**Self-supervised Contrastive Learning on Attribute and Topology Graphs for Predicting Relationships among lncRNAs, miRNAs and diseases**

Lan Huang1, Nan Sheng1,\*, Ling Gao1, Lei Wang1, Wenju Hou1, Jie Hong1, Yan Wang1,\*

*1* *Key laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China.*

\* Corresponding authors.

E-mail: nan\_sheng@yeah.net (Sheng N), wy6868@jlu.edu.cn (Wang Y)

# 1. Baselines

**GCLMTP** [1]: A multi-task prediction method based on self-supervised learning, designed to simultaneously extract node embeddings from lncRNA-miRNA-disease heterogeneous networks using graph contrastive learning. It predicts LDAs, MDAs and LMIs using multiple classifiers.

**GAERF** [2]: This method extracts low-dimensional representations of nodes from lncRNA-miRNA-disease heterogeneous networks using graph autoencoder. It predicts LDAs using Random Forest (RF) and has been adapted by us as a multi-task prediction model to infer MDAs and LMIs.

**LDAEXC** [3]: This method utilizes a deep autoencoder to extract low-dimensional features from integrated lncRNA and disease similarity matrix. Potential disease-associated lncRNAs are then predicted using an XGBoost classifier.

**MLGCNET** [4]: This method learns potential feature representations of nodes from lncRNA-disease heterogeneous networks using GCN. Extra Trees are then employed to calculate LDA scores.

**IPCARF** [5]: A machine learning-based computational method that using incremental principal component analysis to decrease the dimension of lncRNA and disease similarities to obtain the optimal feature subspace. RF is then performed to predict LDAs.

**LDNFSGB** [6]: This method learns optimal node features by reducing the dimensionality of integrated multiple lncRNA and disease similarity using an autoencoder. The gradient boosting algorithm is then employed to discover disease-related lncRNAs.

**SMALF** [7]: This method learns potential features of miRNAs and diseases from the original MDA matrix using stacked autoencoders. It then integrates miRNA (disease) potential features and similarity to construct node-pair feature vectors, using XGBoost to infer disease-associated miRNAs.

**PMDFI** [8]: This method extracts meaningful higher-order features from the original similarity matrix using stacked autoencoders, performs feature interactive learning, and predicts MDAs utilizing an integrated model consisting of multiple RFs and Logistic Regression (LR).

**GBDT\_LR** [9]: This method applies Gradient Boosted Decision Tree (GBDT) model to extract node-pair features and feeds them into a LR model for predicting final MDAs.

**VAEMDA** [10]: This method constructs two splicing matrices by combining integrated miRNA similarity and integrated disease similarity with known MDAs. Node features are then extracted and association scores predicted using a variational autoencoder (VAE).

