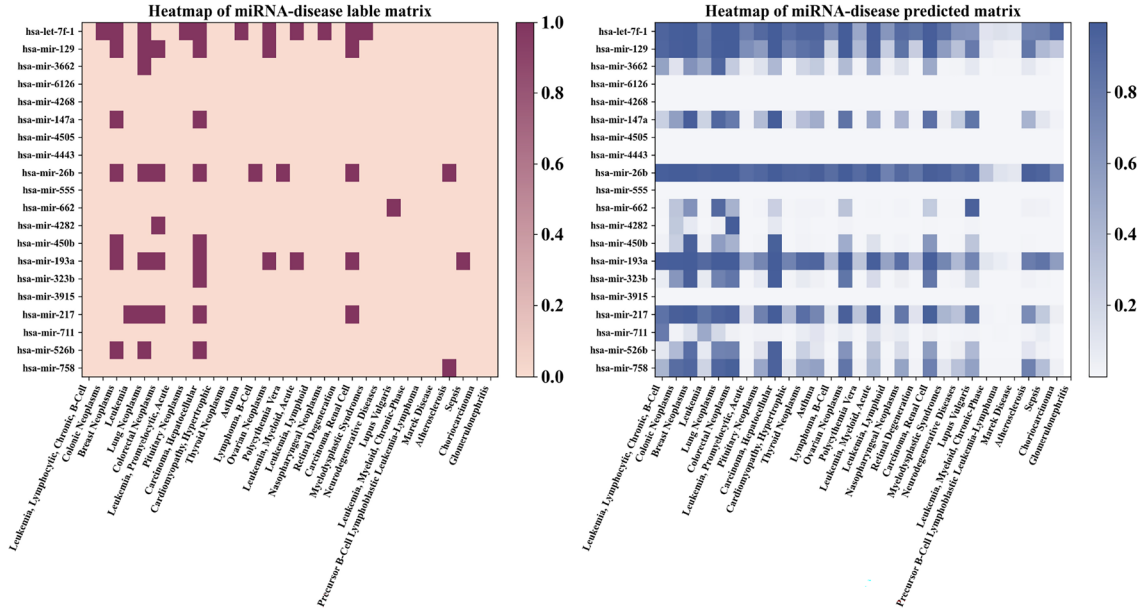

Figure 3. Visualization of predicted score matrix and label matrix heatmaps. The subgraphs respectively depict known and anticipated relationships between miRNAs and diseases. In these heatmaps, the rows represent miRNAs while the columns correspond to various diseases.

MINIMDA[44] constructs an integrated network through multi-source information, obtains embedding representations of miRNAs and diseases through integration of multimodal network's higher-order neighbourhood information, and finally uses a multilayer perceptron (MLP) to predict the latent associations between miRNAs and diseases.

AMHMDA[39] combines the information of multiple similarity networks constructed by extracting the attention mechanism, and then introduces supernodes to construct a heterogeneous hypergraph to enrich the node information, and learns miRNA-disease features through graph convolutional networks.

VGAMF[36] integrates multiple perspectives on miRNAs and diseases through linear weighted fusion, while combining matrix decomposition and variational autoencoder to extract linear and nonlinear features of miRNAs and diseases, and then predict potential miRNA-disease associations.

CGHCN[45] uses a graph convolutional network to capture initial features of miRNAs and diseases, which is combined with a hypergraph convolutive machine network to further learn complex higher-order interaction information.

HFHLMDA[46] constructs hyper-edges for miRNA-disease pairs and their k most relevant neighbours to obtain a hypergraph by the nearest neighbour (KNN) method, and trains a projection matrix to predict the association scores between them.

MGADAE[35] predicts the correlation between miRNAs and diseases by fusing their similarity using multi-core learning. It constructs a heterogeneous network, learns representations through graph convolution, and introduces an attention mechanism to integrate multi-layer representations.

To ensure fairness in comparing results, all methods utilize identical similarity data, encompassing miRNA semantic and Gaussian similarities, as well as disease semantic and Gaussian similarities. AMHMDA[39] incorporates three modalities, while MHXGMDA uses two. To maintain consistency, we utilize miRNA and disease semantic similarities as the third modality for AMHMDA[39] Single-modal models CGHCN[45] and HFHLMDA[46] are trained solely on the miRNA-disease semantic similarity matrix.

Validation set	AUC	PRC	F1-Score	Accuracy	Recall	Specificity	Precision
1	0.9598	0.9551	0.8961	0.8935	0.9150	0.8719	0.8779
2	0.9627	0.9578	0.8994	0.8963	0.9306	0.8622	0.8702
3	0.9586	0.9528	0.8890	0.8860	0.9249	0.8480	0.8558
4	0.9560	0.9499	0.8892	0.8860	0.9198	0.8525	0.8605
5	0.9581	0.9508	0.8881	0.8835	0.9362	0.8320	0.8446
Avg	0.9594 ± 0.0034	0.9539 ± 0.0040	0.8938 ± 0.0056	0.8899 ± 0.0064	0.9256 ± 0.0106	0.8520 ± 0.0200	0.8613 ± 0.0166

Table 1. The 5-fold cross-validation test results of MHXGMDA on VG-DATA.

Experimental setup

In order to verify the generalisation ability of the model, we divided the two benchmark datasets into training (80%) and testing (20%) samples. For the training set, we employed 5-fold cross-validation (5-CV) to fine-tune model parameters and structure. During training, we set the hidden channels to 64, attention heads to 8, and epochs to 2000. We employ the Adam optimiser with an optimal learning rate of 0.01 and a weight decay rate of 0.002. Additionally, dropout of 0.5 was applied to randomly omit neurons, preventing overfitting. Evaluation metrics included AUC, PRC, F1-score, accuracy, recall, specificity, and precision. Table 1 summarizes mean values across multiple experiments on the VG-data dataset, where MHXGMDA achieved AUC and PRC scores of 0.9594 and 0.9539, respectively.

Furthermore, we also performed model testing on DA-data, as shown in Table 2, the AUC and PRC reached 0.9601 and 0.9545, respectively, demonstrating its superior performance.

Parameter discussion

In this study, we learn biological knowledge such as meta-paths in heterogeneous graphs through the Heterogeneous Graph Transformer (HGT) model, and the parameter num-layers regulates the number of layers in the HGT. With the goal of deeply investigating the impact of the parameter num-layers on our model performance, we set different values with the search range of 2, 4, 6, 8, and 10, and tested them on two benchmark datasets, the results are shown in Fig. 4. In general, increasing HGT layers can gradually abstract higher-level feature representations and extract more biological information, but as the number of layers increases, the gradient may gradually disappear or explode during the backpropagation process, resulting in a model that is difficult to train or unstable to train. Eventually, we found that the model performance reaches the best when num-layers is set to 6 on VG-data, therefore, we set num-layers to 6 in all other experiments on VG-data. In addition, we also experimented with 5-CV on DA-data, and the performance reached the best when num-layers is set to 4, similarly, we set num-layers to 6 in all other experiments on DA-data all other experiments set num-layers to 4.

