BMC Genomic Data

RESEARCH Open Access

Establishing a GRU-GCN coordination-based prediction model for miRNA-disease associations

Kai-Cheng Chuang[1], Ping-Sung Cheng[2], Yu-Hung Tsai[2] and Meng-Hsiun Tsai[2,3*]

Abstract

Background miRNAs (microRNAs) are endogenous RNAs with lengths of 18 to 24 nucleotides and play critical roles in gene regulation and disease progression. Although traditional wet-lab experiments provide direct evidence for miRNA-disease associations, they are often time-consuming and complicated to analyze by current bioinformatics tools. In recent years, machine learning (ML) and deep learning (DL) techniques are powerful tools to analyze large-scale biological data. Hence, developing a model to predict, identify, and rank connections in miRNAs and diseases can significantly enhance the precision and efficiency in investigating the relationships between miRNAs and diseases.

Results In this study, we utilized miRNA-disease association data obtained by biotechnological experiments to develop a DL model for miRNA-disease associations. To improve the accuracy of prediction in this model, we introduced two labeling strategies, weight-based and majority-based definitions, to classify miRNA-disease associations. After preprocessing, data was trained with a novel model combining gated recurrent units (GRU) and graph convolutional network (GCN) to predict the level of miRNA-disease associations. The miRNA-disease association datasets were from HMDD (the Human miRNA Disease Database) and categorized by two distinct labeling approaches, weight-based definitions and majority-based definitions. We classified the miRNA-disease associations into three groups, "upregulated", "downregulated" and "nonspecific", by regression analysis and multiclass classification. This GRU-GCN coordinated model achieved a robust area under the curve (AUC) score of 0.8 in all datasets, demonstrating the efficacy in predicting potential miRNA-disease relationships.

Conclusions By introducing innovative label-preprocessing methods, this study addressed the relationships between miRNAs and diseases, and improved the ambiguity of the results in different experiments. Based on these refined label definitions, we developed a DL-based model to refine and predict the results of associations between miRNAs and diseases. This model offers a valuable tool for complementing traditional experimental methods and enhancing our understanding of miRNA-related disease mechanisms.

Keywords miRNAs, GRU (gated recurrent unit), GCN (graph convolutional network), miRNA-disease assosications

*Correspondence:
Meng-Hsiun Tsai
mht@nchu.edu.tw
[1]Department of Life Sciences, National Chung Hsing University, Taichung 402, Taiwan
[2]Department of Management Information Systems, National Chung Hsing University, Taichung 402, Taiwan
[3]Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung 402, Taiwan

Introduction

MicroRNAs (miRNAs) are endogenous noncoding RNAs with 18 to 24 nucleotides in lenhths, and play a major role in many important physiological processes [1]. miRNAs also play crucial roles in cancer- or disease-related pathways. The abnormalities in miRNAs may cause dysfunctions in target gene expression, which eventually leads to the proliferation of cancer cells [2]. Over 60% of human protein-coding genes selectively pair with miRNAs [3]. In general, miRNAs target complementary sequences on the 3'UTR of mRNAs to silence gene expression [4]. Pri-miRNAs (primary transcript miRNAs) are first transcribed and edited to pre-miRNAs in the nucleus, and then pre-miRNAs (precursor miRNAs) are further edited to mature miRNAs in the cytoplasm to identify specific sequences of mRNAs [5]. miRNAs are linked to various diseases such as miR-143/145 family in hypertension and cardiac failure, let-7 and miR-103/107 family in glucose metabolism, miR-29 in diabetes, and a cluster of oncomiR of miR-17 to 92 family in various cancers [6].

Today, various databases for miRNAs, such as miRbase, TargetScan, miRDB, MirGeneDB, and HMDD have been constructed [7–12]. Both miRbase and HMDD (the Human MicroRNA Disease Database) are continually updated. miRbase is a database containing both hairpin and mature miRNA sequences, annotations and functional information about specific microRNAs. miRbase contains pre-miRNA sequences with stem-loops, the positions and sequences of mature miRNAs, and related literatures of specific miRNAs [10, 13]. HMDD was established and used to provide cumulative knowledge on the associations with miRNAs and diseases from life science- or medical publications. HMDD is a database that collects and continually updates results of miRNA-disease associations [11, 12]. HMDD v3.0 is a database that manually collects 32,281 miRNA-disease associations including 1,102 miRNA genes, and 850 diseases from 17,412 publications [11]. Most studies on the relationships between miRNAs and diseases are based on experimental or wetlab assays which are expensive and time-consuming but highly accurate. However, due to experimental conditions, some controversial or inconsistent results exist in different publications. Developing a tool to complement traditional experimental methods will enhance our understanding of miRNA-related disease mechanisms.

Various models were used to delve into miRNA-disease associations. For examples, using semantic information and heterogeneous disease-related interaction data [14], metric learning [15], node2vec-based neural collaborative filtering [16], deep-belief network [17], and employing graph convolutional networks with a learning graph spatial operated paths for predicting miRNA-disease associations [18]. Employing regularized least squares to uncover the relationship between miRNAs and diseases [19]. Matrix factorization-based models such as similarity-based matrix factorization framework (SMAP) [20], framework of predicting miRNA disease associations via matrix factorization (MDMF) [21], and inferring miRNA-disease interactions using probabilistic matrix factorization (IMIPMF) [22] were used to predict the association between miRNAs and diseases.