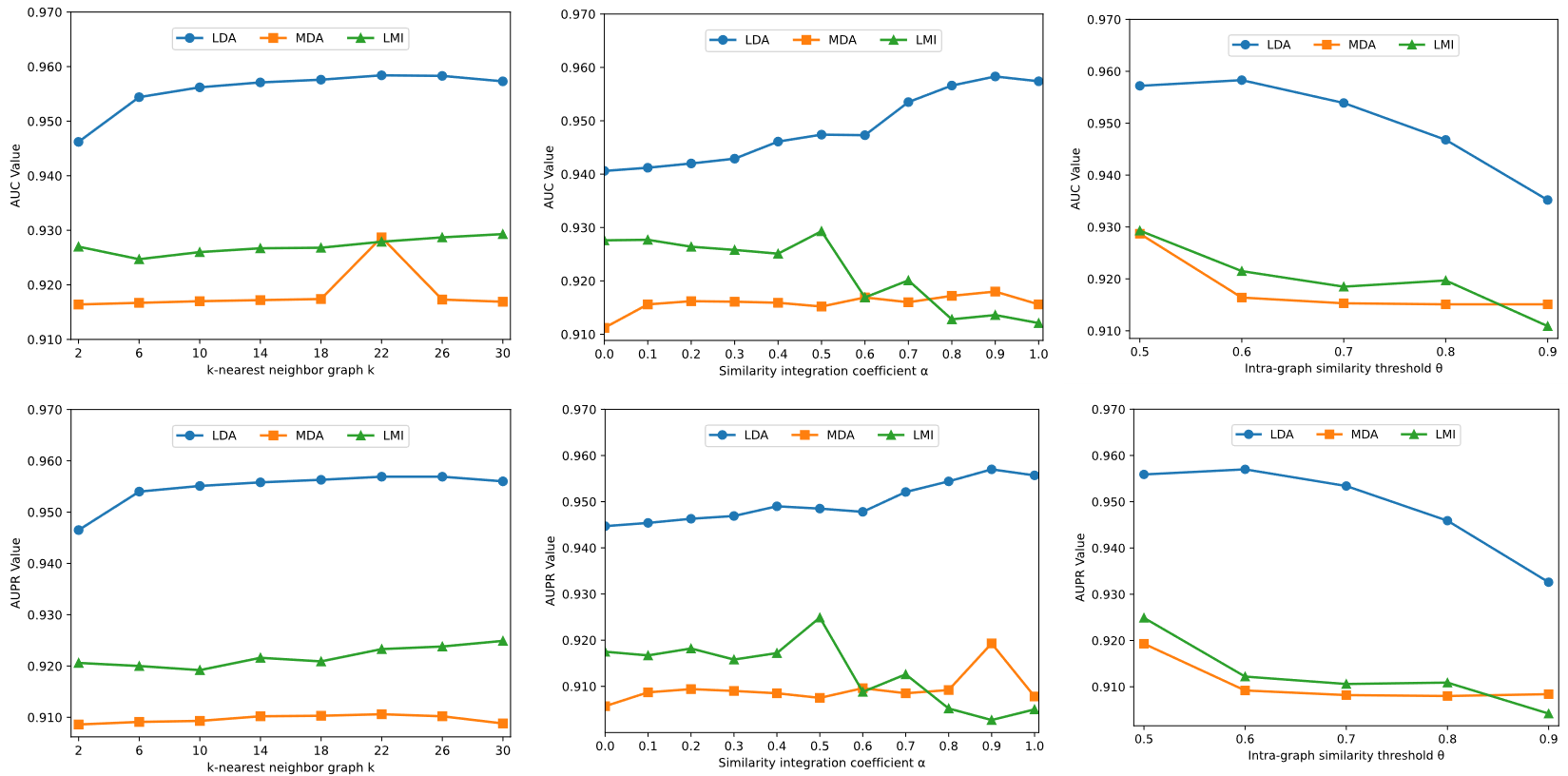
**SEBGLMA** [11]: This method extracts attribute features of RNA sequences using k-mer and word2vec. GCN is then applied to extract node embedding from lncRNA-miRNA heterogeneous networks, which are input into a rotation forest for identifying potential LMIs.

**GCLMI** [12]: An end-to-end model combining graph convolution and aoto-encoder techniques to extract features from lncRNA-miRNA heterogeneous graphs and obtain association scores.

**NDALMA** [13]**:** A computational model that leverages network distance analysis to predict potential LMIs. The method performs network distance analysis on integrated similarity networks, and obtain final scores after confidence calculation and score conversion.

# 2. Analysis of hyper-parameters

**Analysis of *k*-nearest neighbor graph .** To determine the optimal number of nearest neighbors for constructing the attribute graph, we evaluate the AUC and AUPR values of SSCLMD by varying the value of in the range . As shown in Figure 1, the AUC and AUPR increase first and then start to decrease on the LDA, MDA and LMI prediction tasks. This may be due to the fact that features are easier to smooth if the graph becomes denser, while larger may introduce more noisy edges. Given these results, we select for the attribute graph to improve the performance of SSCLMD in LDA, MDA, and LMI prediction tasks, respectively.