Classifier selection

For the sake of selecting the best classifier adapted to the MHXGMDA model framework during the decoding phase, this section adjusts the relevant parameters of seven machine learning models, including XGBoost, SVM, Random Forest, KNN, Decision Tree, Logistic Regression, and Plain Bayes, to evaluate the performance on 5-CV. On VG-data, our model obtains a comparably high AUC value of up to 0.9594 when XGBoost is used as a classifier, while at the same time, the rest of the evaluated metrics reach high levels compared to other classifiers. We also performed a 5-fold cross-validation on DA-data, and the highest AUC value for XGBoost is 0.9601, which is 0.0066 higher compared to SVM with the second highest score. The average experimental results are shown in Tables 3 and 4, where SVM stands for Support Vector Machine, RF for Random Forest, KNN for K-Nearest Neighbors, LR for Logistic Regression, DT for Decision Tree, and NB for Naive Bayesian.

In conclusion, XGBoost's ability to extract information from miRNA-disease splicing features on our model outperforms other classifiers, and therefore, we chose XGBoost as the best classifier in the MHXGMDA framework.

Comparative analysis of performance with other models

We compare MHXGMDA with seven other state-of-the-art models on two benchmark datasets, for all experimental setups with 5-CV training. Figure 5 shows the AUC, PRC for each model.

Tables 5 and 6 show the average AUC, PRC, F1-score, accuracy, recall, specificity, precision, and running time per epoch for each model. It can be observed that although MHXGMDA fails to outperform the MGADAE method in terms of specificity, precision, and running time, it shows better performance in all other metrics, with average AUCs of 0.0138 and 0.0106 higher than the MGADAE method on VG-data and DA-data, respectively. Compared with the other seven methods in multiple cross-validations, MHXGMDA attained the highest AUC and AUPR values, validating its superiority in association discovery compared to other methods.

Ablation experiments with different network architectures

To verify the effectiveness of heterogeneous graph representation encoding in the MHXGMDA model framework, we propose three model variants, MHXGMDA-w/o Last, MHXGMDA-used HAN, and MHXGMDA-w/o Linear, in which we validate the roles of the one-dimensional splicing network layer, the heterogeneous graph Transformer, and the linear layer, respectively. Among them, MHXGMDA-w/o Last is the model that excludes

Validation set	AUC	PRC	F1-Score	Accuracy	Recall	Specificity	Precision
1	0.9583	0.9523	0.9009	0.8931	0.9636	0.8213	0.8458
2	0.9592	0.9548	0.8943	0.8879	0.9482	0.8276	0.8462
3	0.9614	0.9552	0.8990	0.8956	0.9426	0.8501	0.8593
4	0.9640	0.9602	0.9046	0.8995	0.9433	0.8548	0.8689
5	0.9561	0.9487	0.8930	0.8885	0.9341	0.8433	0.8554
Avg	0.9601 ± 0.0040	0.9545 ± 0.0058	0.8988 ± 0.0058	0.8937 ± 0.0058	0.9489 ± 0.0148	0.8381 ± 0.0168	0.8574 ± 0.0116

Table 2. The 5-fold cross-validation test results of MHXGMDA on DA-DATA.

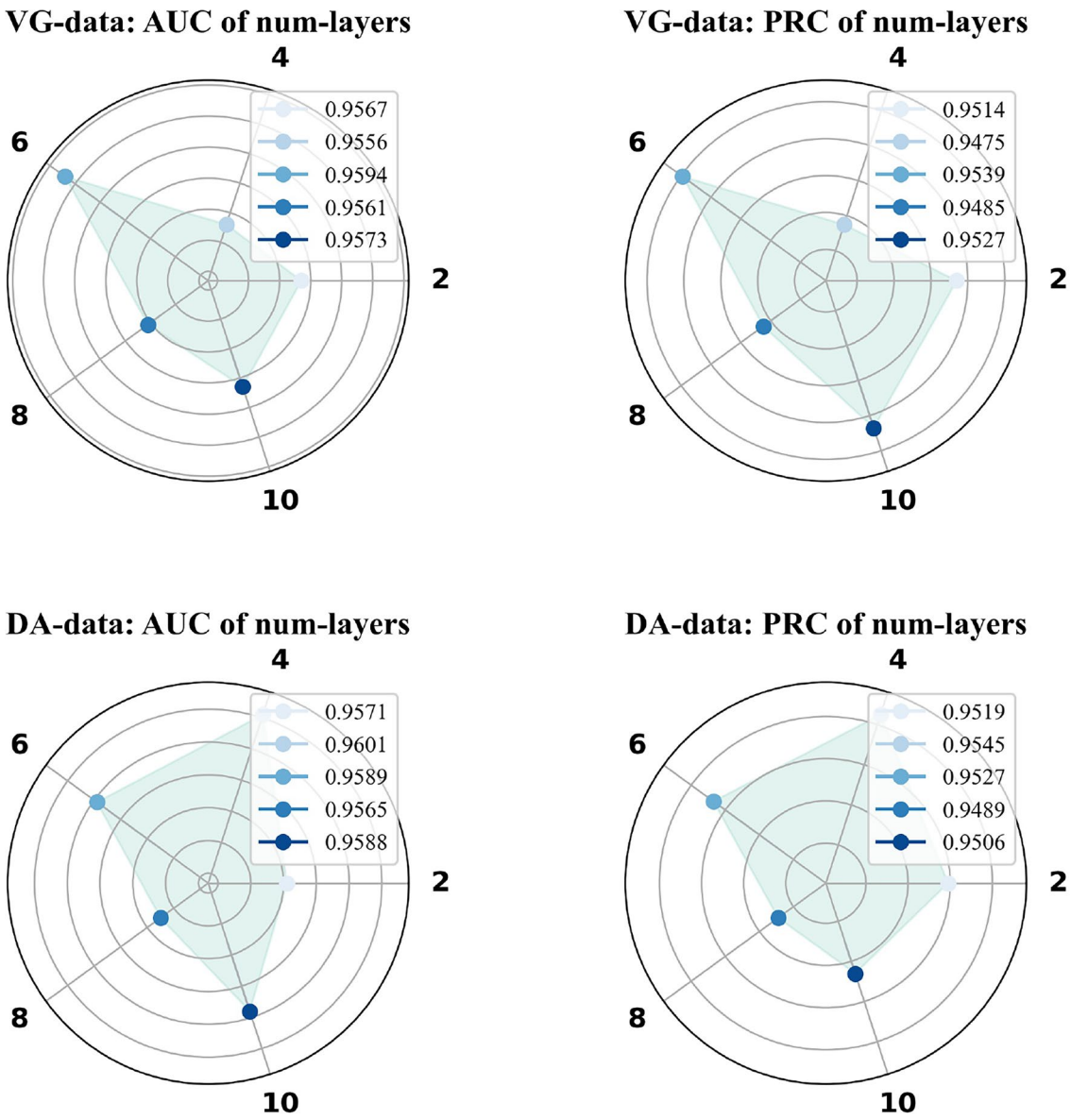

Figure 4. Parameter analysis for num-layers. They denote the AUC and PRC values corresponding to the number of heterogeneous layers of 2, 4, 6, 8 and 10 under the two benchmark datasets, respectively. As the data value increases on each axis, the data point moves further away from the center point.

