BMC Bioinformatics

RESEARCH Open Access

MAGCNSE: predicting lncRNA-disease associations using multi-view attention graph convolutional network and stacking ensemble model

Ying Liang[1†], Ze-Qun Zhang[1†], Nian-Nian Liu[1], Ya-Nan Wu[1], Chang-Long Gu[2] and Ying-Long Wang[1*]

†Ying Liang and Ze-Qun Zhang have contributed equally to this work and should be considered co-first authors

*Correspondence: wangyl@jxau.edu.cn

[1] College of Computer and Information Engineering, Jiangxi Agricultural University, Nanchang, China
Full list of author information is available at the end of the article

Abstract

Background: Many long non-coding RNAs (lncRNAs) have key roles in different human biologic processes and are closely linked to numerous human diseases, according to cumulative evidence. Predicting potential lncRNA-disease associations can help to detect disease biomarkers and perform disease analysis and prevention. Establishing effective computational methods for lncRNA-disease association prediction is critical.

Results: In this paper, we propose a novel model named MAGCNSE to predict underlying lncRNA-disease associations. We first obtain multiple feature matrices from the multi-view similarity graphs of lncRNAs and diseases utilizing graph convolutional network. Then, the weights are adaptively assigned to different feature matrices of lncRNAs and diseases using the attention mechanism. Next, the final representations of lncRNAs and diseases is acquired by further extracting features from the multi-channel feature matrices of lncRNAs and diseases using convolutional neural network. Finally, we employ a stacking ensemble classifier, consisting of multiple traditional machine learning classifiers, to make the final prediction. The results of ablation studies in both representation learning methods and classification methods demonstrate the validity of each module. Furthermore, we compare the overall performance of MAGCNSE with that of six other state-of-the-art models, the results show that it outperforms the other methods. Moreover, we verify the effectiveness of using multi-view data of lncRNAs and diseases. Case studies further reveal the outstanding ability of MAGCNSE in the identification of potential lncRNA-disease associations.

Conclusions: The experimental results indicate that MAGCNSE is a useful approach for predicting potential lncRNA-disease associations.

Keywords: LncRNA-disease associations, Multi-view, Graph convolutional network, Attention mechanism, Convolutional neural network, Stacking ensemble model

Background

Long non-coding RNAs (lncRNAs) are a type of non-coding RNA with the length of more than 200 nucleotides, which cannot encode proteins [1]. The lncRNAs play important roles in many human biologic processes, such as oncogenesis, gene regulation, protein translation, expression, tissue development and immune regulation [2]. In recent years, cumulative research has proved many lncRNAs to be associated with various diseases, including lung cancer [3, 4], breast cancer [5, 6],prostate cancer [7, 8], gastric cancer [9, 10],colon cancer [11, 12], Alzheimer's disease [13, 14] and others.

Predicting underlying association between lncRNAs and different diseases has extremely important significance and value, since it can help to analyze and prevent diseases, identify disease biomarkers and reveal the mechanism of lncRNA levels in diseases. However, many biological experiments suffer from the long time and high cost. As a result, a growing number of computational methods have been recently developed to identify lncRNA-disease associations (LDAs). These methods can roughly be classified into two categories: biological network-based methods and machine learning (ML)-based methods.

Biological network-based methods are premised on the notion that functionally comparable lncRNAs are frequently linked to the similar diseases. In these methods, heterogeneous networks of diseases and lncRNAs are constructed, then LDAs are identified via different methods, such as matrix decomposition or random walk, etc. For example, SIMCLDA [15] first used principal component analysis (PCA) to select features from similarity matrices, then predicted LDAs via inductive matrix completion. BiWalkLDA [16] fused the data from gene ontology and interaction profiles, then utilized the bi-random walks algorithm for prediction. WMFLDA [17] firstly assigned weights to the gene, lncRNA and disease association matrices, then decomposed the rank of these matrices and employed the optimized matrices and weights for prediction. DMFLDA [18] was a deep matrix factorization model which obtained the latent representations through non-linear hidden layers, then used a fully connected layer to connect the representations and finally generates the predictions. MHRWR [19] firstly constructed a heterogeneous network in accordance with six network relevant to lncRNA, gene and disease, then predict LDAs by utilizing a random walk with restart. However, the above-mentioned models based on matrix decomposition or random walk face difficulty in mining the topological information from nodes in the lncRNA-disease network.

ML-based methods generally use feature extraction techniques on lncRNAs and diseases to generate their representations, then identify potential LDAs by applying ML classifiers. ML-based methods here do not only refer to the traditional ML methods, but also to deep learning methods. For example, LDAP [20] used the Karcher mean of the matrices to integrate different biological data and utilized bagging support vector machine to predict LDAs. LDAPred [21] predicted LDAs through a dual convolutional neural network (CNN) and information flow propagation. iLncRNA-dis-FB [22] used the lncRNA-disease similarity matrix to generate three-dimensional feature blocks and fed them into CNN for prediction. RFLDA [23] extracted features using the random forest (RF) variable importance score and then used a RF regression model for prediction. SDLDA [24] first utilized a neural network with singular value decomposition to separately obtain the disease and lncRNA representations, then calculated Hadamard product of them and predict LDAs using a sigmoid activation function.

Although these methods for identifying LDAs have yielded promising results, there is still space for improvement. Firstly, for the representation learning methods, more advanced deep learning methods could be considered, such as the technique of graph convolutional networks (GCNs) for feature extraction, which has recently achieved outstanding performance. For example, GAMCLDA [25] used GCN to get the representations of diseases and lncRNAs, and the inner product of them was computed to reconstruct lncRNA-disease associations. GAERF [26] first created a heterogeneous network by fusing the interaction of lncRNA, miRNA and disease, then a graph autoencoder was leveraged to acquire low-dimensional features, finally used a RF classifier for LDA prediction. PANDA [27] applied a graph autoencoder for feature extraction and utilized a neural network to predict LDAs. In addition, some models in the field of LDA prediction use single lncRNA data and disease data, and many models do not consider the lncRNA sequence information. The fusion of multisource data has recently been extensively embraced in many studies [28–30]. Moreover, the studies of LDAs that involve the integration of multi-view data of lncRNAs and diseases do not consider the contribution weight of different data. Furthermore, for the final classification methods, many studies only use an individual traditional ML classifier, which has its strengths as well as weaknesses.

In this study, a novel method named MAGCNSE is proposed to predict LDAs. First, the GCN is used to extract features from the similarity graphs of different views of lncRNAs and diseases to obtain multiple feature matrices. For views of diseases, MAGCNSE uses disease semantic similarity (DSS) and disease Gaussian interaction profile kernel similarity (DGS), and for views of lncRNAs, MAGCNSE uses lncRNA functional similarity (LFS), lncRNA sequence similarity (LSS) and lncRNA Gaussian interaction profile kernel similarity (LGS). Then, MAGCNSE leverages attention mechanism for adaptively assigning weights to different feature matrices of lncRNAs and diseases. Next, MAGCNSE uses the CNN to further extract features from multi-channel feature matrices to acquire the final representations of lncRNAs and diseases. The representation learning processes were partially inspired by the study [31]. MAGCNSE then concatenates the representations of lncRNAs and diseases according to the lncRNA-disease association matrix to form the positive and negative lncRNA-disease pairs. Finally, a stacking ensemble classifier, which consists of multiple traditional classifiers, is leveraged to identify LDAs. To demonstrate the effectiveness of MAGCNSE, we firstly perform ablation studies in both representation learning methods and classification methods to demonstrate the validity of each module of our model, and we compare GCN with two graph neural network models to illustrate the validity of GCN in this study. In addition, we compare MAGCNSE with six state-of-the-art models on the same datasets of lncRNAs and diseases using 5-fold cross-validation (5-CV) to observe the overall performance of the entire model. Furthermore, we test the performance of MAGCNSE using multi-view data of lncRNAs and diseases. Finally, we implement two types of case studies to validate the performance of MAGCNSE in predicting LDAs for specific diseases. All the results indicate the great capacity of MAGCNSE in identifying LDAs. Compared with previous models in the field of LDA prediction, the main innovations and contributions of this study are summarized as follows:

(1) Multi-view data of lncRNAs and diseases were used in this study and MAGCNSE incorporated the lncRNA sequence information.
(2) MAGCNSE used deep learning methods that synthesize the techniques of GCN, attention mechanism and CNN to fuse the multi-view data to learn the low-dimensional representations of lncRNAs and diseases.
(3) After getting the positive and negative lncRNA-disease pairs by concatenating the representations of lncRNAs and diseases according to the lncRNA-disease association matrix, MAGCNSE applied a stacking ensemble model that integrates multiple machine learning classifiers for the prediction task.
(4) A series of experiments were performed to demonstrate that MAGCNSE is competitive and reliable in the field of LDA prediction.