Considering the labor-intensive nature of analyzing the results of associations between miRNAs and diseases from traditional biotechnological methods, we aimed to utilize the miRNA-disease data in HMDD v3.2 to develop a pairwise association model and predict potential miRNAs-diseases correlation based on machine learning (ML) or deep learning (DL) methods. miRNA-disease heterogeneous information networks were analyzed with graph neural network (GNN)-based approach to predict their relationships [23, 24]. DL is a branch of ML that uses the back-propagation algorithm to mine the structure of datasets and processes multiple layers to recognize data with multiple levels of abstraction [25]. Because of various tasks and data structures, many different algorithms, such as convolutional neural network (CNN), recurrent neural network (RNN), and graph convolutional network (GCN) have been developed for DL. CNN is widely used in processing image data such as images and videos, whereas RNN has made great progress in analyzing sequential data such as text and gene sequences. GCN is modified from CNN and properly handles tasks involving graph-type data, such as social networks, citation networks, and recommendation systems [26]. In this study, we proposed a GRU-GCN coordinated model to predict the association between miRNAs and diseases. We established two modules in this model, miRNA representation module and disease representation module, to evaluate the relationships between miRNAs and diseases. We used secondary structure and sequence of miRNAs as an input feature in the miRNA representation module. We converted MeSH descriptor to a directed acyclic graph combining with the GloVe word embedding transformation as a feature of node into disease vectors in the disease representation module. Furthermore, we utilized secondary structure and sequence of miRNAs in this model to provide another window to investigate the relationships between miRNAs and diseases in fields of bioinformatics, bio-simulation, and epigenetics. We expected this model could accelerate the identification of miRNA-disease association from the results of traditional experimental methods in this GRU-GCN coordination-based model.

Methods

Database

The parameters which are attributes in HMDD v3.2 include the code, miRNA name, disease name, PMID,

and description [11]. The "code" attribute in HMDD is assigned to a specific miRNA and disease pair. The number of each code was listed in Supplementary Table 1. The microRNA database (miRBase) is currently maintained by the Griffiths-Jones lab at the Faculty of Biology, Medicine, and Health, University of Manchester [13]. The sequences and corresponding names of the miRNAs were downloaded from miRbase.

Data preprocessing

HMDD label definition

HMDD database organized the data into six generalized categories (genetics, epigenetics, target, circulation, tissue, and others). Among the categories, "Genetics", "Circulation" and "Tissue" could be further divided into three classes: "upregulate", "nonspecific" and "downregulate". The detailed definitions of the classes were listed in Supplementary Table 2. The relationships between miRNAs and diseases in HMDD was accumulated from PubMed. However, the relationships between miRNAs and a diseases didn't represent consistant results in different experiments in different studies. Therefore, in addition to the digitization of labels, we considered the different results drawn by other publications. Assume that a specific miRNA and a specific disease in HMDD has p positively correlated results, n negatively correlated results, and ns irrelevant studies. For multiclass classification problems, we used majority-determination labels to address different conclusions, and deleted data where p and n are equal:

$$class = \begin{cases} max\left(p,\ n, ns\right),\ \ if\ p \neq\ n \neq\ ns \\ p,\ \ if\ p \neq\ ns > n \\ n,\ \ if\ n \neq\ ns > p \\ delete\ data,\ \ if\ p = n > ns \end{cases}$$

For the multiple regression problem, we defined weight labels by weights to address different results. The weight label is formulated as follows.

$$lable = \frac{p \times\ 1 + ns \times\ 0 + n \times\ (-1)}{p + n + ns}$$

miRNA data preprocessing

The raw data of the HMDD database contains the names of pre-miRNAs without mature miRNA gene sequences. Therefore, it is necessary to identify the mature miRNA sequences corresponding to pre-miRNAs in the HMDD. After obtaining the mature miRNA sequences from miRbase (Supplementary Fig. 2), the model needs the digital type data for input. We used K-merge algorithm to convert the mature miRNA sequence into a sequenced number. In this study, we set al.l the miRNAs to a fixed length of 28 with zero-padding. K-merge algorithm is illustrated in Supplementary Fig. 2. We also followed the methods of PDMDA [27] to extract pre-miRNAs' secondary structures as the input of the model. All extracted features and descriptions are shown in Supplementary Table 3.

Disease data preprocessing

We used the Medical Subject Headings (MeSH) descriptor for disease data preprocessing. MeSH descriptor was further transformed into a directed acyclic graph. An example of a directed acyclic graph of MeSH descriptor was illustrated in Supplementary Fig. 3. The diseases of MeSH descriptor was first converted into a directed acyclic graph, and then combined with the GloVe word embedding transformation [28] to convert the text in the node into disease vectors. Finally, the input graph of the GCN was constructed. More specifically, every disease had a different directed acyclic graph and was defined as $G = (V,\ E,\ X_V)$, where V was the set of MeSH descriptors, the edges $E = \{(i, j)|\ when\ v_i\ is\ adjecent\ to\ v_j\}$ represented the hierarchical relationship of MeSH descriptor, and nodes vector $X_E = \left\{X_{(i,j)}\middle|\ (i, j) \in\ E\right\}$ represented the disease vector of MeSH descriptor which was transformed by GloVe word embedding transformation.

Deep learning models

After the raw data has been labeled, preprocessed, and mapped to the original HMDD database, the processed data was input into the feature learning model. In this study, we employed GRU to learn the features of miRNA sequence and GCN to learn the features of disease graphs from MeSH descriptors. The multiple regression model in this study was based on MLP. The features of both miRNA and disease were simultaneously input to the final MLP-based multiple regression model. We used multilayer perceptron (MLP) [29], 1-dimensional convolutional neural network (Conv1D) [30], bidirectional encoder representations from transformers (BERT) [31], long short-term memory (LSTM) [32], GRU [33], and GCN [34] to train and test the miRNA feature learning module or disease feature learning module. Importantly, the preprocessing steps were various in the disease representation module because of the different models used. Only GRU used GloVe embedding as the data preprocessing method, and only BERT used tokenization as the data preprocessing method. After the raw data were labeled, preprocessed, and mapped to the original HMDD database, the processed data was input into the representation modules. We used MLP, Conv1D, BERT, LSTM, GRU, and GCN to learn the features of miRNA sequences and disease graphs from MeSH descriptors. We used the MLP classifier and MLP-based multiple regression

model to output the results. We randomly selected 640 data points, which were not included in the training set, as the testing set. All the testing data for each experiment was different. MLP, Conv1D, BERT, LSTM, and GRU were implemented using Pytorch, and GNN was implemented using Pytorch Geometric. Training hyperparameters and other details were shown in Supplementary Table 4. Both representation modules of miRNA and disease were then input to the final MLP-based multiple regression model. The features X_m^F learned from the miRNA representation module and the features X_d^F learned from the disease representation module were input into the MLP-based multiple regression model for regression analysis. The weight of the l-th layer was represented as $W^{(l)}|\,(l \in 1,2,3,\ldots,n)$, and the bias of the l-th layer wa represented as $b^{(l)}\,|(\,l \in 1,2,3,\ldots,n)$. The algorithm of regression model was as $\hat{y} = W^l ReLU\left(\ldots\ ReLU\left(W^1\left(X_m^F \oplus X_d^F\right) + b^1\right)\ldots\right) + b^l$, where $\oplus$ was matrix connection.