Classifier	AUC	PRC	F1-Score	Accuracy	Recall	Specificity	Precision
XGBoost	**0.9594±0.0034**	**0.9539±0.0040**	**0.8938±0.0056**	**0.8899±0.0064**	0.9256±0.0106	**0.8520±0.0200**	**0.8613±0.0166**
SVM	0.9550±0.0023	0.9495±0.0028	0.8901±0.0043	0.8829±0.0051	**0.9434±0.0206**	0.8322±0.0230	0.8521±0.0178
RF	0.9540±0.0042	0.9479±0.0064	0.8873±0.0076	0.8814±0.0107	0.9116±0.0172	0.8456±0.0331	0.8568±0.0246
KNN	0.9533±0.0049	0.9459±0.0083	0.8877±0.0059	0.8838±0.0069	0.9219±0.0091	0.8418±0.0195	0.8549±0.0140
LR	0.9445±0.0041	0.9380±0.0073	0.8789±0.0046	0.8753±0.0060	0.9107±0.0131	0.8378±0.0206	0.8475±0.0162
DT	0.9425±0.0061	0.8931±0.0534	0.8724±0.0133	0.8677±0.0149	0.9048±0.0262	0.8250±0.0351	0.8409±0.0202
NB	0.8251±0.0510	0.8648±0.0249	0.8310±0.0219	0.8110±0.0390	0.9221±0.0412	0.6997±0.1202	0.7621±0.0646

Table 3. The evaluation indicators of MHXGMDA with different classifiers on VG-DATA. The optimal values of evaluation indicators are in bold.

Classifier	AUC	PRC	F1-Score	Accuracy	Recall	Specificity	Precision
XGBoost	**0.9601±0.0040**	**0.9545±0.0058**	**0.8988±0.0058**	**0.8937±0.0058**	0.9489±0.0148	0.8381±0.0168	**0.8574±0.0116**
SVM	0.9535±0.0043	0.9421±0.0023	0.8926±0.0079	0.8844±0.0086	**0.9586±0.0019**	0.8099±0.0143	0.8352±0.0124
RF	0.9534±0.0020	0.9467±0.0032	0.8846±0.0042	0.8804±0.0058	0.9102±0.0210	**0.8397±0.0158**	0.8561±0.0213
KNN	0.9503±0.0036	0.9438±0.0033	0.8823±0.0061	0.8731±0.0081	0.9486±0.0108	0.7974±0.0262	0.8251±0.0189
LR	0.9415±0.0046	0.9373±0.0059	0.8716±0.0022	0.8627±0.0116	0.9299±0.0113	0.7952±0.0111	0.8202±0.0128
DT	0.9338±0.0054	0.9269±0.0130	0.8694±0.0151	0.8630±0.0174	0.9089±0.0227	0.8173±0.0397	0.8346±0.0285
NB	0.8286±0.0113	0.8599±0.0069	0.8316±0.0130	0.8174±0.0083	0.8997±0.0224	0.7347±0.0113	0.7731±0.0335

Table 4. The evaluation indicators of MHXGMDA with different classifiers on DA-DATA. The optimal values of evaluation indicators are in bold.