Results and discussion

Experimental settings

To evaluate the performance of our model, we used 5-CV for prediction comparison. We treated the known 1569 LDAs as positive samples. To eliminate the impact of data imbalance between positive samples and negative samples, many previous studies [32–36] randomly selected the same number of negative samples from the unknown LDAs. We followed the same strategy and randomly selected 1569 LDAs from all the unknown LDAs to be the negative samples. For 5-CV, the dataset was divided into 5 disjoint subsets, among which 4 subsets were utilized to train the model and the remaining subset was utilized for testing in each round. We used all three views of lncRNAs and two views of diseases in this study. To learn the representations, we applied the Adam optimizer and set the learning rate to 0.001, and we trained MAGCNSE for 250 epochs. Other important hyperparameters will be discussed in subsequent sections.

Area under the receiver-operating characteristic (ROC) curve (AUC) and area under the precision-recall (PR) curve (AUPR) were utilized as two comprehensive performance evaluation metrics for performance evaluation of MAGCNSE. Other six evaluation metrics are also used, including Accuracy, Sensitivity, Specificity, Precision, $F1$- *score* and Matthews correlation coefficient (MCC). These metrics are calculated as follows:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$F1\text{-}\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \tag{6}$$

where TP, FN, TN, FP denote the number of true positives, false negatives, true negatives and false positives, respectively.

To reduce the bias caused by random sample splitting, we implemented 5 times 5-CV and used the average values of the evaluation metrics.

Effect of parameters

Since the selection of hyperparameters affects the final prediction results, it's necessary to find the relatively optimal hyperparameters, including the GCN embedding size, number of filters in CNN, number of GCN layers and number of base classifiers. The embedding size of lncRNAs and diseases in GCN could affect their final representations to a large extent, the dimension of the ultimate representations of lncRNAs and diseases was decided by the number of CNN filters, the number of GCN layers affects the number of feature matrices extracted by GCN, the number of base classifiers in the stacking ensemble model determines the input dimension of the LogisticRegression classifier. GCN embedding size was chosen from {16,32,64,128,256}, number of filters in CNN was chosen from {16,32,64,128,256}, number of GCN layers was chosen from {1,2,3,4,5}, number of base classifiers was chosen from {5,10,15,20,25}. We compared the performance of MAGCNSE using different values of hyperparameters under 5-CV, such that only one of the hyperparameters was changed each time, the results are shown in Fig 1. When the AUC value reached the maximum, we selected corresponding value of hyperparameters. In this paper, we set the GCN embedding sizes, numbers of filters in CNN and numbers of GCN layers to 128, 128, 2, respectively. Specifically, the AUC value was slightly influenced by the number of base classifiers. Aiming to reduce complexity and the running time of MAGCNSE, we used 5 base classifiers in this paper.

Ablation studies

For the representation learning, to validate the necessity of using multiple GCN layers and adding the attention mechanism and CNN, we used 5-CV to compare MAGCNSE with the following four variants. (1) MAGCNSE-fgl: uses only the feature matrices generated by the first GCN layer and ignores the subsequent GCN layers, while the attention mechanism and CNN are still applied. (2) MAGCNSE-natt: uses multiple GCN layers and applies CNN to fuse them but does not use the attention mechanism; different feature matrices of lncRNAs and diseases extracted from GCN are given the same weights. (3) MAGCNSE-nattcnn: removes both the attention mechanism and CNN and only uses multiple GCN layers, then assigns the same weights to them. (4) MAGCNSE-ncnn: the

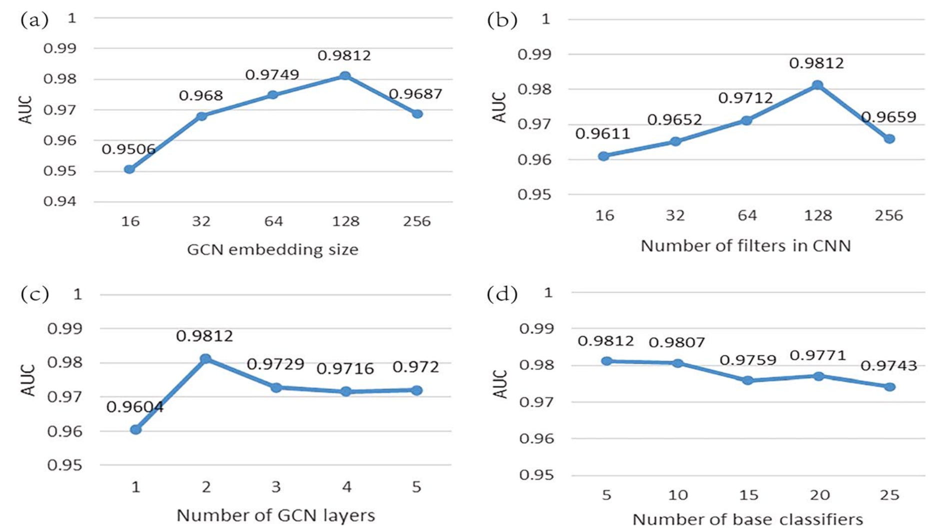

Fig. 1 Performance of MAGCNSE using different parameters. (a) Comparison of the AUC values under different GCN embedding sizes. (b) Comparison of the AUC values under different number of filters in CNN. (c) Comparison of the AUC values under different number of GCN layers. (d) Comparison of the AUC values under different number of base classifiers

feature matrix generated by multiple GCN layers is still applied, and attention mechanism is also applied, but CNN is not used for fusion.

It can be seen from Fig 2 and Table 1 that MAGCNSE achieved a superior prediction performance to its variants on all evaluation metrics. Compared with MAGCNSE-fgl, MAGCNSE uses multiple GCN layers rather than one GCN layer, so it gets more

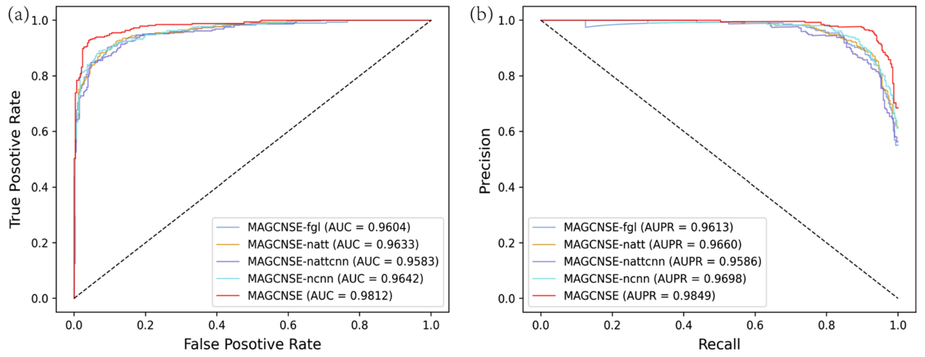

Fig. 2 ROC curves (a) and PR curves (b) of MAGCNSE and its variants

Table 1 Comparison of the evaluation metrics between MAGCNSE and its four variants

Method	Accuracy	Sensitivity	Specificity	Precision	*F*1- *score*	MCC
MAGCNSE-fgl	0.9029	0.9013	0.9043	0.8984	0.8998	0.8056
MAGCNSE-natt	0.9013	0.9068	0.8959	0.8952	0.901	0.8026
MAGCNSE-nattcnn	0.8885	0.9003	0.8783	0.8647	0.8822	0.7771
MAGCNSE-ncnn	0.9013	0.896	0.907	0.9128	0.9043	0.8025
MAGCNSE	**0.9395**	**0.9192**	**0.9626**	**0.9654**	**0.9417**	**0.88**

The bold number is the highest value of each column and its clarifies the superiority of our model

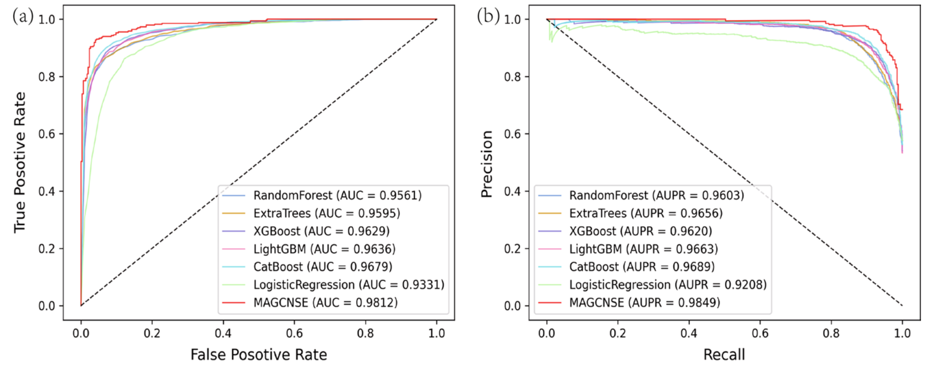

Fig. 3 ROC curves (a) and PR curves (b) of MAGCNSE and traditional ML classifiers

Table 2 Comparison of the evaluation metrics between MAGCNSE and six traditional machine learning classifiers