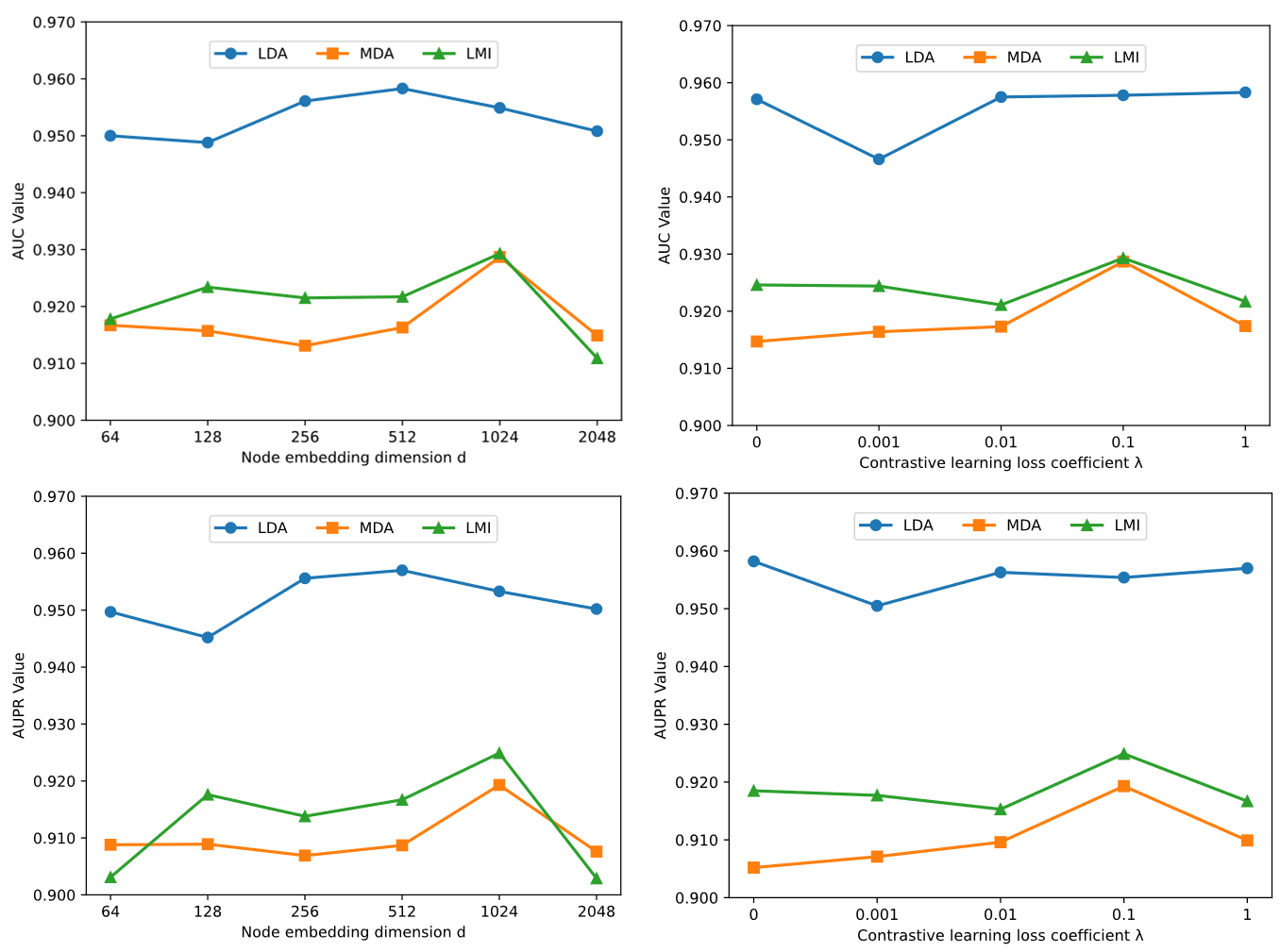


**Figure 1.** Parameter sensitivity analysis of k-nearest neighbor graph , similarity integration coefficient , and intra-graph similarity threshold on dataset 1.

**Analysis of similarity integration coefficient** . The similarity integration coefficient is adopted to balance the contribution of intra-graph node similarities derived from two sources: the disease-related lncRNA/miRNA network and the lncRNA-miRNA network. We set coefficient from 0 to 1 and report the AUC and AUPR on the three prediction tasks in Figure 1. For the LDA and MDA tasks, the performance peaks at , indicating that lncRNA and miRNA GIPK similarities based on lncRNA-disease network and miRNA-disease network are more informative. In contrast, for the LMI task, suggesting that the lncRNA-disease and miRNA-disease networks can provide complementary information to improve LMI prediction.

**Analysis of intra-graph similarity threshold** . The threshold determines the number of edges in the intra-layer, which reflects the strength of similarity between the same type nodes in the topology graph. We conduct experiments by setting them with different values ranging from 0.5 to 0.9, with results shown in Figure 1. We can find that the performance decreases slowly as increases. Optimal values are 0.5 for MDA and LMI prediction, producing the best results by retaining more edges. However, slightly improves LDA prediction, indicating some dissimilar edges may act as noise in this task.

**Analysis of node embedding dimension** . The initialize node embedding dimension is chosen from 64 to 2048. After two graph convolutional layers the embedding dimension is reduced to , . We can notice that when , the model performance is best in LDA prediction, while , the model performance is best in MDA and LMI prediction (see Figure 2). Hence, appropriately increasing the embedding dimension size can enhance prediction performance, but not always being performance gain because of model overfitting.