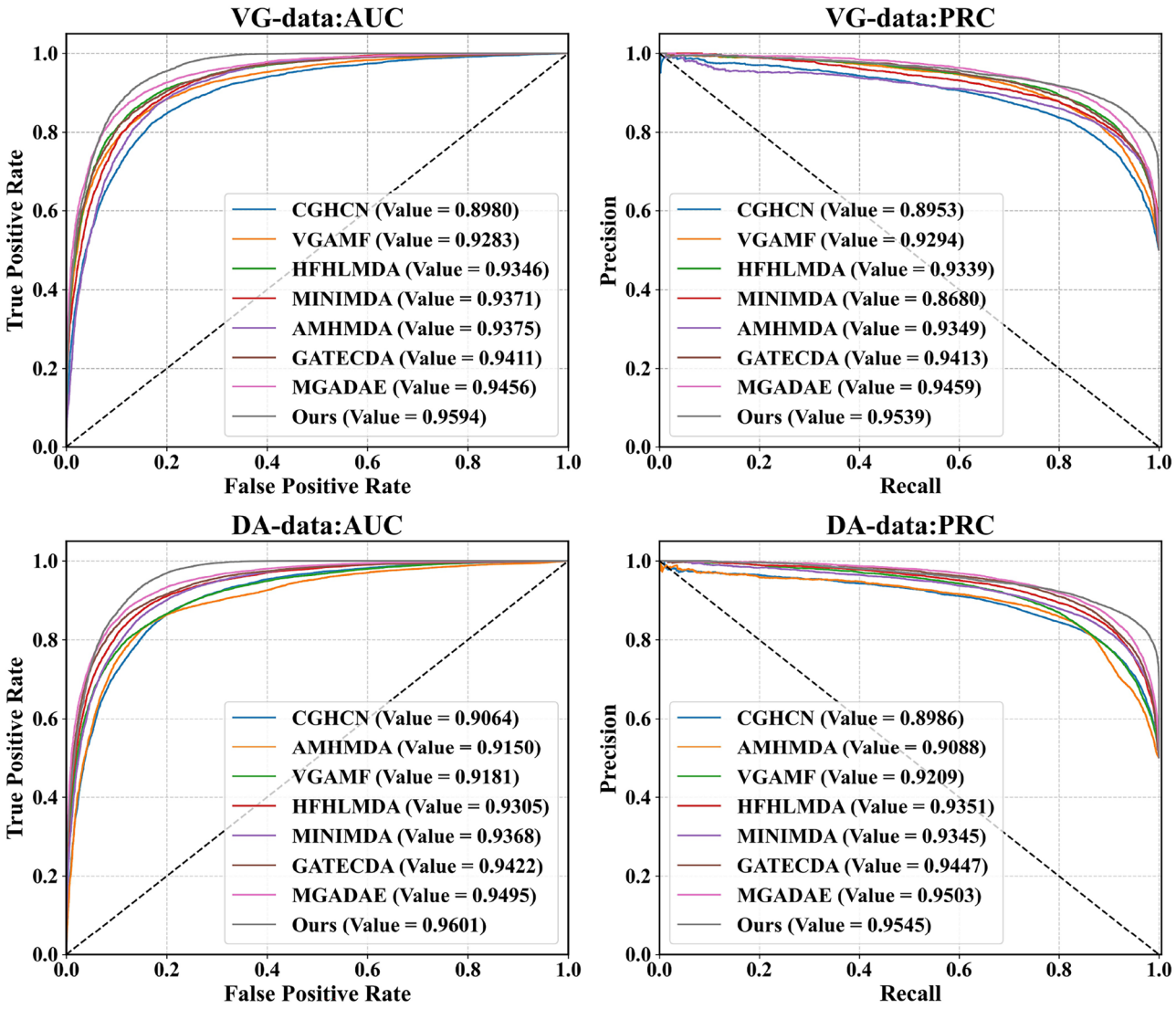

Figure 5. ROC curves and PRC curves plotted by utilizing the cross validation results of different models. ROC curve represents Receiver Operating Characteristic Curve, PRC curve represents Precision-Recall Curve.

the one-dimensional splicing network layer, MHXGMDA-used HAN replaces the Heterogeneous Graph Transformer model (HGT) with the Heterogeneous Graph Attention Network model (HAN), and MHXGMDA-wo Linear refers to the removal of the linear layer in the feed-forward neural network that precedes the multi-layer heterogeneous encoder. As shown in Fig. 6, Tables 7 and 8, MHXGMDA outperforms the other three variants of the model, with AUCs of 0.8886, 0.9212, and 0.9135 for the variants on VG-data, and 0.9578, 0.9595, and 0.9572 on DA-data, respectively. In addition, in all of the evaluated metrics, MHXGMDA-used HAN can significantly

Model	AUC	PRC	F1-Score	Accuracy	Recall	Specificity	Precision	Time (s)
CGHCN	0.8980±0.0055	0.8953±0.0032	0.8311±0.0060	0.8228±0.0097	0.8718±0.0144	0.7746±0.0193	0.7950±0.0185	**0.0074**
VGAMF	0.9283±0.0024	0.9294±0.0019	0.8586±0.0026	0.8678±0.0254	0.8881±0.0289	0.8075±0.0517	0.8390±0.0330	0.0428
HFHLMDA	0.9346±0.0056	0.9339±0.0095	0.8660±0.0082	0.8620±0.0110	0.8933±0.0167	0.8310±0.0125	0.8407±0.0116	104.8440
MINIMDA	0.9371±0.0090	0.8680±0.0125	0.8656±0.0152	0.8835±0.0090	0.8477±0.0189	**0.8530±0.0110**	0.8455±0.0244	0.1937
AMHMDA	0.9375±0.0071	0.9349±0.0068	0.8675±0.0045	0.8610±0.0127	0.9086±0.0097	0.8131±0.0101	0.8305±0.0107	0.2231
GATECDA	0.9411±0.0075	0.9413±0.0082	0.8716±0.0074	0.8694±0.0087	0.8873±0.0131	0.8510±0.0096	0.8569±0.0236	1.6756
MGADAE	0.9456±0.0032	0.9459±0.0061	0.8772±0.0080	0.8743±0.0091	0.8978±0.0113	0.8509±0.0101	0.8582±0.0198	0.1084
MHXGMDA	**0.9594±0.0034**	**0.9539±0.0040**	**0.8938±0.0056**	**0.8899±0.0064**	**0.9256±0.0106**	0.8520±0.0200	**0.8613±0.0166**	0.1530

Table 5. Performance of MHXGMDA with other seven models on VG-DATA. The optimal values of evaluation indicators are in bold.

Model	AUC	PRC	F1-Score	Accuracy	Recall	Specificity	Precision	Time(seconds)
CGHCN	0.9064±0.0036	0.8986±0.0058	0.8391±0.0073	0.8323±0.0092	0.8747±0.0121	0.7900±0.0107	0.8064±0.0182	**0.0077**
AMHMDA	0.9150±0.0032	0.9088±0.0047	0.8548±0.0062	0.8493±0.0046	0.8870±0.0069	0.8118±0.0036	0.8259±0.0147	0.5699
VGAMF	0.9181±0.0072	0.9209±0.0054	0.8402±0.0101	0.8369±0.0122	0.8528±0.0319	0.8249±0.0375	0.8221±0.0399	0.1159
HFHLMDA	0.9305±0.0061	0.9351±0.0057	0.8660±0.0132	0.8620±0.0123	0.8916±0.0165	0.8324±0.0132	0.8419±0.0192	514.2976
MINIMDA	0.9368±0.0060	0.9345±0.0066	0.8643±0.0089	0.8611±0.0087	0.8848±0.0117	0.8373±0.0082	0.8448±0.0099	0.3648
GATECDA	0.9422±0.0088	0.9447±0.0079	0.8709±0.0092	0.8690±0.0107	0.8835±0.0097	0.8545±0.0093	0.8587±0.0111	2.1156
MGADAE	0.9495±0.0056	0.9503±0.0077	0.8815±0.0087	0.8799±0.0102	0.8934±0.0112	**0.8663±0.0101**	**0.8701±0.0165**	0.1583
MHXGMDA	**0.9601±0.0040**	**0.9545±0.0058**	**0.8988±0.0058**	**0.8937±0.0058**	**0.9489±0.0148**	0.8381±0.0168	0.8574±0.0116	0.1649

Table 6. Performance of MHXGMDA with other seven models on DA-DATA. The optimal values of evaluation indicators are in bold.

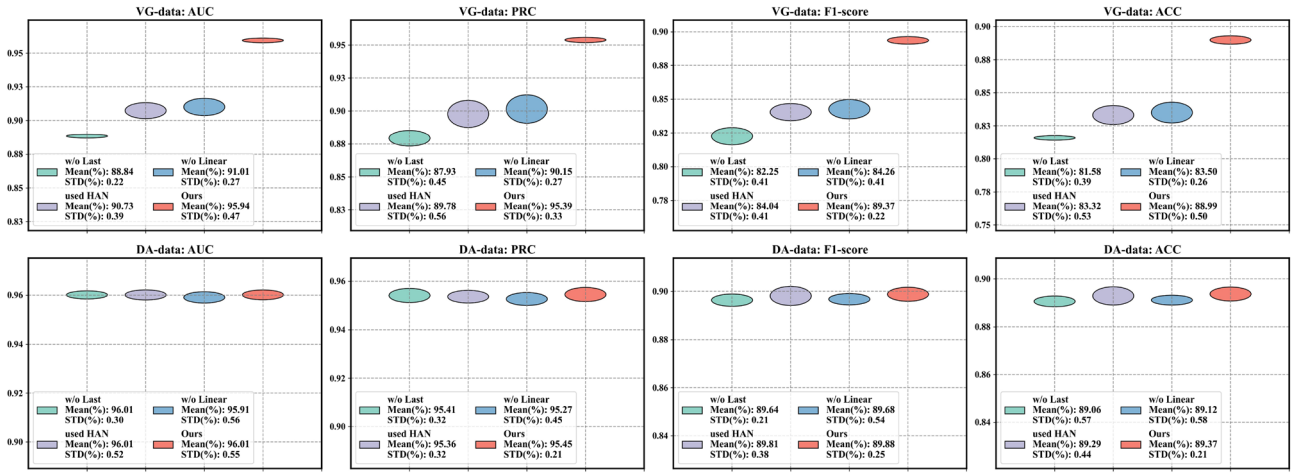

Figure 6. Ablation experience results on different network architectures of MHXGMDA. Mean denotes the mean, STD denotes the standard deviation, the vertical axis denotes the corresponding values of the evaluation indicators under each variant.