Predicted regulation-level model for miRNA-disease associations

miRNA sequence information $X_m = \{x_m^1, x_m^2, \ldots, x_m^n\}$ and disease vectors $X_d = \{x_d^1, x_d^2, \ldots, x_d^n\}$ were used as input features into a MLP-based multiple regression model (or a classifier) R. The MLP-based multiple regression model R performed regression analysis based on the input miRNA sequence information X_m and disease X_d and was trained to output the regulation-level $Y = \{y^1, y^2, \ldots, y^n\}$ which was extracted from HMDD. According to the notation above, the loss function could be defined as:

$$\arg\min_{\theta_R} \frac{1}{n}\sum_{i=1}^{n}\left(Y - R\left(x_m^i, x_d^i\right)\right)^2$$

where θ_R were the parameters of R. Assuming that the miRNA feature learning model F_m was used to learn the expression of X_m, the disease feature learning model F_d was used to learn the correct expression of X_d, and $\theta_{F_m}, \theta_{F_d}$ were the parameters of F_m and F_d, respectively. The loss function of the model could be thus formulated as follows.

$$\underset{\theta_R, \theta_{F_m}, \theta_{F_d}}{arg\,min} \frac{1}{n}\sum_{i=1}^{n}(Y - R(F_m(x_m^i), F_d(x_d^i)))^2$$

Model evaluation

We used mean square error (MSE) to evaluate the average squared difference between the estimated values and the actual values. The mean absolute error (MAE) was used to evaluate the sum of absolute errors. The confusion matrix was used to evaluate the values of Precision, Recall, and F1-score in different models. The ROC (receiver operating characteristic) and AUC (area under the curve) were used to evaluate the discrimination ability in different models. The definition of MSE and MAE was $\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$ and $\frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$. The formula of confusion matrix was $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$, and $F1\ Score = \frac{2\times Recall\times Precision}{Recall+Precision}$.

Results

Gate Recurrent Unit (GRU) had better performance in the miRNA representation module

Before constructing the GRU-GCN coordinated model to predict associations of miRNAs and diseases, we first estimated the performance of GRU-GCN coordinated model with different models in the miRNA representation module and MLP in the disease representation model. To evaluate the best performance of the miRNA representation module among the different DL models, we trained MLP, LSTM, Conv1D, GRU, and BERT as miRNA sequence feature extractors. Unexpectedly, during the data preprocessing stage, we found many repetitive, contradictory or mislabeled results in HMDD database. To correct biases, we used miRbase database to screen out mislabeled miRNAs to address these problems. After training, the combinations of features in the miRNA sequences and miRNA secondary structures were input to the MLP to obtain different miRNA representation modules. The features of diseases were input the MLP to obtain the representation of diseases. Then the modules of miRNAs and diseases were subsequently input into the regression model to predict regulation-level of miRNA-disease associations (Fig. 1). Table 1 showed that both GRU and BERT had better feature extracting power and achieved similar value in MSE (0.2319 and 0.2329), but GRU surpassed other models in terms of MAE (0.3451).

Graph Convolutional Network (GCN) had better performance in disease representation

In the Fig. 1; Table 1, we revealed the best performance of the miRNA representation module, we next evaluated the performance of GRU-GCN coordinated model with GRU-MLP-coordinated model in the miRNA representation module and different models in the disease representation model. To estimate the best performance of the disease representation modules among the different DL models, we trained MLP, LSTM, GRU, BERT, and GCN as disease feature extractors. After training, the GRU-MLP-coordinated miRNA representation module combined with different disease representation modules were input into the regression model for predicting regulation-level of miRNA-disease associations (Fig. 2).