Method	Accuracy	Sensitivity	Specificity	Precision	*F*1- *score*	MCC
RandonForest	0.8945	0.877	0.9120	0.9089	0.8926	0.7896
ExtraTrees	0.8958	0.8859	0.9057	0.9042	0.8948	0.7921
XGBoost	0.9076	0.9101	0.9050	0.9056	0.9078	0.8153
LightGBM	0.9037	0.9031	0.9044	0.9052	0.9036	0.8085
CatBoost	0.9108	0.9146	0.9070	0.9079	0.9111	0.8218
LogisticRegression	0.8652	0.8470	0.8834	0.8792	0.8627	0.7312
MAGCNSE	**0.9395**	**0.9192**	**0.9626**	**0.9654**	**0.9417**	**0.88**

The bold number is the highest value of each column and its clarifies the superiority of our model

feature matrices. The results support the conclusion that different information may lie in the neighbors with different distances in the similarity network, and the performance may thus be enhanced by integrating their information. Compared with MAGCNSE-natt, MAGCNSE assigns weights to different feature matrices of lncRNAs and diseases through the attention mechanism. The results indicate the importance of using different feature matrices extracted from GCN, which is different when different views are applied, and the performance can be improved by importing the attention mechanism. Compared with MAGCNSE-ncnn, MAGCNSE uses CNN to fuse data and further extract the representations, the results show the effectiveness of CNN in processing multi-channel feature matrices. Compared with MAGCNSE-nattcnn, MAGCNSE does not only use the attention mechanism, but also employs the CNN. It can be noted that MAGCNSE-natt and MAGCNSE-ncnn outperform MAGCNSE-nattcnn, which further shows the effectiveness of both the attention mechanism and CNN in this study.

For the classification task, we compared the entire stacking ensemble model with single base classifiers and the LogisticRegression classifier under 5-CV.

From Fig 3 and Table 2, we can learn that the stacking ensemble model outperforms the six single classifiers on all evaluation metrics. It proves that the stacking ensemble model can achieve more robust performance than single traditional ML classifiers. The reason for the improvement in the MAGCNSE performance lies in the ability of the stacking ensemble model to average out noise from different single models and thus enhance the generalizable signal. Each individual classifier may have its weaknesses and biases on the datasets, but they can be countered with the strengths of other classifiers in the stacking ensemble model [37].

Comparison of GCN and other graph neural network models

Many graph neural network (GNN) models have been recently applied in the field of bioinformatics. Hence, we selected two advanced GNN models, graph attention network (GAT) [38] and graph sample and aggregate (GraphSAGE) [39] to compare with GCN. The difference between GCN and GAT lies in that GCN explicitly assigns non-parametric weights to the neighbor nodes, while GAT implicitly captures the different weights to neighbor nodes via the neural network architecture during the aggregation process. GraphSAGE proposes a batch-training algorithm and adopts sampling to obtain a fixed number of neighbors for each node, while training GCN usually requires using the whole graph data [40]. We used these three GNN models to extract features from the similarity graphs of different views of lncRNAs and diseases, and kept the subsequent modules of MAGCNSE unchanged for a fair comparison. Table 3 illustrated that GCN performs better than GAT and GraphSAGE for our task, which verifies the effectiveness of GCN for feature extraction in this study.

Comparison with other state-of-the-art methods

To evaluate the overall performance of MAGCNSE, we compared it with six recently proposed state-of-the-art models: LDNFSGB [33], IPCARF [36], VGAELDA [41], RSWF-BLP [42], LDASR [32], GCRFLDA [43]. To be fair, we evaluated all the above-mentioned methods utilizing 5-CV on the same datasets of lncRNAs and diseases, and we used AUC and AUPR value as the evaluation metrics. As shown in Fig 4, the LDNFSGB, IPCARF, VGAELDA, RSWF-BLP, LDASR and GCRFLDA models achieved AUC values of 0.9573, 0.9505, 0.9325, 0.9654, 0.8908 and 0.9547, respectively, while MAGCNSE achieved the highest AUC value of 0.9812, outperforming other models by 1.58–9.04%. Besides, the LDNFSGB, IPCARF, VGAELDA, RSWF-BLP, LDASR and GCRFLDA models achieved AUPR values of 0.9543, 0.9607, 0.9547, 0.9686, 0.9102 and 0.9611, respectively, while MAGCNSE achieved the highest AUPR value of 0.9849, outperforming other models by 1.63–7.47%. The superiority of MAGCNSE over the other

Table 3 Comparison of the AUC values and AUPR values of MAGCNSE using GCN and other graph models

Method	GAT	GraphSAGE	GCN
AUC	0.9668	0.9713	**0.9812**
AUPR	0.9713	0.9723	**0.9849**
Accuracy	0.9045	0.9188	**0.9395**
Sensitivity	0.8929	**0.9231**	0.9192
Specificity	0.9156	0.9142	**0.9626**
Precision	0.9106	0.9202	**0.9654**
*F*1- *score*	0.9016	0.9217	**0.9417**
MCC	0.8089	0.8374	**0.88**

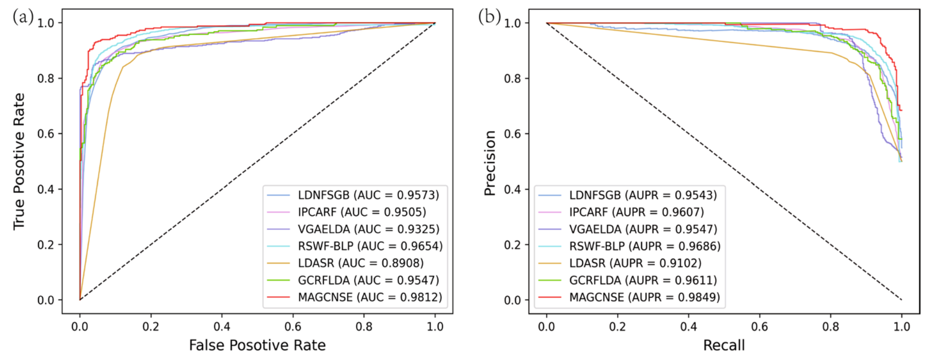

Fig. 4 ROC curves (a) and PR curves (b) of MAGCNSE and other state-of-the-art methods

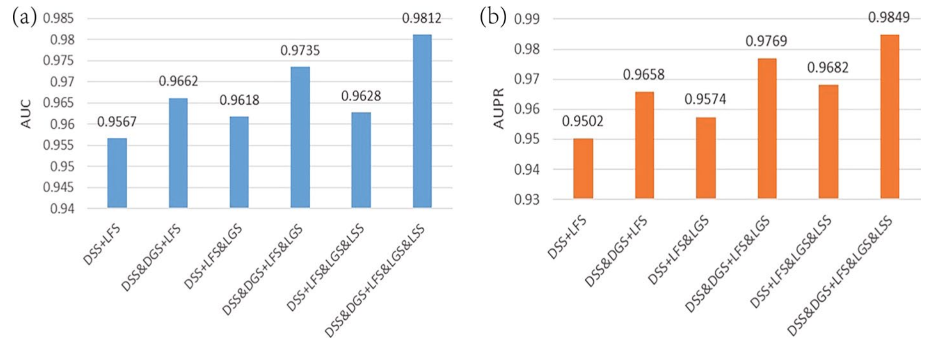

Fig. 5 AUC values and AUPR values of MAGCNSE using different views

state-of-the-art methods further proves that MAGCNSE is competent and reliable in predicting underlying LDAs. The detailed parameters used by the seven methods are added into the Additional file 1: Table S1.

Effect of different views

In order to confirm whether the results are better as expected after using multi-view features, we applied 5-CV to compare the AUC value and AUPR value of MAGCNSE under different views.

It can be known from Fig 5 that using multi-view features can generally enhance the performance of MAGCNSE, and MAGCNSE achieves the best performance when all views of lncRNAs and diseases were leveraged in this study. In most cases, as the number of views increased, the AUC and AUPR values also increased. The possible reason could be that different views contain different information, and the node features are enriched by fusing different views.

Case studies

In order to further verify the performance of MAGCNSE in predicting the associations between lncRNAs and some specific diseases, we conducted two types of case studies. Our data were all obtained from LncRNADisease v2.0 [44] (http://www.rnanut.net/lncrnadisease/) and used for the model training. The PubMed literature and two external databases of Lnc2Cancer 3.0 [45] (http://bio-bigdata.hrbmu.edu.cn/lnc2cancer/) and MNDR v3.1 [46] (https://www.rna-society.org/mndr/) were used for verifying the results.

In the first type of case studies, we aimed at verifying the performance of MAGCNSE for unknown LDAs prediction. For a specific disease, the detailed steps of case studies are as follows. Step 1: use all known LDAs as the positive samples, and randomly select the same number of negative samples from the unknown LDAs, the negative samples do not involve the specific disease. Step 2: select all unknown associations between lncRNAs and the specific disease as the testing samples. Step 3: after training MAGCNSE using the positive and negative samples, use it to test the lncRNA-disease testing samples, and record the prediction scores of the testing samples. Step 4: sort the prediction scores from the highest to the lowest, and find the top 10 lncRNAs related to that disease. Step 5: validate the results according to Lnc2Cancer 3.0 and MNDR v3.1. If no evidence is found in the two databases, then refer it to PubMed literature. Here, we selected colon cancer and lung cancer as the research subjects.