Model	AUC	PRC	F1-Score	Accuracy	Recall	Specificity	Precision
MHXGMDA-w/o Last	0.8884±0.0028	0.8793±0.0159	0.8225±0.0105	0.8158±0.0052	0.8758±0.0361	0.7507±0.0498	0.7910±0.0226
MHXGMDA-w/o Linear	0.9073±0.0128	0.8978±0.0215	0.8404±0.0143	0.8332±0.0157	0.8989±0.0281	0.7820±0.0320	0.8060±0.0232
MHXGMDA-used HAN	0.9101±0.0118	0.9015±0.0205	0.8426±0.0127	0.8350±0.0143	0.9127±0.0363	0.7874±0.0500	0.8227±0.0254
MHXGMDA	**0.9594±0.0034**	**0.9539±0.0040**	**0.8938±0.0056**	**0.8899±0.0064**	**0.9256±0.0106**	**0.8520±0.0200**	**0.8613±0.0166**

Table 7. Ablation experiment results on different network architectures of MHXGMDA on VG-DATA. The optimal values of evaluation indicators are in bold.

Model	AUC	PRC	F1-Score	Accuracy	Recall	Specificity	Precision
MHXGMDA-w/o Linear	0.9591±0.0046	0.9527±0.0053	0.8968±0.0047	0.8912±0.0041	0.9356±0.0165	0.8340±0.0167	0.8483±0.0117
MHXGMDA-w/o Last	0.9601±0.0033	0.9541±0.0058	0.8964±0.0051	0.8906±0.0045	0.9437±0.0176	0.8384±0.0253	0.8506±0.0190
MHXGMDA-used HAN	0.9601±0.0041	0.9536±0.0052	0.8981±0.0079	0.8929±0.0076	0.9395±0.0083	**0.8568±0.0100**	**0.8625±0.0149**
MHXGMDA	**0.9601±0.0040**	**0.9545±0.0058**	**0.8988±0.0058**	**0.8937±0.0058**	**0.9489±0.0148**	0.8381±0.0168	0.8574±0.0116

Table 8. Ablation experiment results on different network architectures of MHXGMDA on DA-DATA. The optimal values of evaluation indicators are in bold.

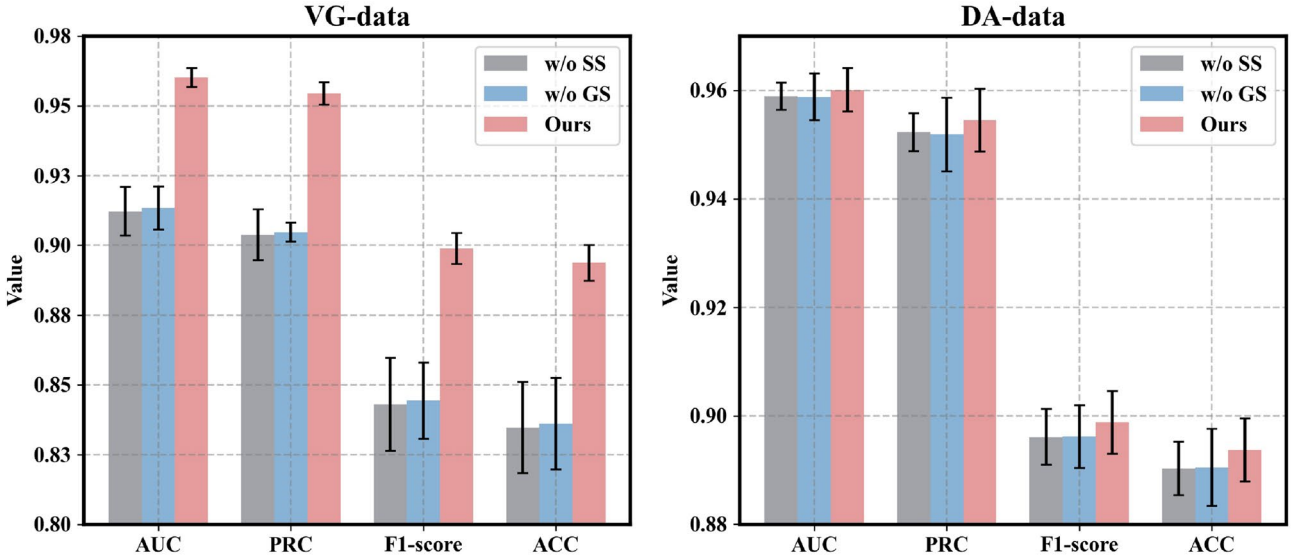

Figure 7. Ablation experience results on different views of MHXGMDA. The figure shows four indicator values of variant models from different views under two benchmark datasets, respectively, where the train data of w/o SS only includes Gaussian similarity matrix, w/o GS only includes semantic similarity matrix, and Ours includes both.

Model	AUC	PRC	F1-Score	Accuracy	Recall	Specificity	Precision
MHXGMDA-w/o SS	0.9122±0.0087	0.9038±0.0091	0.8430±0.0167	0.8347±0.0164	0.8947±0.0423	0.7753±0.0428	0.7944±0.0247
MHXGMDA-w/o GS	0.9133±0.0077	0.9047±0.0034	0.8444±0.0137	0.8361±0.0164	0.8921±0.0330	0.7771±0.0296	0.8092±0.0193
MHXGMDA	**0.9594±0.0034**	**0.9539±0.0040**	**0.8938±0.0056**	**0.8899±0.0064**	**0.9256±0.0106**	**0.8520±0.0200**	**0.8613±0.0166**

Table 9. Ablation experiment results on different views of MHXGMDA on VG-DATA. The optimal values of evaluation indicators are in bold.