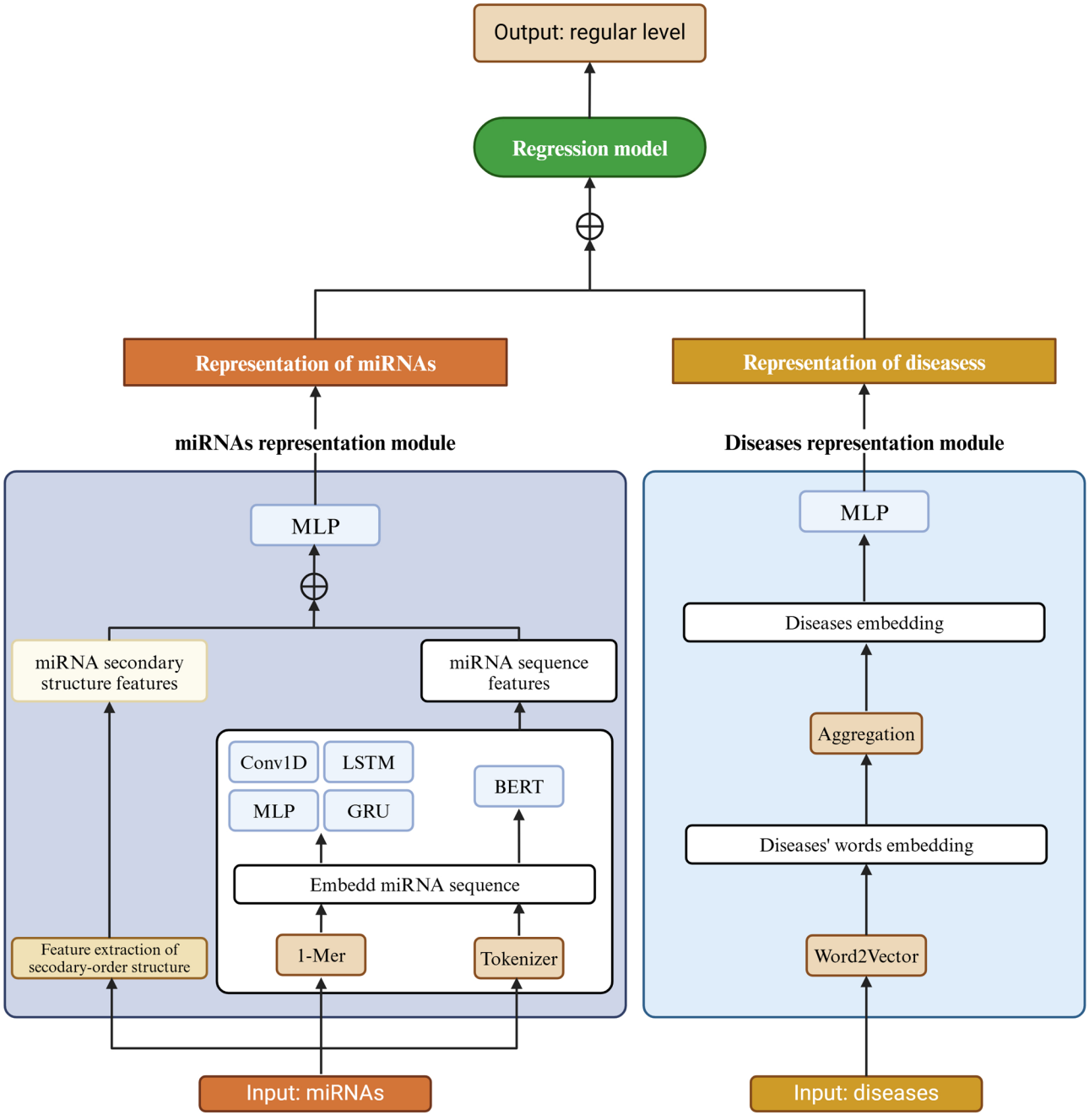

Fig. 1 The architecture of regular level by different miRNA sequence feature extractors. The figure was created with BioRender.com

Table 1 Performance of miRNA sequence representation modules

F_m	MLP	Conv1D	BERT	GRU	LSTM
F_d	MLP	MLP	MLP	MLP	MLP
MSE(L2)	0.2664	0.2904	0.2329	0.2319	0.2359
MAE(L1)	0.3908	0.4076	0.3649	0.3451	0.3610

MSE: mean square error. MAE: mean absolute error

Table 2 displayed that GCN had better feature extraction power and lower MSE (0.2406), but GRU surpassed other models in terms of MAE (0.3515).

1-merge had better performance in GRU-GCN coordination-based prediction model for miRNA-disease association

In the results of Tables 1 and 2, we demonstrated that GRU-GCN coordination-based prediction model had better performance for predicting the regular level of

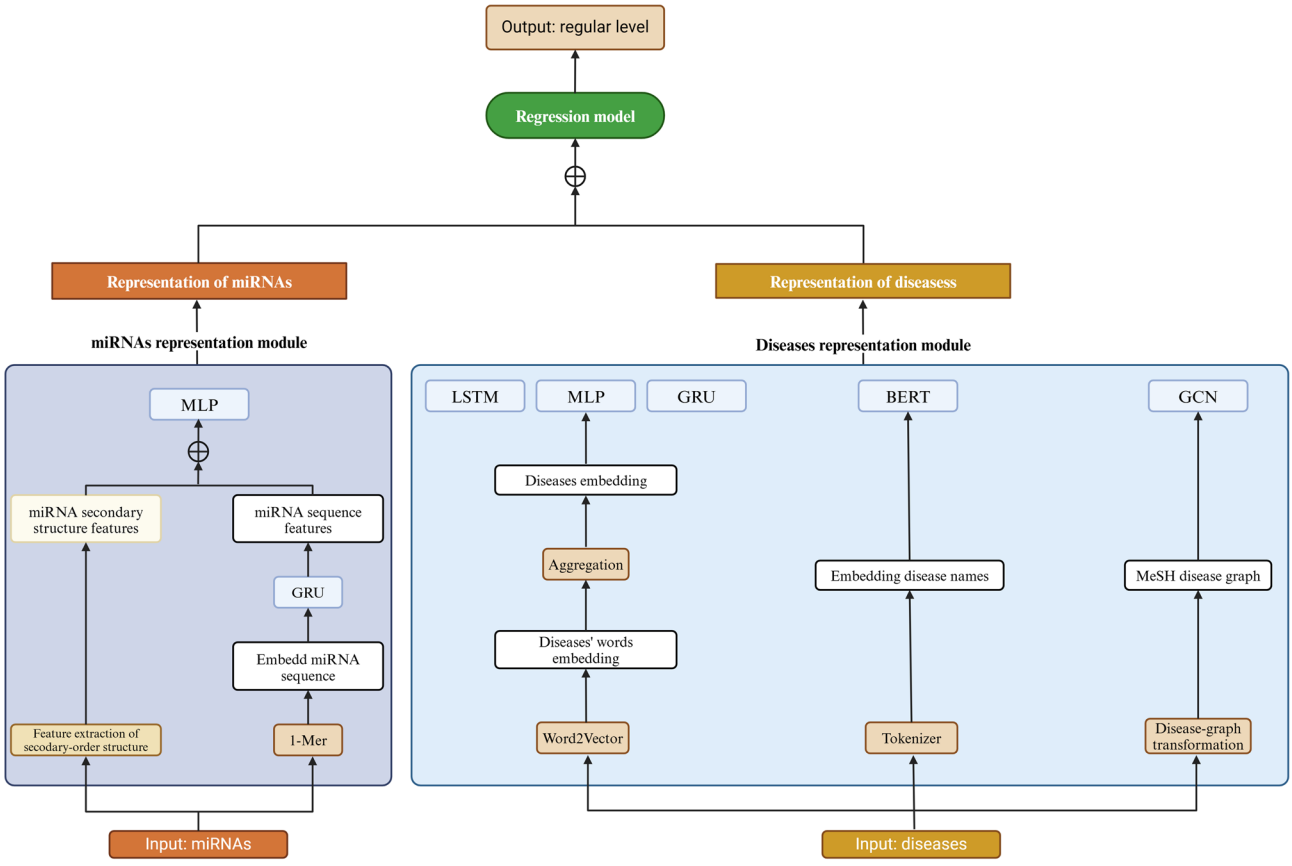

Fig. 2 The architecture of regular level by different disease feature extractors. The figure was created with BioRender.com