**Figure 2.** Parameter sensitivity analysis of node embedding dimensions , and contrastive learning loss coefficient on dataset 1.

**Analysis of contrastive learning loss weight** . We evaluate the effect of contrastive learning loss weight in Eq.18, varying it from 0 to 1. As illustrated in Figure 2, the optimal value depends on the task. For LDA prediction task, which has sparse known labels, maximizes performance. This finding underlines the effectiveness of incorporating self-supervised contrastive learning in enhancing LDA prediction results, particularly in scenarios with limited labeled data. For MDA and LMI, is best, suggesting less influence from contrastive regularization. Overall, self-supervised contrastive learning can improve prediction performance by designing an auxiliary task, especially for tasks with limited labeled data.

# 2. Supplementary Table and Figure

**Supplementary Table 1 (ST1).** Hyper-parameters on LDA, MDA and LMI prediction tasks under dataset 1 and dataset 2.

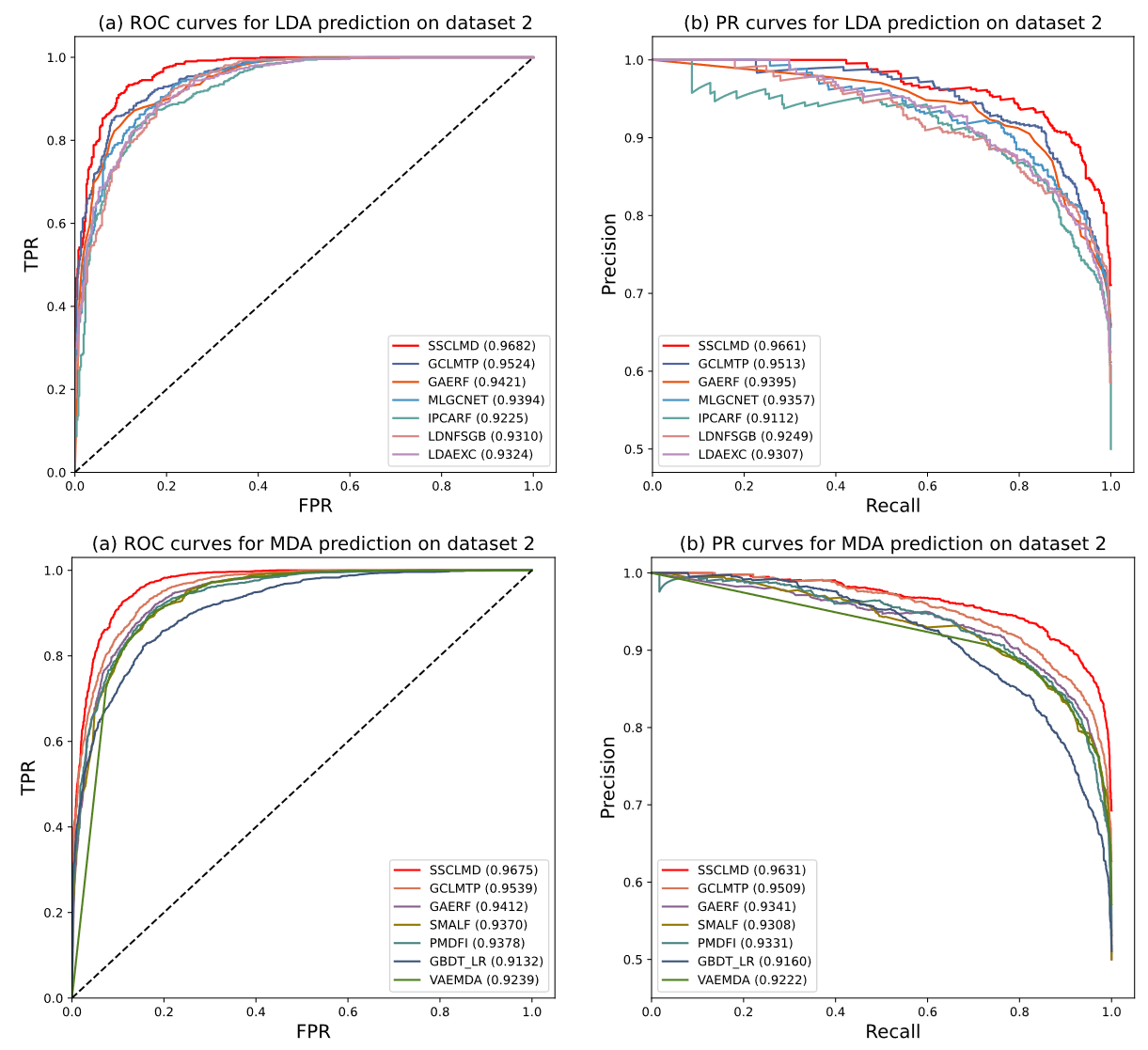
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Tasks** |  |  |  | **Nhid1** | **Nhid2** |  | **Epoch** | ***lr*** | **Dropout** |
| Dataset 1 | LDA | 22 | 0.9 | 0.6 | 256 | 128 | 1 | 80 | 0.0005 | 0.5 |
| MDA | 26 | 0.9 | 0.5 | 512 | 256 | 0.1 | 80 | 0.0005 | 0.5 |
| LMI | 30 | 0.5 | 0.5 | 512 | 256 | 0.1 | 80 | 0.0005 | 0.5 |
| Dataset 1 | LDA | 18 | 0.9 | 0.5 | 256 | 128 | 1 | 80 | 0.0005 | 0.5 |
| MDA | 26 | 0.9 | 0.7 | 512 | 256 | 0.1 | 80 | 0.0005 | 0.5 |
| LMI | 34 | 0.5 | 0.5 | 512 | 256 | 1 | 80 | 0.0005 | 0.5 |