Model	AUC	PRC	F1-Score	Accuracy	Recall	Specificity	Precision
MHXGMDA-w/o SS	0.9589±0.0025	0.9522±0.0035	0.8961±0.0051	0.8903±0.0049	**0.9507±0.098**	0.8226±0.0111	0.8430±0.0129
MHXGMDA-w/o GS	0.9588±0.0043	0.9518±0.0068	0.8962±0.0058	0.8905±0.0071	0.9382±0.0126	**0.8394±0.0232**	0.8554±0.0119
MHXGMDA	**0.9601±0.0040**	**0.9545±0.0058**	**0.8988±0.0058**	**0.8937±0.0058**	0.9489±0.0148	0.8381±0.0168	**0.8574±0.0116**

Table 10. Ablation experiment results on different views of MHXGMDA on DA-DATA. The optimal values of evaluation indicators are in bold.

outperform MHXGMDA-w/o Last, which indicates that the one-dimensional splicing network layer has the ability to fully learn the node representations, while the multi-layer heterogeneous graph Transformer can further enhance the model performance.

Ablation experiments with different views

In order to access the rationality of including multimodal training data in MHXGMDA, we implemented two variants of the model ignoring multiple modalities, MHXGMDA-w/o SS and MHXGMDA-w/o GS. Specifically, MHXGMDA-w/o SS is trained without the miRNA semantic similarity matrix, the disease semantic similarity matrix. While the training data of MHXGMDA-w/o GS only excludes miRNA, disease Gaussian similarity matrix. The experimental results are shown in Fig. 7, Tables 9 and 10. On both datasets, almost all the metrics tested by the MHXGMDA model are significantly better than the single-modal variants MHXGMDA-w/o SS and MHXGMDA-w/o GS, which implies that combining the multimodal data is significant to the prediction of miRNA-disease relationships.

Case study

To evaluate the accuracy of MHXGMDA in predicting miRNA-disease associations in real cases, we chose three different diseases: Lung Neoplasms, Carcinoma, Hepatocellular and Glioblastoma as case study subjects. Firstly, we deleted all miRNAs associated with the above three diseases during training, Subsequently, the model's ability to recover deleted associations during the prediction process is evaluated. Then, we ranked the association scores of the three disease-related miRNAs predicted by MHXGMDA and chose the top 20 miRNAs. For the sake of simplicity, we abbreviated HMDD v4.0 as 'H4' in Table 11.

Numerous studies have demonstrated a close association between alterations in miRNA expression levels and the progression of diverse diseases. One of these diseases is lung tumours, one of the common malignant tumours, and the co-expression of hsa-miR-182 and hsa-miR-126 helps to differentiate between primary lung tumours and lung metastases[47]. The second group of diseases in the case study is hepatocellular carcinoma, one of the deadliest forms of cancer in the world, and it has been found that decreased levels of miR-16 and miR-199a expression in the serum of patients exhibit a robust linkage with the progression of hepatocellular carcinoma[48]. In addition, glioblastoma is one of the most common types of fatal brain tumours, where the tumour compresses, infiltrates, and destroys brain tissue, leading to local symptoms and neurological impairment. It has been shown that the microRNA-302-367 cluster effectively leads to the destruction of glioma initiating cells and their tumorigenic properties[49]. Experiments showed that nearly all possible associations forecasted by the model could be verified, which sufficiently demonstrated the excellent performance and reliability of MHXGMDA in actually exploring miRNA-disease associations.

Finally, we focused on the three diseases mentioned above and used them as the central nodes to construct the network by carefully selecting the miRNAs that ranked in the top ten of their respective scores. As shown in Fig. 8, the finding that lung tumours and glioblastoma exhibited the highest number of identical miRNAs in the top ten scores is quite striking. However, it is more noteworthy that despite their significant commonalities

Lung neoplasms			Carcinoma, Hepatocellular			Glioblastoma		
Ranking	miRNA	Evidence	Ranking	miRNA	Evidence	Ranking	miRNA	Evidence
1	hsa-mir-31	H4	1	hsa-mir-21	H4	1	hsa-mir-223	H4
2	hsa-mir-150	H4	2	hsa-mir-10b	H4	2	hsa-mir-21	H4
3	hsa-mir-542	H4	3	hsa-mir-132	H4	3	hsa-mir-34a	H4
4	hsa-mir-146a	H4	4	hsa-mir-222	H4	4	hsa-mir-155	H4
5	hsa-mir-34a	H4	5	hsa-mir-146a	H4	5	hsa-mir-146a	H4
6	hsa-mir-155	H4	6	hsa-mir-215	H4	6	hsa-mir-150	Unconfirmed
7	hsa-mir-25	H4	7	hsa-mir-30b	H4	7	hsa-mir-20a	H4
8	hsa-mir-200b	H4	8	hsa-mir-29c	H4	8	hsa-mir-19a	H4
9	hsa-mir-223	H4	9	hsa-mir-211	H4	9	hsa-mir-18a	H4
10	hsa-mir-29a	H4	10	hsa-mir-421	H4	10	hsa-mir-140	H4
11	hsa-mir-200c	H4	11	hsa-mir-296	H4	11	hsa-mir-16-1	H4
12	hsa-mir-21	H4	12	hsa-mir-455	H4	12	hsa-mir-30e	H4
13	hsa-mir-200a	H4	13	hsa-mir-199a	H4	13	hsa-mir-193a	H4
14	hsa-mir-149	H4	14	hsa-mir-23b	H4	14	hsa-mir-92b	H4
15	hsa-mir-590	H4	15	hsa-mir-204	H4	15	hsa-mir-29b	H4
16	hsa-mir-9	H4	16	hsa-mir-101	H4	16	hsa-mir-542	H4
17	hsa-mir-424	H4	17	hsa-mir-195	H4	17	hsa-mir-149	H4
18	hsa-let-7f-1	H4	18	hsa-mir-34a	H4	18	hsa-mir-29a	H4
19	hsa-mir-141	H4	19	hsa-mir-9	H4	19	hsa-mir-373	H4
20	hsa-mir-483	H4	20	hsa-mir-218-1	H4	20	hsa-mir-129	H4

Table 11. TOP 20 miRNA-disease resistance associations predicted by MHXGMDA.