Table 2 Performance of disease representation disease representation modules

F_m	MLP	MLP	MLP	MLP	MLP
F_d	MLP	BERT	LSTM	GCN	GRU
MSE(L2)	0.2664	0.3670	0.3311	0.2406	0.2565
MAE(L1)	0.3908	0.4510	0.4187	0.3606	0.3515

MSE: mean square error. MAE: mean absolute error

miRNA-disease association. miRNAs are usually composed of 18–24 nucleotides and inputting different lengths of miRNAs into the miRNA presentation module would result in differences. For this reason, we used K-merge algorithm to assess the appropriate input length of a miRNA. We converted the mature miRNA sequence into a sequenced number and then input it into the GRU-MLP-coordinated miRNA representation for miRNA-disease association and examined the performance from 1-merge to 7-merge in this model (Fig. 3). Table 3 indicated that 1-merge coding had the lowest MSE (0.1901) and MSA (0.3386) for predicting the regular level of miRNA-disease association.

GRU-GCN coordination-based prediction model had the best performance the in predicting model for miRNA-disease association

The above results indicated that GRU and GCN had the best performance in the miRNA and disease representation modules respectively. Here, we replaced GRU and GCN with MLP in the miRNA and disease representation modules respectively to evaluate the performance in GRU-GCN, MLP-GCN, and GRU-MLP coordination-based prediction models. The representations of both miRNAs and diseases would all be input into the final multiple regression model. The model architecture was illustrated in Fig. 4. Table 4 showed that GRU-GCN coordination-based prediction model had the lower MSE (0.1901) and MAE (0.3386), because it had the best performance the in predicting model for miRNA-disease association.

Evaluation of GRU-GCN coordination-based prediction model for miRNA-disease association

In the results of Fig. 4; Table 4, we demonstrated that incorporating the GRU-MLP-based miRNA representation module and the GCN-based disease representation module into regression model had the best performance in miRNA-disease associations. However, it was

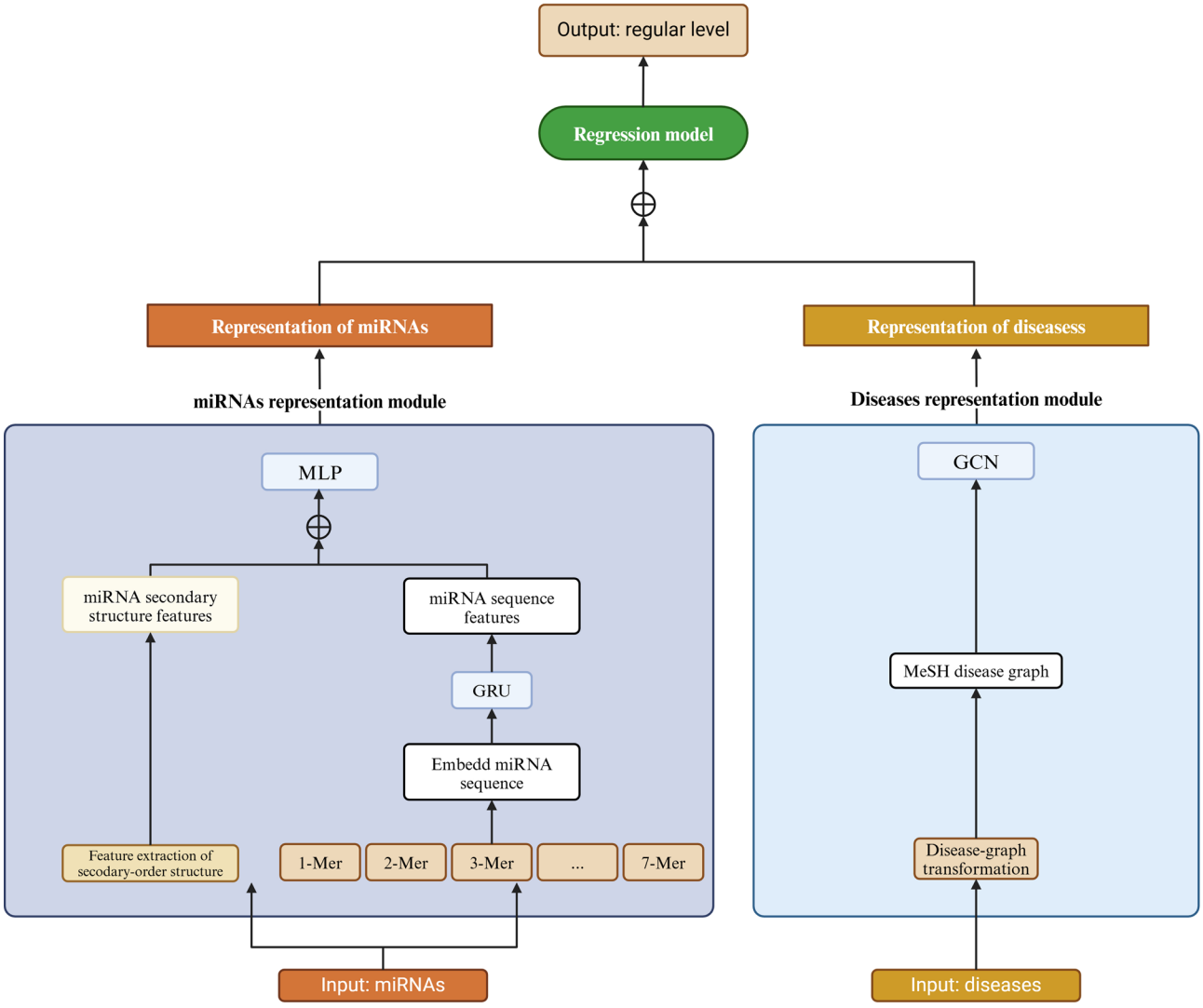

Fig. 3 The architecture of regular level by different K-merge coding in GRU-GCN coordination-based prediction model for miRNA-disease associations. The figure was created with BioRender.com