Colon cancer is one of the most serious cancers that is related to the digestive system [47]. Table 4 illustrates that eight of the top 10 lncRNAs were confirmed. For example, colon cancer's epithelial-mesenchymal transition process is affected by AFAP1-AS1 [48]. The capacity of colon cancer cells to proliferate and migrate is impaired when PCAT1 expression is suppressed [49].

Lung cancer is a common cause for death globally, which includes non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC) [50]. Table 5 illustrates that eight of the top 10 lncRNAs were confirmed. For example, through targeting miR-150-5p/HMGA2 signaling, lncRNA-ZFAS1 knockdown inhibits NSCLC progression [51]. CRNDE acts as an oncogene to sponge miR-338-3p, which plays a crucial regulatory role in regulating NSCLC development [52].

To demonstrate whether MAGCNSE is capable of accurately retrieving known LDAs for a specific disease, we conducted the second type of case studies. For a specific disease, the detailed steps are as follows. Step 1: remove all associations related the specific disease from the known LDAs to treat it as a new disease, use the remaining known LDAs as the positive samples, and randomly select the same number of negative samples from the unknown LDAs, the negative samples do not involve the

Table 4 The top 10 predicted colon cancer-associated lncRNAs

Rank	lncRNA name	Evidence
1	CDKN2B-AS1	MNDR v3.1
2	NPTN-IT1	Unconfirmed
3	HOXA11-AS	Unconfirmed
4	AFAP1-AS1	Lnc2Cancer 3.0, MNDR v3.1
5	PCAT1	PMID:33277833
6	GAS5	Lnc2Cancer 3.0, MNDR v3.1
7	CRNDE	MNDR v3.1
8	CASC2	PMID:32655801
9	SNHG16	Lnc2Cancer 3.0, MNDR v3.1
10	SPRY4-IT1	PMID:28651500

Table 5 The top 10 predicted lung cancer-associated lncRNAs

Rank	lncRNA name	Evidence
1	ZFAS1	PMID: 31692094
2	LINC-ROR	Lnc2Cancer 3.0, MNDR v3.1
3	CRNDE	PMID: 30554121
4	HOXA11-AS	Lnc2Cancer 3.0, MNDR v3.1
5	CYTOR	MNDR v3.1
6	PTENP1	Unconfirmed
7	XIST	MNDR v3.1
8	DRAIC	PMID: 30544991
9	NEAT1	Lnc2Cancer 3.0, MNDR v3.1
10	NPTN-IT1	Unconfirmed

Table 6 The top 10 predicted cervical cancer-associated lncRNAs

Rank	lncRNA name	Evidence
1	CCAT2	LncRNADisease v2.0
2	MALAT1	LncRNADisease v2.0
3	H19	LncRNADisease v2.0
4	TUG1	LncRNADisease v2.0
5	CDKN2B-AS1	LncRNADisease v2.0
6	UCA1	LncRNADisease v2.0
7	HOTAIR	LncRNADisease v2.0
8	MEG3	LncRNADisease v2.0
9	CCAT1	LncRNADisease v2.0
10	GAS5	LncRNADisease v2.0

specific disease. Step 2: select the sample pairs between all lncRNAs and the specific disease as the testing samples. Step 3: after MAGCNSE is trained using the positive and negative samples, use it to test the lncRNA-disease testing samples, and record the prediction scores of the testing samples. Step 4: sort the prediction scores from the highest to the lowest, and find the top 10 lncRNAs related to that disease. Step 5: validate the results by referring to LncRNADisease v2.0. If no evidence is found in this database, then refer it to Lnc2Cancer 3.0, MNDR v3.1 and PubMed literature. Here, cervical cancer was chosen as the research subject.

Cervical cancer is a very prevalent condition in women [53]. Table 6 shows that all of the top 10 lncRNAs were confirmed by LncRNADisease v2.0, which means that MAGCNSE could retrieve known LDAs for a single disease with a high accuracy. For example, knockdown of CCAT2 could trigger the apoptosis of cervical cancer cells and CCAT2 have promotive effect on cervical cancer cells' proliferation and survival [54]. Overexpression of HOTAIR is related to cervical cancer progression; thus, it could be further investigated for diagnosis and gene therapy [55].

The detailed prediction scores of all predicted lncRNAs with the above-mentioned diseases are given in Additional file 1: Table S2, Additional file 1: Table S3 and Additional file 1: Table S4.

Conclusions

The prediction of potential LDAs can help to detect disease biomarkers and perform disease analysis and prevention, using computational methods to efficiently predict LDAS is of great importance. In this study, we developed a novel model called MAGCNSE to identify potential LDAs. MAGCNSE first uses GCN to fuse multi-view similarity graphs of lncRNAs and diseases and obtain multiple feature matrices. Then, it applies the attention mechanism to adaptively assign the weights to different feature matrices. Next, it further extracts features with the use of the CNN to get the final representations of lncRNAs and diseases. Finally, it utilizes a stacking ensemble classifier to make the predictions. Compared with previous models in the field of LDA prediction, multi-view data of lncRNAs and diseases were used in this study, and MAGCNSE used lncRNA sequence similarity, then MAGCNSE utilized deep learning methods rather than linear methods for data fusion to learn the representations of lncRNAs and diseases, and MAGCNSE employed a stacking ensemble model rather than single ML classifiers for the final prediction task. We performed experiments on the effect of parameters, ablation studies in both representation learning methods and classification methods, experiments comparing GCN with two other GNN models, comparison studies with other state-of-the-art methods, experiments on the effect of different views and two types of case studies. All results demonstrate the outstanding performance of MAGCNSE in predicting potential LDAs.

However, there are still some aspects in our study that can be further investigated. Firstly, we only use the information of lncRNAs and diseases, there are some other biological information such as miRNA, protein and drug could also be considered for further research. In addition, the way to select, integrate and extract the features of lncRNAs and diseases for by more effective and superior deep learning methods is a long-term challenge in the future.

Methods

Human lncRNA-disease associations

In this study, we retrieved known LDAs from LncRNADisease v2.0, which includes 10564 experimentally validated associations between 6105 lncRNAs and 451 diseases among several species. First, we selected only human LDAs and removed duplicated records, then filtered out lncRNAs with no sequence information from NONCODE v6.0 [56] (http://www.noncode.org/) and diseases with no DOID from Disease Ontology [57] (https://disease-ontology.org/). Finally, we obtained 1569 human LDAs between 489 lncRNAs and 251 diseases. We define an adjacency matrix $LD \in R^{l\times d}$ to represent LDAs, such that $LD(i,j) = 1$ if lncRNA l_i interacts with disease d_j, otherwise $LD(i,j) = 0$.

Disease semantic similarity

In studies of ncRNA-disease associations, DSS has been extensively used in recent years and has been proved to be effective. It is calculated by Wang's method [58], in which the Medical Subject Headings (MeSH) descriptions of diseases is downloaded from the National Library of Medicine (https://www.nlm.nih.gov/), and the directed acyclic graphs (DAGs) for diseases can be constructed afterwards. The disease d_i is defined such that $DAG(d_i) = (d_i, D(d_i))$, where $D(d_i)$ represents all ancestor nodes of d_i and node d_i itself. For each disease t that belongs to $D(d_i)$, its contribution to disease d_i can be computed as follows:

$$\begin{cases} DS_{d_i}(t) = 1 & if\ t = d_i \\ DS_{d_i}(t) = max\left\{\xi \times DS_{d_i}(t') \mid t' \in D(d_i)\right\} otherwise \end{cases} \tag{7}$$

where ξ denotes a contribution factor, it's generally set to 0.5.

The total contributions of $D(d_i)$ to disease d_i can be computed as follows:

$$DC(d_i) = \sum_{t \in D(d_i)} DS_{d_i}(t) \tag{8}$$

Then, the DSS matrix can be computed as follows:

$$DSS(d_i, d_j) = \frac{\sum_{t \in D(d_i) \cap D(d_j)} (DS_{d_i}(t) + DS_{d_j}(t))}{DC(d_i) + DC(d_j)} \tag{9}$$

We used the DOSE software package [59] to calculate the DSS. We obtained the unique DOID of each disease from Disease Ontology, and then utilized the function doSim of the DOSE software and selected the measure method of "Wang" to get the DSS matrix.

LncRNA functional similarity

It has been previously observed that functionally comparable lncRNAs are frequently linked with similar diseases. We followed the previous works [60] to calculate LFS in this work. Given that lncRNAs l_i and l_j are relevant to p diseases and q diseases, respectively, then the LFS can be calculated as:

$$LFS(l_i, l_j) = \frac{\sum_{d \in D(l_j)} S(d, D(l_i)) + \sum_{d \in D(l_i)} S(d, D(l_j))}{p + q} \tag{10}$$

$$S(d_m, D(l_i)) = \max_{d \in D(l_i)} (DSS(d_m, d)) \tag{11}$$

where $D(l_i)$ represents the disease set associated with lncRNA l_i.