**Supplementary Table 2 (ST2).** The top 10 predicted BC-related lncRNA and miRNA candidates on dataset 2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rank** | **LncRNA** | **Evidence** | **Rank** | **MiRNA** | **Evidence** |
| 1 | MIR17HG | PMID: 36943627 | 1 | hsa-mir-150 | HMDD v4.0,  dbDEMC v3.0 |
| 2 | BANCR | LncRNADisease v3.0,  Lnc2Cancer v3.0 | 2 | hsa-mir-15b | HMDD v4.0,  dbDEMC v3.0 |
| 3 | HULC | LncRNADisease v3.0,  Lnc2Cancer v3.0 | 3 | hsa-mir-142 | HMDD v4.0,  dbDEMC v3.0 |
| 4 | TUG1 | LncRNADisease v3.0,  Lnc2Cancer v3.0 | 4 | hsa-mir-192 | HMDD v4.0,  dbDEMC v3.0 |
| 5 | WT1-AS | LncRNADisease v3.0 | 5 | hsa-mir-181c | HMDD v4.0,  dbDEMC v3.0 |
| 6 | MIR155HG | Unconfirmed | 6 | hsa-mir-144 | HMDD v4.0,  dbDEMC v3.0 |
| 7 | TUSC7 | LncRNADisease v3.0 | 7 | hsa-mir-181d | HMDD v4.0,  dbDEMC v3.0 |
| 8 | DLEU2 | PMID: 38296962 | 8 | hsa-mir-106a | HMDD v4.0,  dbDEMC v3.0 |
| 9 | GHET1 | LncRNADisease v3.0,  Lnc2Cancer v3.0 | 9 | hsa-mir-378a | HMDD v4.0,  dbDEMC v3.0 |
| 10 | LINC01133 | LncRNADisease v3.0,  Lnc2Cancer v3.0 | 10 | hsa-mir-424 | HMDD v4.0,  dbDEMC v3.0 |

**Supplementary Table 3 (ST3).** The top 10 predicted BC-related lncRNAs and miRNA candidates on dataset 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rank** | **LncRNA** | **Evidence** | **Rank** | **MiRNA** | **Evidence** |
| 1 | FER1L4 | PMID: 31332783 | 1 | hsa-mir-211 | HMDD v4.0  dbDEMC v3.0 |
| 2 | MIR4435-2HG | PMID: 36105009 | 2 | hsa-mir-186 | HMDD v4.0  dbDEMC v3.0 |
| 3 | FTX | Unconfirmed | 3 | hsa-mir-28 | HMDD v4.0  dbDEMC v3.0 |
| 4 | MIR100HG | PMID: 33088216 | 4 | hsa-mir-19b-2 | Unconfirmed |
| 5 | LINC-PINT | PMID: 32632453 | 5 | hsa-mir-181d | HMDD v4.0  dbDEMC v3.0 |
| 6 | TUSC7 | PMID: 35296964 | 6 | hsa-mir-454 | HMDD v4.0  dbDEMC v3.0 |
| 7 | LINC00261 | PMID: 33274565 | 7 | hsa-mir-216a | HMDD v4.0  dbDEMC v3.0 |
| 8 | HNF1A-AS1 | PMID: 32319789 | 8 | hsa-mir-136 | HMDD v4.0  dbDEMC v3.0 |
| 9 | MIR17HG | PMID: 36943627 | 9 | hsa-mir-181a-1 | Unconfirmed |
| 10 | DGCR5 | PMID: 32521856 | 10 | hsa-mir-138-2 | Unconfirmed |