Table 3 Performance of different K-merge in GRU-GCN coordinated model

F_m	GRU	GRU	GRU	GRU	GRU	GRU	GRU
F_d	GCN	GCN	GCN	GCN	GCN	GCN	GCN
K-merge	1-mer	2-mer	3-mer	4-mer	5-mer	6-mer	7-mer
MSE(L2)	0.1901	0.2287	0.2279	0.2392	0. 2408	0.2892	0.2311
MAE(L1)	0.3386	0.3658	0.3588	0.3720	0.3860	0.4035	0.3713

MSE: mean square error. MAE: mean absolute error

still challenging to assess the effectiveness of this model through MSE and MAE. Here, we further defined the input dataset as a multiclass classification through majority-determination labels, and evaluated the value of Precision, Recall, F1-Score, ROC, and AUC in Table 5. The ROC curves of the GRU-GCN coordination-based prediction model for the 'Circulation,' 'Tissue,' and 'Genetics' datasets, as listed in Supplementary Table 2, are displayed in Fig. 5A and C, with the AUC for each exceeding 0.8. These results revealed that this model could effectively classify the associations between miRNAs and diseases. Notably, the AUC of the Circulation dataset had the best performance (above 0.9), and the AUC of the other datasets had lower performance (between 0.80 and 0.86). The ROC curve of the GRU-GCN coordination-based prediction model with a multiclass classifier on the "All" dataset was shown in Fig. 5D, and the AUC of the model was also greater than 0.8. This result indicated that the model had

Fig. 4 The architecture of regular level by GRU-GCN coordination-based prediction model for miRNA-disease associations. The figure was created with BioRender.com

Table 4 Performance of GRU-GCN, MLP-GCN and GRU-MLP coordinated model

F_m	GRU	MLP	GRU
F_d	GCN	GCN	MLP
MSE(L2)	0.1901	0.2406	0.2319
MAE(L1)	0.3386	0.3606	0.3451

MSE: mean square error. MAE: mean absolute error

Table 5 Evaluation of GRU-GCN coordination-based prediction model for miRNA-disease association

Dataset	Class	AUC	Precision	Recall	F1-Score
Circulation	Down	0.9331	0.7959	0.6964	0.7429
	NS	0.9103	0.8413	0.8689	0.8548
	Up	0.9353	0.7805	0.7901	0.7853
Tissue	Down	0.8144	0.6289	0.6455	0.6371
	NS	0.8341	0.6650	0.6782	0.6716
	Up	0.8021	0.7042	0.6787	0.6912
Genetics	Down	0.8648	0.8819	0.8971	0.8894
	Up	0.8653	0.8133	0.7888	0.8009
All	Down	0.8071	0.4070	0.6250	0.4930
	NS	0.8138	0.8595	0.5683	0.6842
	Up	0.8334	0.5221	0.7284	0.6582

Down: downregulated, NS: nonspecific, Up: upregulated

good identification ability in the face of the integrated data.

Discussion

In this study, we established a model for predicting the level of miRNA and disease associations. After comparing the performance of different model pairs, we demonstrated that the GRU-GCN coordination-based prediction model was the best model pair in Figs. 1 and 2; Tables 1 and 2. miRNA is a kind of sequential data