LncRNA sequence similarity

Following previous studies [61, 62], we utilized Levenshtein distance [63] to calculate LSS. The Levenshtein distance means the minimum cost of converting one string to another string through the insertion, deletion, or replacement of a single character. In previous studies, the editing cost was set to 2, while the insertion cost and deletion cost were set to 1, and we followed the same criterion in our study. The LSS is calculated as follows:

$$LSS(l_i, l_j) = 1 - \frac{dist}{len(l_i) + len(l_j)} \tag{12}$$

where *dist* denotes the minimum cost of converting lncRNA l_i sequence to l_j sequence, *len* is length of lncRNA sequence.

Gaussian interaction profile kernel similarity for lncRNAs and diseases

Based on previous works [58], LGS can be computed as:

$$LGS(l_i, l_j) = \exp(-\eta_l \left\| LD(i,:) - LD(j,:) \right\|^2) \tag{13}$$

$$\eta_l = 1 \Big/ \left(\frac{1}{N_l} \sum_{i=1}^{N_l} \| LD(i,:) \|^2 \right) \tag{14}$$

where η_l denotes the standardized core bandwidth for lncRNA similarity calculation which is generally set to 1, and N_l denotes the number of lncRNAs.

Similarly, for diseases, DGS is computed as follows:

$$DGS(d_i, d_j) = \exp(-\eta_d \left\| LD(:,i) - LD(:,j) \right\|^2) \tag{15}$$

$$\eta_d = 1/(\frac{1}{N_d} \sum_{i=1}^{N_d} \| LD(:,i) \|^2) \tag{16}$$

where η_d denotes the standardized core bandwidth for disease similarity calculation, and N_d denotes the number of diseases.

Model framework

The main workflow of MAGCNSE is shown in Fig 6, consisting of four steps. (1) Since the similarity matrices between lncRNAs and diseases can be regarded as graph structures, we extracted the features from similarity graphs of different views of lncRNAs and diseases via GCN to obtain multiple feature matrices. (2) Attention mechanism was applied on the acquired feature matrices of lncRNAs and diseases to adaptively capture the importance and assign weights to them. (3) We used the CNN to further extract features from multi-channel feature matrices to acquire the final representations of lncRNAs and diseases. During the above-mentioned procedures, a temporary matrix was calculated in each training epoch, such that each element of it was the corresponding dot product of each lncRNA representation and disease representation. Then, the difference between the lncRNA-disease adjacency matrix and temporary matrix was obtained, and the Frebious norm of it was later computed. Subsequently, the parameters of the model were updated in each training epoch by minimizing the Frebious norm. (4) For the positive and negative lncRNA-disease pairs concatenated by the representaion of

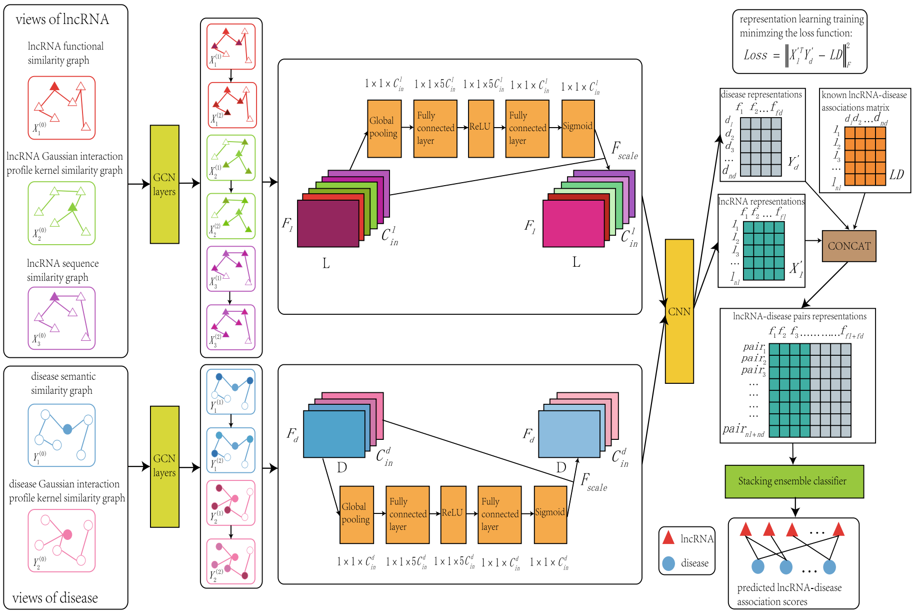
views of lncRNA
lncRNA functional similarity graph
lncRNA Gaussian interaction profile kernel similarity graph
lncRNA sequence similarity graph
disease semantic similarity graph
disease Gaussian interaction profile kernel similarity graph
views of disease
GCN layers
GCN layers
Global pooling
Fully connected layer
ReLU
Fully connected layer
Sigmoid
1 × 1 × C_in^l
1 × 1 × 5C_in^l
1 × 1 × 5C_in^l
1 × 1 × C_in^l
1 × 1 × C_in^l
F_scale
F_l
C_in^l
L
F_d
C_in^d
D
1 × 1 × C_in^d
1 × 1 × 5C_in^d
1 × 1 × 5C_in^d
1 × 1 × C_in^d
1 × 1 × C_in^d
CNN
representation learning training minimzing the loss function:
Loss = ‖X'_l^T Y'_d − LD‖_F^2
disease representations
f_1 f_2…f_fd
d_1 d_2 d_3 … d_nd
Y'_d
known lncRNA-disease associations matrix
d_1 d_2 … d_nd
l_1 l_2 l_3 … l_nl
LD
lncRNA representations
f_1 f_2 … f_fl
l_1 l_2 l_3 … l_nl
X'_l
CONCAT
lncRNA-disease pairs representations
f_1 f_2 f_3……….f_fl+fd
pair_1 pair_2 pair_3 … pair_nl+nd
Stacking ensemble classifier
lncRNA
disease
predicted lncRNA-disease association scores

Fig. 6 The flowchart of MAGCNSE. Step 1: extract features from the 3 views of similarity graphs of lncRNAs and 2 views of similarity graphs diseases utilizing GCN. Step 2: leverage attention mechanism for adaptively assigning weights to different feature matrices of lncRNAs and diseases. Step 3: acquire the final representations of lncRNAs and diseases by further extracting features from the multi-channel feature matrices of lncRNAs and diseases using the CNN. Step 4: employ a stacking ensemble classifier to make LDA predictions

lncRNAs and diseases, the stacking ensemble classifier, consisting of multiple traditional ML classifiers was leveraged to perform LDA predictions.

Multi-view graph convolutional network

Due to its excellent capacity of data processing and suitability for data with a graph structure, GCN has been extensively used in bioinformatics and other fields in recent years [64]. GCN can aggregate the information of neighbor nodes to obtain the dependency relationship between nodes and extract the data features. In our work, GCN was applied with the purpose of extracting features of the lncRNA and disease similarity matrices under diverse views, as illustrated in Fig 6. G_l^r and G_d^s denote the specific view of the lncRNA and disease, respectively. Given that lncRNA l_i is denoted as $x_i \in R^{1\times p}$, the neighbors of the lncRNA in view r are represented as $\{i_1, i_2, \ldots, i_t\}$, and the related features of the neighbors are represented as $\left\{x_{i_1}, x_{i_2}, \ldots, x_{i_t}\right\}$. When the embedding of a lncRNA node is learned, the similarity with the neighbor nodes should be considered. Then, the representation of the i-th lncRNA under view r can be calculated by the following formula:

$$x_i' = ReLU\left(\left(\widetilde{r_{i,i}}\, x_i + \sum_{j=1}^{t} \widetilde{r_{i,j}}\, x_{ij}\right) W_i\right) \tag{17}$$

where $\widetilde{r_{i,j}}$ denotes the normalized similarity weights between the i-th lncRNA and its neighbor i_j under view r, while $W_i \in R^{p \times F_l}$ denotes the weight parameters that project the original feature of the i-th lncRNA into the latent feature.