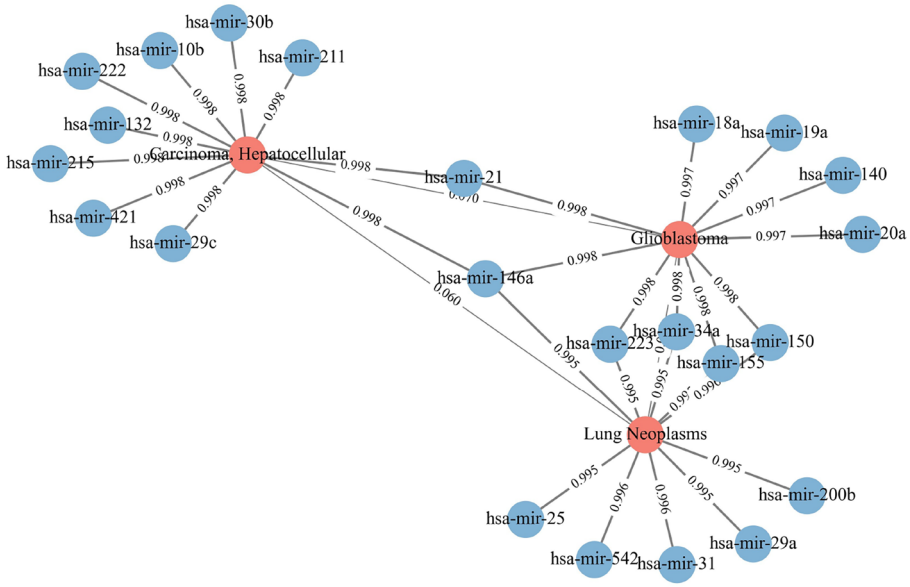

Figure 8. miRNA-disease association subnetwork. The red nodes represent the three diseases and the blue nodes represent the top10 miRNAs associated with the diseases.

at the miRNA level, lung tumours and glioblastomas show relatively little similarity in disease characteristics, and thus the same miRNAs may play different roles in different diseases.

Conclusion

In this work, we propose MHXGMDA, a computational method based on multi-layer heterogeneous encoder—machine learning decoder structure for miRNA-disease association prediction. Compared with existing prediction models, MHXGMDA not only captures different types of biometric knowledge from multi-view similarity of miRNAs and diseases, but also incorporates a multi-layer heterogeneous graph Transformer at the encoding stage to explore the dynamic information of miRNA-disease associations. In the decoding process, MHXGMDA applies XGBoost to learn miRNA-disease key features from multi-layer HGTs to deeply fuse the embedding information. Finally, we tested the model experimentally on two association datasets. The experiments demonstrate that our model surpasses state-of-the-art methods in exploring overlooked miRNA-disease associations, validating the proficiency of MHXGMDA in identifying miRNA-disease associations and helping to pioneer new disease diagnosis and treatment options. However, the fact that negative samples with sufficient experimental evidence of weak correlation between miRNAs and diseases are difficult to collect, and the random sampling of negative samples in the MHXGMDA dataset may have unintended negative impacts, and in our future work, we will construct a balanced dataset with more reliable negative samples to optimise the prediction of miRNA-disease associations.

Key Points

- We construct a multi-layer heterogeneous graph Transformer model based on similarity matrices from multi-views, covering miRNA semantic similarity, disease similarity, spectral kernel similarity of miRNA Gaussian interactions, and spectral kernel similarity of disease Gaussian interactions. In the heterogeneous graph Transformer, we traverse different meta-paths with miRNAs and diseases as nodes to capture richer dynamic information.
- In order to fully aggregate miRNA-disease embedding features, we spliced all the representational matrices of the multilayer HGT outputs as inputs to the XGBoost machine learning model, making maximum use of the encoding-decoding process information.
- We applied MHXGMDA to compute the association scores of missing miRNAs with diseases at 5-CV. The findings reveal that, in comparison to other advanced methods, our method provides a more promising approach to predict the association between miRNAs and diseases.

Data availibility

The datasets used in this article are all based on publicly available datasets mentioned in the Materials, which are available at https://github.com/yinboliu-git/MHXGMDA.

Code availability

The code is publicly available at https://github.com/yinboliu-git/MHXGMDA.

Received: 13 March 2024; Accepted: 29 July 2024
Published online: 03 September 2024