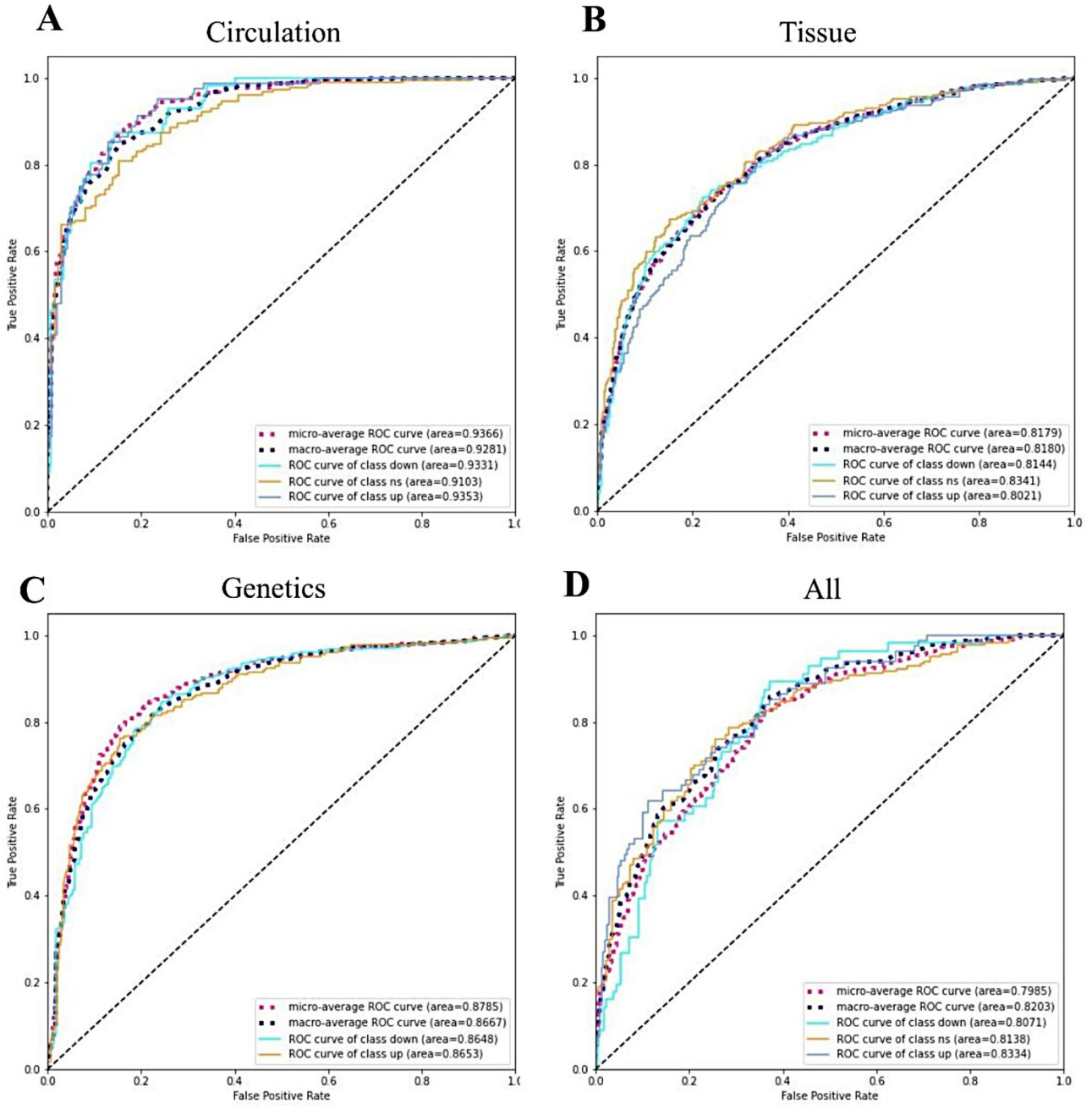

Fig. 5 ROCs of GCN & GRU regulation prediction using multi-class classifier using (**A**) Circulation", (**B**) Tissue, (**C**) Genetics and (**D**) All datasets

composed of "A", "U", "C", and "G" and converting these this sequential data to digital types in the preprocessing step is important. In the results of Fig. 3; Table 3, we showed that 1-merge coding had the lowest MSE and MAE for predicting the level of miRNA-disease association. In the results of Figs. 4 and 5; Tables 4 and 5, we demonstrated that the GRU-GCN coordination-based prediction model had optimal identification ability in multiclass classification.

In the miRNA representation module, both GRU and BERT had good feature extraction power and achieved similar MSE results. However, GRU had the lowest MSE and MAE compared with other models. Although BERT is the most common model for sequence data, the performance of BERT was not comparable to that of GRU. A possible reason could be insufficient pre-training data for BERT. In the disease representation module, we compared MLP, LSTM, GRU, BERT, and GCN as disease feature extractors and determined that GCN had the lowest MSE and lower MAE. This is likely because the inputting data by MeSH disease graph in GCN contained more information compared to other preprocessing methods.

More specifically, the disease graph provided by MeSH includes all related disease descriptors into the child nodes of the disease. Consequently, GCN had higher sensitivity to the relationships between diseases during training, better understanding for the classification and relationships of diseases and improved the feature expression of the diseases.

HMDD congregates large amounts of miRNA-related literature. However, the labels in the HMDD appear as repetitive or contradictory collections. These factors would cause the problem as a binary classification to distinguish whether diseases were downregulation, upregulation or not by miRNAs in ML. To overcome the inadequacies in predicting potential associations between miRNAs and diseases in ML methods, we proposed a method of deep regulation-level for miRNA-disease associations model with GRU and GNN to define the data preprocessing method and to solve the problem of label duplication in the HMDD database.

The predicting results of ML are highly dependent on the representations of the data. Directly inputting raw data without feature extraction might cause the regression model to analyze the irrelevant features. Extracting the features of X_m and X_d were the main factors for the quality of prediction in the model. Thus, we used DL models to learn the proper representation of X_m and X_d. These representation learning modules could convert X_m and X_d into features which were more suitable for inputting into regression model and decreased the labor consumption of manual feature extraction.

Zheng et al. using K-mer sparse matrix to extract miRNA sequence information, miRNA functional similarity, disease semantic similarity and Gaussian interaction profile kernel similarity information to predict miRNA-disease associations [35]. Differently, we linked the features of secondary structure and sequence of miRNAs in the miRNA representation module and the features of MeSH descriptor to node into disease vectors in the disease representation module to predict miRNA-disease associations. The secondary structure of miRNA was responsible for the inhibitory ability of miRNA. In the miRNA representation module, we extracted features of the secondary structure and sequence in miRNAs. The Mesh descriptors could improve the connection between clinically defined diseases and miRNAs. In the diseases representation module, we used MeSH descriptors to extract features of diseases. By combining these two modules, we could predict the associations of miRNAs and clinically defined diseases through considering the influence of the structure and sequence of miRNA. The predictive results in this model were based on inputting datasets by supervised learning. Whether unknow associations between miRNAs and diseases could be identified requires further investigation. The innovation of this study was to establish a prediction model for the associations of miRNAs and diseases, which had potential benefits for clinical applications.

Conclusion

In this GRU-GCN-coordinated model, we combined two innovative label-preprocessing methods to define the relationships between miRNAs and diseases and improve the ambiguity of the results from different experiments. On the dasis of these definitions, we proposed a deep learning-based model to refine and predict the results of associations between miRNAs and diseases. Through connecting the features of the secondary structure and sequence of miRNAs in the miRNA representation module and the Mesh descriptors in the disease representation module in the GRU-GCN coordination-based model to predict the relationshps between miRNAs and diseasesm this model showed good identification ability in miRNA-disease association. We hope this model could effectively help understand the potential associations between miRNAs and diseases, reduce the redundant analyzing processes and assist biological researchers to select the trustworthy miRNA-disease pairs. In addition, identifying the associations between miRNAs and diseases could help researchers further understand the relationships between pathogenesis and miRNAs in diseases, and therefore provide significant contributions to medical applications such as disease treatment, diagnosis, prevention, and drug development.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-024-01293-z.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

Not applicable.

Author contributions

Kai-Cheng Chuang: conceptualization, validation, investigation, project administration, and writing—original draft preparation. Ping-Sung Cheng: data curation and formal analysis. Yu-Hung Tsai: methodology and formal analysis. Meng-Hsiun Tsai: conceptualization, project administration, resources, writing—review and editing, supervision, and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the grant from the National Science and Technology Council in Taiwan (NSTC110-2511-H-006-013-MY3 and NSTC111-2622-H-006-004), and the Taichung Veterans General Hospital Research Program (TCVGH-NCHU1127602 and TCVGH-NCHU1137630).