Given the propagation formula of single lncRNA nodes in view r, the representations of the lncRNA nodes on the graph G_l^r can be acquired as follows:

$$X_r^{(l+1)} = ReLU\left(\widetilde{D_r}^{-\frac{1}{2}}\widetilde{R}\widetilde{D_r}^{-\frac{1}{2}}X_r^{(l)}W_r^{(l)}\right) \tag{18}$$

$$\widetilde{R} = I + R \tag{19}$$

where $X_r^{(l)} \in R^{L \times F_l}$ denotes the F_l embedding of L lncRNAs in the l-th GCN layer in view r. Specifically, the value of the initial embedding $X_r^{(0)}$ is randomly generated.$W_r^{(l)} \in R^{F_l \times F_l}$ denotes the weight parameters, R denotes the similarity matrix of all lncRNAs,$\widetilde{R}$ is the normalized similarity weights of lncRNAs in view r, and $\widetilde{D_r}$ is the diagonal matrix which is computed as follows:

$$\widetilde{D_r}(i,i) = \sum\nolimits_j \widetilde{R}(i,j) \tag{20}$$

Similarly, the representations of the disease nodes on graph G_d^s can be calculated as follows:

$$Y_s^{(l+1)} = ReLU\left(\widetilde{D_s}^{-\frac{1}{2}}\widetilde{S}\widetilde{D_s}^{-\frac{1}{2}}Y_s^{(l)}W_s^{(l)}\right) \tag{21}$$

$$\widetilde{S} = I + S \tag{22}$$

where $Y_s^{(l)} \in R^{T \times F_d}$ denotes the F_d embedding of T diseases in the l-th GCN layer in view s. Specifically, $Y_s^{(0)}$ denotes the initial embedding value, which is randomly generated. $W_s^{(l)} \in R^{F_d \times F_d}$ denotes the weight parameters, $\widetilde{S}$ denotes the normalized similarity weights of diseases in view s, and $\widetilde{D_s}$ is the corresponding diagonal matrix.

Given the embeddings of lncRNAs and diseases in multiple GCN layers in diverse views and that the GCN has l layers, the embeddings of lncRNAs in view r and those of diseases in view s can be denoted as follows:

$$\left\{X_r^{(1)}, X_r^{(2)}, \ldots, X_r^{(l)}\right\} \tag{23}$$

$$\left\{Y_s^{(1)}, Y_s^{(2)}, \ldots, Y_s^{(l)}\right\} \tag{24}$$

Finally, the features of lncRNAs in R views and the features of diseases in S views extracted by the GCN are as follows:

$$\left\{\left\{X_1^{(1)}, X_1^{(2)}, \ldots, X_1^{(l)}\right\}, \left\{X_2^{(1)}, X_2^{(2)}, \ldots, X_2^{(l)}\right\}, \ldots, \left\{X_R^{(1)}, X_R^{(2)}, \ldots, X_R^{(l)}\right\}\right\} \tag{25}$$

$$\left\{\left\{Y_1^{(1)}, Y_1^{(2)}, \ldots, Y_1^{(l)}\right\}, \left\{Y_2^{(1)}, Y_2^{(2)}, \ldots, Y_2^{(l)}\right\}, \ldots, \left\{Y_S^{(1)}, Y_S^{(2)}, \ldots, Y_S^{(l)}\right\}\right\} \tag{26}$$

Attention mechanism

We found the multiple feature matrices under different views to be similar to multiple channels of an image, but with potentially different importance. With reference to the study [65], we applied the technique of the attention mechanism to adaptively capture the importance and assign weights to feature matrices of lncRNAs and diseases. First, channel-wise statistics were obtained through a global average pooling operation. For lncRNA, we define a statistic $Z \in R^{1 \times 1 \times C_{in}^{l}}$, which can be obtained by squeezing the lncRNA feature matrices set $X_l \in R^{F_l \times L \times C_{in}^{l}}$ via the spatial dimensions of $F_l \times L$, where $X_l = [x_1, x_2, \ldots, x_{C_{in}^{l}}]$. The k-th element of Z was calculated as:

$$z_k = F_{sq}(x_k) = \frac{1}{F_l \times L} \sum_{i=1}^{F_l} \sum_{j=1}^{L} x_k(i,j) \tag{27}$$

where x_k is the k-th feature matrices of the lncRNA.

Then, the attention weights for the feature matrices of lncRNA can be calculated as follows:

$$Z_{att} = F_{atten}(Z, W_{in}^{l}) = \sigma(W_2 \delta(W_1 Z)) \tag{28}$$

where σ and δ represent the Sigmoid function and ReLU function, respectively, $W_1 \in R^{(C_{in}^{l} \times \mu) \times C_{in}^{l}}$ and $W_2 \in R^{C_{in}^{l} \times (C_{in}^{l} \times \mu)}$ denote the weight parameters in the first and second fully connected layers, respectively. The μ is a hyperparameter, we chose the value of μ from {2,3,4,5,6} and kept other parameters in MAGCNSE unchanged to find the relatively optimal value of μ in this study. The AUC value and AUPR value of MAGCNSE using different values of μ are given in Additional file 1: Table S5, from which we can see that MAGCNSE achieves the best performance when the value of μ is 5, so we set μ to 5 in this study.

Given the weight of each feature matrix of lncRNA, each normalized feature matrix of lncRNA can be obtained as follows:

$$\widetilde{x_k} = F_{scale}(x_k, z_k^{att}) = z_k^{att} \bullet x_k \tag{29}$$

Therefore, the entire normalized feature matrices of lncRNA can be denoted as $\widetilde{X_l} = [\widetilde{x_1}, \widetilde{x_2}, \ldots, \widetilde{x_{C_{in}^{l}}}]$. Analogously, the entire normalized feature matrices of disease $\widetilde{Y_d} = [\widetilde{y_1}, \widetilde{y_2}, \ldots, \widetilde{y_{C_{in}^{d}}}]$ can be obtained by the same above-mentioned steps.

Convolutional neural network

The normalized multiple channel's feature matrices of lncRNAs and diseases can be regarded as an image of lncRNAs and an image of diseases, respectively. In the bioinformatics field, CNNs have become extensively exploited due to their excellent image processing abilities in recent years [66, 67]. Therefore, we utilized the CNN to further extract the features of lncRNAs and diseases. Given that $\widetilde{X_l} = [\widetilde{x_1}, \widetilde{x_2}, \ldots, \widetilde{x_{C_{in}^l}}]$, the embedding of the q-th output channel can be calculated as follows:

$$Lout_q = \sum_{i=1}^{C_{in}^l} \widetilde{x_i} \otimes w_q^l + b_q \tag{30}$$

where $\otimes$ means the convolution operation, $w_q^l \in R^{F_l \times 1}$ denotes the q-th convolution filter, while b_q denotes the q-th bias.

Then, the final lncRNA representations $X_l' \in R^{C_{out}^l \times L}$ can be obtained by stacking the embeddings of all channels, it is defined as:

$$X_l' = stack(Lout_q) \tag{31}$$

Analogously, the final disease representations Y_d' can be obtained.

During the above-mentioned procedures, MAGCNSE calculates a temporary matrix LD' in each training epoch, which is defined as:

$$LD' = X_l^{'\,T} \bullet Y_d^{'} \tag{32}$$

Each element of LD' represents the dot product of each corresponding lncRNA representation and disease representation. Then, the difference between LD and LD' is obtained, we define the Frebious norm of it as the $Loss$, which can be computed as follows:

$$Loss = \left\| LD' - LD \right\|_F^2 \tag{33}$$

The parameters of the model are updated in each training epoch by minimizing the $Loss$ term.

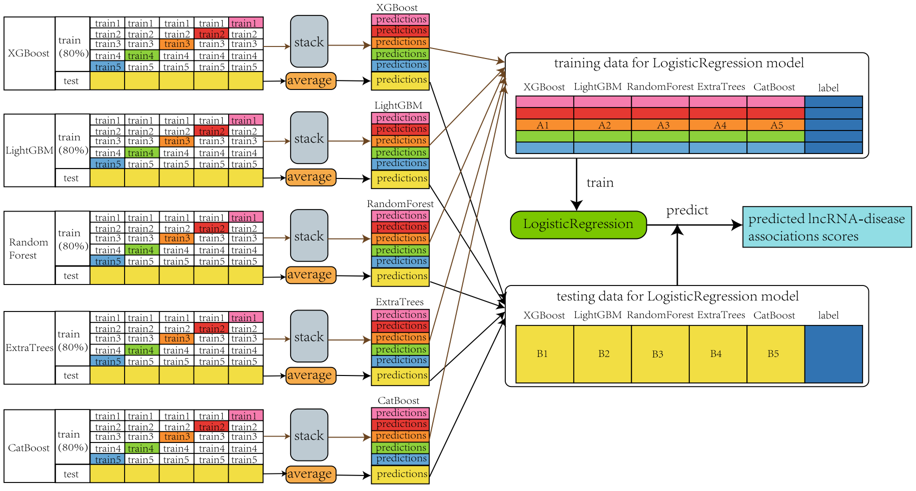
XGBoost
LightGBM
Random Forest
ExtraTrees
CatBoost
train (80%)
test
train1
train2
train3
train4
train5
stack
average
predictions
training data for LogisticRegression model
XGBoost LightGBM RandomForest ExtraTrees CatBoost label
A1
A2
A3
A4
A5
train
LogisticRegression
predict
predicted lncRNA-disease associations scores
testing data for LogisticRegression model
B1
B2
B3
B4
B5

Fig. 7 The flowchart of the stacking ensemble classifier

Table 7 Key hyperparameters of the six traditional classifiers and their optimal value after grid search

Method	Optimal hyperparameters
RandonForest	max_feature=10; min_sample_split=2; n_estimators=2000
ExtraTrees	max_feature=10; min_sample_split=2; n_estimators=2000
XGBoost	learning_rate=0.05; max_depth=4; gamma=0; n_estimators=1000
LightGBM	learning_rate=0.15; max_depth=10; num_leaves=31; n_estimators=200
CatBoost	depth=3; iteration=800; learning_rate=0.1; border_count=32; l2_leaf_reg=5
LogisticRegression	C=20.0; max_iter=40; penalty='l2'

Stacking ensemble classifier

Fig 7 shows the stacking ensemble framework, containing two layers. The base classifiers were five classic tree-based classifiers (XGBoost, LightGBM, RandomForest, ExtraTrees, CatBoost), which is generally capable of processing unnormalized features well [68]. Meanwhile, LogisticRegression was applied as the meta classifier for the results of the five above-mentioned base classifiers. For base classifiers, we used a grid search approach with 5-CV to identify the optimal hyperparameters. In the following, we explain the detailed process of the stacking ensemble.