References

1. Bartel, D. P. Micrornas: Genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
2. Gulyaeva, L. F. & Kushlinskiy, N. E. Regulatory mechanisms of microrna expression. *J. Transl. Med.* **14**, 143 (2016).
3. Ciafre, S. *et al.* Extensive modulation of a set of micrornas in primary glioblastoma. *Biochem. Biophys. Res. Commun.* **334**, 1351–1358 (2005).
4. Amiel, J., de Pontual, L. & Henrion-Caude, A. Mirna, development and disease. *Adv. Genet.* **80**, 1–36 (2012).
5. Ladd, A. N. New insights into the role of RNA-binding proteins in the regulation of heart development. *Int. Rev. Cell Mol. Biol.* **324**, 125–185 (2016).
6. Geekiyanage, H. & Galanis, E. Mir-31 and mir-128 regulates poliovirus receptor-related 4 mediated measles virus infectivity in tumors. *Mol. Oncol.* **10**, 1387–1403 (2016).
7. Hollams, E. M., Giles, K. M., Thomson, A. M. & Leedman, P. J. Mrna stability and the control of gene expression: Implications for human disease. *Neurochem. Res.* **27**, 957–980 (2002).
8. Seko, Y., Cole, S., Kasprzak, W., Shapiro, B. A. & Ragheb, J. A. The role of cytokine mRNA stability in the pathogenesis of autoimmune disease. *Autoimmun. Rev.* **5**, 299–305 (2006).
9. Palanichamy, J. K. & Rao, D. S. miRNA dysregulation in cancer: Towards a mechanistic understanding. *Front. Genet.* **5**, 81746 (2014).
10. Kawahara, Y. Human diseases caused by germline and somatic abnormalities in microrna and microrna-related genes. *Congenit. Anomalies* **54**, 12–21 (2014).
11. Szymczyk, A., Macheta, A. & Podhorecka, M. Abnormal microrna expression in the course of hematological malignancies. *Cancer Manag. Res.* 4267–4277 (2018).
12. Ali Syeda, Z., Langden, S. S. S., Munkhzul, C., Lee, M. & Song, S. J. Regulatory mechanism of microrna expression in cancer. *Int. J. Mol. Sci.* **21**, 1723 (2020).
13. Keller, P. *et al.* Gene-chip studies of adipogenesis-regulated microrna s in mouse primary adipocytes and human obesity. *BMC Endocr. Disord.* **11**, 1–11 (2011).
14. Zhu, E. *et al.* mirtools: microrna profiling and discovery based on high-throughput sequencing. *Nucleic Acids Res.* **38**, W392–W397 (2010).
15. Jiang, Q. *et al.* Prioritization of disease microrna s through a human phenome-microrna ome network. *BMC Syst. Biol.* **4**, 1–9 (2010).
16. 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 (2010).
17. Chen, X. & Yan, G.-Y. Novel human lncrna-disease association inference based on lncrna expression profiles. *Bioinformatics* **29**, 2617–2624 (2013).
18. Zeng, X., Liu, L., Lü, L. & Zou, Q. Prediction of potential disease-associated micrornas using structural perturbation method. *Bioinformatics* **34**, 2425–2432 (2018).
19. Chen, X., Wang, C.-C., Yin, J. & You, Z.-H. Novel human mirna-disease association inference based on random forest. *Mol. Ther. Nucleic Acids* **13**, 568–579 (2018).
20. Zhao, Y., Chen, X. & Yin, J. Adaptive boosting-based computational model for predicting potential mirna-disease associations. *Bioinformatics* **35**, 4730–4738 (2019).
21. Xuan, P. *et al.* Prediction of potential disease-associated micrornas based on random walk. *Bioinformatics* **31**, 1805–1815 (2015).
22. Pal, M. K. *et al.* Microrna: A new and promising potential biomarker for diagnosis and prognosis of ovarian cancer. *Cancer Biol. Med.* **12**, 328 (2015).
23. Galvão-Lima, L. J., Morais, A. H., Valentim, R. A. & Barreto, E. J. mirnas as biomarkers for early cancer detection and their application in the development of new diagnostic tools. *Biomed. Eng. Online* **20**, 21 (2021).
24. Yang, M., Wu, G., Zhao, Q., Li, Y. & Wang, J. Computational drug repositioning based on multi-similarities bilinear matrix factorization. *Brief. Bioinform.* **22**, 267 (2021).
25. Yin, M.-M., Liu, J.-X., Gao, Y.-L., Kong, X.-Z. & Zheng, C.-H. Ncplp: A novel approach for predicting microbe-associated diseases with network consistency projection and label propagation. *IEEE Trans. Cybern.* **52**, 5079–5087 (2020).
26. Ji, C. *et al.* Aemda: Inferring mirna-disease associations based on deep autoencoder. *Bioinformatics* **37**, 66–72 (2021).
27. Ji, C. *et al.* A semi-supervised learning method for mirna-disease association prediction based on variational autoencoder. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **19**, 2049–2059 (2021).
28. Liu, J., Kuang, Z. & Deng, L. Gcnpca: mirna-disease associations prediction algorithm based on graph convolutional neural networks. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **20**, 1041–1052 (2022).
29. Li, G. *et al.* Predicting mirna-disease associations based on graph attention network with multi-source information. *BMC Bioinform.* **23**, 244 (2022).
30. Zhou, F. *et al.* Predicting mirna-disease associations through deep autoencoder with multiple kernel learning. *IEEE Trans. Neural Netw. Learn. Syst.* **34**, 5570–5579 (2021).
31. Jin, Z. *et al.* Predicting mirna-disease association via graph attention learning and multiplex adaptive modality fusion. *Comput. Biol. Med.* **169**, 107904 (2024).
32. Liu, Y. *et al.* mirna-disease association prediction based on heterogeneous graph transformer with multi-view similarity and random auto-encoder. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 885–888 (IEEE, 2023).
33. Yang, Y. *et al.* Mgcnrf: Prediction of disease-related mirnas based on multiple graph convolutional networks and random forest. In *IEEE Transactions on Neural Networks and Learning Systems* (2023).
34. Chang, Z., Zhu, R., Liu, J., Shang, J. & Dai, L. Hgsmda: mirna-disease association prediction based on hypergcn and sørensen-dice loss. *Non-coding RNA* **10**, 9 (2024).
35. Jiao, C.-N. *et al.* Multi-kernel graph attention deep autoencoder for mirna-disease association prediction. *IEEE J. Biomed. Health Inform.* (2023).
36. Ding, Y., Lei, X., Liao, B. & Wu, F.-X. Predicting mirna-disease associations based on multi-view variational graph auto-encoder with matrix factorization. *IEEE J. Biomed. Health Inform.* **26**, 446–457 (2021).
37. Bai, T., Yan, K. & Liu, B. Damirlocgnet: mirna subcellular localization prediction by combining mirna—Disease associations and graph convolutional networks. *Brief. Bioinform.* **24**, bbad212 (2023).
38. Hu, Z., Dong, Y., Wang, K. & Sun, Y. Heterogeneous graph transformer. *Proc. Web Conf.* **2020**, 2704–2710 (2020).
39. Ning, Q. *et al.* Amhmda: Attention aware multi-view similarity networks and hypergraph learning for mirna—disease associations identification. *Brief. Bioinform.* **24**, bbad094 (2023).
40. Li, W., Yin, Y., Quan, X. & Zhang, H. Gene expression value prediction based on xgboost algorithm. *Front. Genet.* **10**, 484931 (2019).
41. Liu, D., Huang, Y., Nie, W., Zhang, J. & Deng, L. Smalf: mirna-disease associations prediction based on stacked autoencoder and xgboost. *BMC Bioinform.* **22**, 219 (2021).

42. Branson, N., Cutillas, P. R. & Bessant, C. Comparison of multiple modalities for drug response prediction with learning curves using neural networks and xgboost. *Bioinform. Adv.* **4**, vbad190 (2024).
43. Deng, L., Liu, Z., Qian, Y. & Zhang, J. Predicting circrna-drug sensitivity associations via graph attention auto-encoder. *BMC Bioinform.* **23**, 160 (2022).
44. Lou, Z. *et al.* Predicting mirna—disease associations via learning multimodal networks and fusing mixed neighborhood information. *Brief. Bioinform.* **23**, bbac159 (2022).
45. Liang, X. *et al.* Predicting mirna—disease associations by combining graph and hypergraph convolutional network. *Interdiscip. Sci. Comput. Life Sci.* 1–15 (2024).
46. Wang, Y.-T., Wu, Q.-W., Gao, Z., Ni, J.-C. & Zheng, C.-H. Mirna-disease association prediction via hypergraph learning based on high-dimensionality features. *BMC Med. Inform. Decis. Mak.* **21**, 1–13 (2021).
47. Barshack, I. *et al.* Microrna expression differentiates between primary lung tumors and metastases to the lung. *Pathol. Res. Pract.* **206**, 578–584 (2010).
48. Yang, N., Ekanem, N. R., Sakyi, C. A. & Ray, S. D. Hepatocellular carcinoma and microrna: New perspectives on therapeutics and diagnostics. *Adv. Drug Deliv. Rev.* **81**, 62–74 (2015).
49. Ahmed, S. P., Castresana, J. S. & Shahi, M. H. Glioblastoma and mirnas. *Cancers* **13**, 1581 (2021).

Acknowledgements

This study was funded by the scientific research projects from Anhui Provincial Engineering Laboratory for Beidou Precision Agriculture Information: the research and implementation of a deep learning-enabled human-computer interaction mode for agricultural equipment (Grant number: BDSY2023002).

Author contributions

W.S.J. developed the experiments and wrote the manuscript, L.Y.B. conceived the study and processed data, Y.G. and C.W.X. collected data, W.H.T. polished the manuscript, Z.X.L. and W.Y.M. conducted data analysis. All authors reviewed the manuscript. These authors contributed equally: W.S.J. and L.Y.B. These authors jointly supervised this work: Z.X.L. and W.Y.M.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-68897-4.

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