Data availability

The datasets generated and/or analyzed during the current study are open and available in the HMDD v3.2 and miRBase repository, [HMDD v3.2: HMDD v4.0 (cuilab.cn); miRbase: miRBase - Downloads]. The code used in this study have been deposited at GitHub (https://github.com/luke-cai/GRU-GCN-for-miRNA-disease)

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 5 September 2024 / Accepted: 19 December 2024
Published online: 14 January 2025

References

1. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116(2):281–97.
2. Xiao Y, Guan J, Ping Y, Xu C, Huang T, Zhao H, Fan H, Li Y, Lv Y, Zhao T, et al. Prioritizing cancer-related key miRNA-target interactions by integrative genomics. Nucleic Acids Res. 2012;40(16):7653–65.
3. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 2009;19(1):92–105.
4. O'Brien J, Hayder H, Zayed Y, Peng C. Overview of MicroRNA Biogenesis, mechanisms of actions, and circulation. Front Endocrinol (Lausanne). 2018;9:402.
5. Cai Y, Yu X, Hu S, Yu J. A brief review on the mechanisms of miRNA regulation. Genomics Proteom Bioinf. 2009;7(4):147–54.
6. Saliminejad K, Khorram Khorshid HR, Soleymani Fard S, Ghaffari SH. An overview of microRNAs: Biology, functions, therapeutics, and analysis methods. J Cell Physiol. 2019;234(5):5451–65.
7. Singh NK. miRNAs target databases: developmental methods and target identification techniques with functional annotations. Cell Mol Life Sci. 2017;74(12):2239–61.
8. Fromm B, Domanska D, Hoye E, Ovchinnikov V, Kang W, Aparicio-Puerta E, Johansen M, Flatmark K, Mathelier A, Hovig E, et al. MirGeneDB 2.0: the metazoan microRNA complement. Nucleic Acids Res. 2020;48(D1):D132–41.
9. Wong N, Wang X. miRDB: an online resource for microRNA target prediction and functional annotations. Nucleic Acids Res. 2015;43(Database issue):D146–152.
10. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47(D1):D155–62.
11. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q. HMDD v3.0: a database for experimentally supported human microRNA-disease associations. Nucleic Acids Res. 2019;47(D1):D1013–7.
12. Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, Cui Q. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Res. 2014;42(Database issue):D1070–1074.
13. Griffiths-Jones S. miRBase: the microRNA sequence database. Methods Mol Biol. 2006;342:129–38.
14. Fu L, Peng Q. A deep ensemble model to predict miRNA-disease association. Sci Rep. 2017;7(1):14482.
15. Ha J, Park C. MLMD: Metric Learning for Predicting MiRNA-Disease associations. Ieee Access. 2021;9:78847–58.
16. Ha J, Park S. NCMD: Node2vec-Based neural collaborative filtering for Predicting MiRNA-Disease Association. Ieee Acm T Comput Bi. 2023;20(2):1257–68.
17. Chen X, Li TH, Zhao Y, Wang CC, Zhu CC. Deep-belief network for predicting potential miRNA-disease associations. Brief Bioinform. 2021; 22(3).
18. Chu S, Duan G, Yan C. PGCNMDA: learning node representations along paths with graph convolutional network for predicting miRNA-disease associations. Methods. 2024;229:71–81.
19. Chen X, Yan GY. Semi-supervised learning for potential human microRNA-disease associations inference. Sci Rep. 2014;4:5501.
20. Ha J. SMAP: similarity-based matrix factorization framework for inferring miRNA-disease association. Knowl-Based Syst. 2023; 263.
21. Ha J. MDMF: Predicting miRNA-Disease Association based on Matrix Factorization with Disease Similarity Constraint. J Pers Med. 2022; 12(6).
22. Ha J, Park C, Park C, Park S. IMIPMF: Inferring miRNA-disease interactions using probabilistic matrix factorization. J Biomed Inf 2020, 102.
23. Chen M, Peng Y, Li A, Li Z, Deng Y, Liu W, Liao B, Dai C. A novel information diffusion method based on network consistency for identifying disease related microRNAs. RSC Adv. 2018;8(64):36675–90.
24. Li J, Zhang S, Liu T, Ning C, Zhang Z, Zhou W. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. Bioinformatics. 2020;36(8):2538–46.
25. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
26. Khemani B, Patil S, Kotecha K, Tanwar S. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. J Big Data-Ger. 2024; 11(1).
27. Yan C, Duan G, Li N, Zhang L, Wu FX, Wang J. PDMDA: predicting deep-level miRNA-disease associations with graph neural networks and sequence features. Bioinformatics. 2022;38(8):2226–34.
28. Jeffrey Pennington RS. Christopher Manning: GloVe: Global Vectors for Word Representation. In: *Proceedings of the* 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Doha, Qatar: Edited by Alessandro Moschitti BP, Walter Daelemans. Association for Computational Linguistics 2014: 1532–1543.
29. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533–6.
30. Serkan Kiranyaz OA, Osama Abdeljaber T, Ince M, Gabbouj DJ. Inman: 1D convolutional neural networks and applications: a survey. Mech Syst Signal Process. 2021; 151.
31. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics. 2021;37(15):2112–20.
32. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
33. Hill ST, Kuintzle R, Teegarden A, Merrill E 3rd, Danaee P, Hendrix DA. A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. Nucleic Acids Res. 2018;46(16):8105–13.
34. Wang T, Bai J, Nabavi S. Single-cell classification using graph convolutional networks. BMC Bioinformatics. 2021;22(1):364.
35. Zheng K, You ZH, Wang L, Zhou Y, Li LP, Li ZW. MLMDA: a machine learning approach to predict and validate MicroRNA-disease associations by integrating of heterogenous information sources. J Transl Med. 2019;17(1):260.

Publisher's note