(1) We use 80% and 20% of the datasets as the training set and testing set, respectively. (2) The base classifier was trained via 5-CV using the training set. For each cross-validation, the base classifier calculated the prediction values in the training and testing datasets, separately. (3) For base classifiers, MAGCNSE integrated the prediction results from the training dataset, which are marked as A1, A2, A3, A4 and A5, they were used as the training dataset of the subsequent LogisticRegression algorithm. Besides, MAGCNSE calculated the average value of the prediction results on the testing dataset, which are marked as B1, B2, B3, B4 and B5, they were used as the testing dataset of the LogisticRegression algorithm. (4) The LogisticRegression classifier searched for the optimal hyperparameters by utilizing a grid search with 5-CV on the integrated training dataset, then we used the integrated training dataset to train it. (5) Finally, the LogisticRegression classifier predicted the testing samples and obtained the final predicted class labels and probabilities for each lncRNA-disease pair.

The key hyperparameters of the six traditional classifiers and their optimal value after grid search are shown in Table 7.

Abbreviations

lncRNAs: Long non-coding RNAs; LDAs: LncRNA-disease associations; ML: Machine learning; PCA: Principal component analysis; CNN: Convolutional neural network; RF: Random forest; GCN: Graph convolutional network; DSS: Disease semantic similarity; DGS: Disease Gaussian interaction profile kernel similarity; LFS: LncRNA functional similarity; LSS: LncRNA sequence similarity; LGS: LncRNA Gaussian interaction profile kernel similarity; 5-CV: 5-fold cross-validation; ROC: Receiver-operating characteristic; AUC: Area under the ROC curve; PR: Precision-recall; AUPR: Area under the PR curve; MCC: Matthews correlation coefficient; GNN: Graph neural network; GAT: Graph attention network; GraphSAGE: Graph sample and aggregate; NSCLC: Non-small-cell lung cancer; SCLC: Small-cell lung cancer; MeSH: Medical subject headings; DAGs: Directed acyclic graphs..

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04715-w.

Additional file 1: Table S1. The detailed parameters of seven state-of-the-art methods in this study. **Table S2**. The detailed prediction scores of all predicted lncRNAs with colon cancer. **Table S3**. The detailed prediction scores of all predicted lncRNAs with lung cancer. **Table S4**. The detailed prediction scores of all predicted lncRNAs with cervical cancer. **Table S5**. AUC and AUPR values of MAGCNSE using different values of μ.

Acknowledgements

The authors thank the editor and anonymous reviewers for their constructive comments and suggestions, which greatly help us improve our manuscript.

Author contributions

Z-QZ, YL, C-LG and Y-LW designed the methodology. Z-QZ and N-NL developed and implemented the method. Z-QZ and Y-NW performed the experiments and analyzed the results. Z-QZ and YL wrote and revised the manuscript. All authors read and approved the final manuscript.

Funding

This work is supported by the National Natural Science Foundation of China under Grant Nos.62163018 and the Natural Science Foundation of Hunan Province under Grant Nos.2020JJ4209.

Availability of data and materials

The source code and datasets analysed during the current study are available at https://github.com/YingLiangjxau/MAGCNSE. All data used in the paper, including the data of lncRNA-disaese associations, the DOIDs of diseases, and the sequences of lncRNAs, were obtained from current public databases and were cited in the text.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]College of Computer and Information Engineering, Jiangxi Agricultural University, Nanchang, China. [2]College of Information Science and Engineering, Hunan University, Changsha, China.

Received: 15 March 2022 Accepted: 5 May 2022

Published online: 19 May 2022

References

1. Nagano T, Fraser P. No-nonsense functions for long noncoding RNAs. Cell. 2011;145(2):178–81.
2. Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: functional surprises from the RNS world. Genes Develop. 2009;23(13):1494–504.
3. Zhang Y, Yuan J, Gao Z, Zhang Z. LncRNA tuc338 promotes invasion of lung cancer by activating mapk pathway. Eur Rev Med Pharmacol Sci. 2018;22(2):443–9.
4. Loewen G, Jayawickramarajah J, Zhuo Y, Shan B. Functions of lncRNA hotair in lung cancer. J Hematol Oncol. 2014;7(1):1–10.
5. Yeh C-C, Luo J-L, Nhut Phan N, Cheng Y-C, Chow L-P, Tsai M-H, Chuang EY, Lai L-C. Different effects of long noncoding RNA ndrg1-ot1 fragments on ndrg1 transcription in breast cancer cells under hypoxia. RNA Biol. 2018;15(12):1487–98.
6. Zheng Y, Wang M, Wang S, Xu P, Deng Y, Lin S, Li N, Liu K, Zhu Y, Zhai Z, et al. LncRNA meg3 rs3087918 was associated with a decreased breast cancer risk in a Chinese population: a case-control study. BMC Cancer. 2020;20(1):1–8.
7. Wu M, Huang Y, Chen T, Wang W, Yang S, Ye Z, Xi X. LncRNA meg3 inhibits the progression of prostate cancer by modulating mir-9-5p/qki-5 axis. J Cell Mol Med. 2019;23(1):29–38.
8. Chakravarty D, Sboner A, Nair SS, Giannopoulou E, Li R, Hennig S, Mosquera JM, Pauwels J, Park K, Kossai M, et al. The oestrogen receptor alpha-regulated lncRNA neat1 is a critical modulator of prostate cancer. Nat Commun. 2014;5(1):1–16.
9. Sun M, Xia R, Jin F, Xu T, Liu Z, De W, Liu X. Downregulated long noncoding RNA meg3 is associated with poor prognosis and promotes cell proliferation in gastric cancer. Tumor Biol. 2014;35(2):1065–73.
10. Yao X, Tang J, Zhu H, Jing Y. High expression of lncRNA casc15 is a risk factor for gastric cancer prognosis and promote the proliferation of gastric cancer. Eur Rev Med Pharmacol Sci. 2017;21(24):5661–7.