**Supplementary Figure 1 (SF1).** ROC and PR curves of SSCLMD and other baseline methods for LDA and MDA prediction on dataset 2.

**Reference**

[1] N. Sheng, Y. Wang, L. Huang, L. Gao, Y. Cao, X. Xie, and Y. Fu, “Multi-task prediction-based graph contrastive learning for inferring the relationship among lncRNAs, miRNAs and diseases,” Briefings in Bioinformatics, vol. 24, no. 5, pp. bbad276, 2023.

[2] Q.-W. Wu, J.-F. Xia, J.-C. Ni, and C.-H. Zheng, “GAERF: predicting lncRNA-disease associations by graph auto-encoder and random forest,” Briefings in Bioinformatics, vol. 22, no. 5, pp. bbaa391, 2021.

[3] C. Lu, and M. Xie, “LDAEXC: LncRNA-Disease Associations Prediction with Deep Autoencoder and XGBoost Classifier,” Interdisciplinary Sciences: Computational Life Sciences, vol. 15, no. 3, pp. 439-451, 2023.

[4] Q. W. Wu, R. F. Cao, J. Xia, J. C. Ni, C. H. Zheng, and Y. Su, “Extra Trees Method for Predicting LncRNA-Disease Association Based on Multi-layer Graph Embedding Aggregation,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 6, pp. 3171-3178, 2021.

[5] R. Zhu, Y. Wang, J.-X. Liu, and L.-Y. Dai, “IPCARF: improving lncRNA-disease association prediction using incremental principal component analysis feature selection and a random forest classifier,” BMC Bioinformatics, vol. 22, no. 1, pp. 1-17, 2021.

[6] Y. Zhang, F. Ye, D. Xiong, and X. Gao, “LDNFSGB: prediction of long non-coding rna and disease association using network feature similarity and gradient boosting,” BMC Bioinformatics, vol. 21, no. 1, pp. 1-27, 2020.

[7] D. Liu, Y. Huang, W. Nie, J. Zhang, and L. Deng, “SMALF: miRNA-disease associations prediction based on stacked autoencoder and XGBoost,” BMC Bioinformatics, vol. 22, no. 1, pp. 1-18, 2021.

[8] M. Tang, C. Liu, D. Liu, J. Liu, J. Liu, and L. Deng, “PMDFI: Predicting miRNA-Disease Associations Based on High-Order Feature Interaction,” Frontiers in Genetics, vol. 12, pp. 656107, 2021.

[9] S. Zhou, S. Wang, Q. Wu, R. Azim, and W. Li, “Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression,” Computational Biology and Chemistry, vol. 85, pp. 107200, 2020.

[10] L. Zhang, X. Chen, and J. Yin, “Prediction of Potential miRNA-Disease Associations Through a Novel Unsupervised Deep Learning Framework with Variational Autoencoder,” Cells, vol. 8, no. 9, pp. 1040, 2019.

[11] Z.-Y. Zhao, J. Lin, Z. Wang, J.-X. Guo, X.-K. Zhan, Y.-A. Huang, C. Shi, and W.-Z. Huang, “SEBGLMA: Semantic Embedded Bipartite Graph Network for Predicting lncRNA-miRNA Associations,” International Journal of Intelligent Systems, pp. 2785436, 2023.

[12] C. Zhao, Y. Qiu, S. Zhou, S. Liu, W. Zhang, and Y. Niu, “Graph embedding ensemble methods based on the heterogeneous network for lncRNA-miRNA interaction prediction,” BMC Genomics, vol. 21, no. 13, pp. 867, 2020.

[13] L. Zhang, P. Yang, H. Feng, Q. Zhao, and H. Liu, “Using Network Distance Analysis to Predict lncRNA-miRNA Interactions,” Interdisciplinary Sciences: Computational Life Sciences, vol. 13, no. 3, pp. 535-545, 2021.