11. Wu Q, Meng W-Y, Jie Y, Zhao H. LncRNA malat1 induces colon cancer development by regulating mir-129-5p/hmgb1 axis. J Cell Physiol. 2018;233(9):6750–7.
12. Li Y, Li C, Li D, Yang L, Jin J, Zhang B. lncRNA kcnq1ot1 enhances the chemoresistance of oxaliplatin in colon cancer by targeting the mir-34a/atg4b pathway. OncoTargets Ther. 2019;12:2649.
13. Doxtater K, Tripathi MK, Khan MM. Recent advances on the role of long non-coding RNAs in Alzheimer's disease. Neural Regen Res. 2020;15(12):2253.
14. Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, Morgan TE, Finch CE, Laurent GS III, Kenny PJ, Wahlestedt C. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase. Nature Med. 2008;14(7):723–30.
15. Lu C, Yang M, Luo F, Wu F-X, Li M, Pan Y, Li Y, Wang J. Prediction of lncRNA-disease associations based on inductive matrix completion. Bioinformatics. 2018;34(19):3357–64.
16. Hu J, Gao Y, Li J, Zheng Y, Wang J, Shang X. A novel algorithm based on bi-random walks to identify disease-related lncRNAs. BMC Bioinform. 2019;20(18):1–11.
17. Wang Y, Yu G, Wang J, Fu G, Guo M, Domeniconi C. Weighted matrix factorization on multi-relational data for lncRNA-disease association prediction. Methods. 2020;173:32–43.
18. Zeng M, Lu C, Fei Z, Wu F, Li Y, Wang J, Li M. Dmflda: A deep learning framework for predicting incRNA–disease associations. IEEE/ACM transactions on computational biology and bioinformatics 2020;
19. Zhao X, Yang Y, Yin M. Mhrwr: Prediction of lncRNA-disease associations based on multiple heterogeneous networks. IEEE/ACM transactions on computational biology and bioinformatics 2020;
20. Lan W, Li M, Zhao K, Liu J, Wu F-X, Pan Y, Wang J. Ldap: a web server for lncRNa-disease association prediction. Bioinformatics. 2017;33(3):458–60.
21. Xuan P, Jia L, Zhang T, Sheng N, Li X, Li J. Ldapred: a method based on information flow propagation and a convolutional neural network for the prediction of disease-associated lncRNAs. Int J Mol Sci. 2019;20(18):4458.
22. Wei H, Liao Q, Liu B. ilncrnadis-fb: identify lncRNA-disease associations by fusing biological feature blocks through deep neural network. IEEE/ACM transactions on computational biology and bioinformatics 2020;
23. Yao D, Zhan X, Kwoh C-K. An improved random forest-based computational model for predicting novel miRNa-disease associations. BMC Bioinform. 2019;20(1):1–14.
24. Zeng M, Lu C, Zhang F, Li Y, Wu F-X, Li Y, Li M. Sdlda: lncRNA-disease association prediction based on singular value decomposition and deep learning. Methods. 2020;179:73–80.
25. Wu X, Lan W, Chen Q, Dong Y, Liu J, Peng W. Inferring lncRNA-disease associations based on graph autoencoder matrix completion. Comput Biol Chem. 2020;87:107282.
26. Wu Q.-W, Xia J.-F, Ni J.-C, Zheng C.-H. Gaerf: predicting lncRNA-disease associations by graph auto-encoder and random forest. Briefings in bioinformatics 2021;
27. Silva A.B.O.V, Spinosa E.J. Graph convolutional auto-encoders for predicting novel lncRNA-disease associations. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2021;
28. Liu P, Luo J, Chen X. mircom: Tensor completion integrating multi-view information to deduce the potential disease-related miRNA-miRNA pairs. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2020;
29. Wang L, You Z-H, Chen X, Li Y-M, Dong Y-N, Li L-P, Zheng K. Lmtrda: Using logistic model tree to predict miRNA-disease associations by fusing multi-source information of sequences and similarities. PLoS Comput Biol. 2019;15(3):1006865.
30. Pan X, Shen H-B. Scoring disease-microRNA associations by integrating disease hierarchy into graph convolutional networks. Pattern Recognit. 2020;105:107385.
31. Tang X, Luo J, Shen C, Lai Z. Multi-view multichannel attention graph convolutional network for miRNA–disease association prediction. Briefings in Bioinformatics 2021;
32. Guo Z-H, You Z-H, Wang Y-B, Yi H-C, Chen Z-H. A learning-based method for lncRNA-disease association identification combing similarity information and rotation forest. Science. 2019;19:786–95.
33. Zhang Y, Ye F, Xiong D, Gao X. Ldnfsgb: prediction of long non-coding RNA and disease association using network feature similarity and gradient boosting. BMC bioinformatics. 2020;21(1):1–27.
34. Madhavan M, Gopakumar G. Dbnlda: Deep belief network based representation learning for lncRNA-disease association prediction. Applied Intelligence,2021; 1–11
35. Zhang Y, Ye F, Gao X. Mca-net: multi-feature coding and attention convolutional neural network for predicting lncRNA-disease association. IEEE/ACM Transactions on Computational Biology and Bioinformatics ;2021
36. Zhu R, Wang Y, Liu J-X, Dai L-Y. Ipcarf: improving lncRNA-disease association prediction using incremental principal component analysis feature selection and a random forest classifier. BMC bioinformatics. 2021;22(1):1–17.
37. Güneş F, Wolfinger R, Tan P.-Y. Stacked ensemble models for improved prediction accuracy. In: Proc. Static Anal. Symp., 2017; pp. 1–19
38. Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks stat. 2017;1050:20.
39. Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. Advances in neural information processing systems 2017;**30**
40. Wu Z, Pan S, Chen F, Long G, Zhang C, Philip SY. A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems. 2020;32(1):4–24.
41. Shi Z, Zhang H, Jin C, Quan X, Yin Y. A representation learning model based on variational inference and graph autoencoder for predicting lncrna-disease associations. BMC Bioinform. 2021;22(1):1–20.
42. Xie G, Huang B, Sun Y, Wu C, Han Y. Rwsf-blp: a novel lncRNA-disease association prediction model using random walk-based multi-similarity fusion and bidirectional label propagation. Molecular Genetics and Genomics. 2021;296(3):473–83.
43. Fan Y, Chen M, Pan X. Gcrflda: scoring lncRNA-disease associations using graph convolution matrix completion with conditional random field. Brief Bioinform. 2022;23(1):361.
44. Bao Z, Yang Z, Huang Z, Zhou Y, Cui Q, Dong D. Lncrnadisease 2.0: an updated database of long non-coding RNA-associated diseases. Nucleic Acid Res. 2019;47(D1):1034–7.

45. Gao Y, Shang S, Guo S, Li X, Zhou H, Liu H, Sun Y, Wang J, Wang P, Zhi H, et al. Lnc2cancer 3.0: an updated resource for experimentally supported lncRNA/circRNA cancer associations and web tools based on rna-seq and scrna-seq data. Nucleic Acid Res. 2021;49(D1):1251–8.
46. Ning L, Cui T, Zheng B, Wang N, Luo J, Yang B, Du M, Cheng J, Dou Y, Wang D. Mndr v3. 0: mammal ncRNA-disease repository with increased coverage and annotation. Nucleic Acid Res. 2021;49(D1):160–4.
47. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.
48. Bo H, Fan L, Li J, Liu Z, Zhang S, Shi L, Guo C, Li X, Liao Q, Zhang W, et al. High expression of lncRNA afap1-as1 promotes the progression of colon cancer and predicts poor prognosis. J Cancer. 2018;9(24):4677.
49. Sun H, Sun X, Zhang H, Yue A, Sun M. LncRNA-pcat1 controls the growth, metastasis and drug resistance of human colon cancer cells. J BU ON Off J Balk Union Oncol. 2020;25(5):2180–5.
50. Fu Y, Li C, Luo Y, Li L, Liu J, Gui R. Silencing of long non-coding RNA miat sensitizes lung cancer cells to gefitinib by epigenetically regulating mir-34a. Front Pharmacol. 2018;9:82.
51. Zeng Z, Zhao G, Rao C, Hua G, Yang M, Miao X, Ying J, Nie L. Knockdown of lncRNA zfas1-suppressed non-small cell lung cancer progression via targeting the mir-150-5p/hmga2 signaling. J Cell Biochem. 2020;121(8–9):3814–24.
52. Jing H, Xia H, Qian M, Lv X. Long noncoding rna crnde promotes non-small cell lung cancer progression via sponging microrna-338-3p. Biomed Pharmacother. 2019;110:825–33.
53. Dong J, Su M, Chang W, Zhang K, Wu S, Xu T. Long non-coding RNAs on the stage of cervical cancer. Oncol Rep. 2017;38(4):1923–31.
54. Le Wu LJ, Zhang W, Zhang L. Medical science monitor: Roles of long non-coding RNA ccat2 in cervical cancer cell growth and apoptosis. Int Med J Exp Clin Res. 2016;22:875.
55. Huang L, Liao L-M, Liu A-W, Wu J-B, Cheng X-L, Lin J-X, Zheng M. Overexpression of long noncoding RNA hotair predicts a poor prognosis in patients with cervical cancer. Arch Gynecol Obstet. 2014;290(4):717–23.
56. Zhao L, Wang J, Li Y, Song T, Wu Y, Fang S, Bu D, Li H, Sun L, Pei D, et al. Noncodev6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. Nucleic Acids Res. 2021;49(D1):165–71.
57. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, Felix V, Jeng L, Bearer C, Lichenstein R, et al. Human disease ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res. 2019;47(D1):955–62.
58. 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. 2010;26(13):1644–50.
59. Yu G, Wang L-G, Yan G-R, He Q-Y. Dose: an r/bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics. 2015;31(4):608–9.
60. Sun J, Shi H, Wang Z, Zhang C, Liu L, Wang L, He W, Hao D, Liu S, Zhou M. Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. Mol BioSyst. 2014;10(8):2074–81.
61. Li M, Liu M, Bin Y, Xia J. Prediction of circRNA-disease associations based on inductive matrix completion. BMC Med Genom. 2020;13(5):1–13.
62. Yang Q, Li X. Bigan: LncRNA-disease association prediction based on bidirectional generative adversarial network. BMC bioinformatics. 2021;22(1):1–17.
63. Levenshtein VI, et al. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet Phys Dokl. 1966;10:707–10.
64. Zhang S, Tong H, Xu J, Maciejewski R. Graph convolutional networks: a comprehensive review. Comput Soc Netw. 2019;6(1):1–23.
65. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018; pp. 7132–7141
66. Zhuang Z, Shen X, Pan W. A simple convolutional neural network for prediction of enhancer-promoter interactions with dna sequence data. Bioinformatics. 2019;35(17):2899–906.
67. Zhao T, Hu Y, Peng J, Cheng L. Deeplgp: a novel deep learning method for prioritizing lncrna target genes. Bioinformatics. 2020;36(16):4466–72.
68. Tang Q, Nie F, Kang J, Chen W. mrnalocater: Enhance the prediction accuracy of eukaryotic mrna subcellular localization by using model fusion strategy. Mol Ther. 2021;29(8):2617–23.

Publisher's Note

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