frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 09 January 2024
DOI 10.3389/fgene.2023.1332273

OPEN ACCESS

EDITED BY
Yuan Zhou,
Peking University, China

REVIEWED BY
Ke Han,
Harbin University of Commerce, China
Jianwei Li,
Hebei University of Technology, China

*CORRESPONDENCE
Dengju Yao,
ydkvictory@hrbust.edu.cn

†These authors have contributed equally to this work

RECEIVED 02 November 2023
ACCEPTED 22 December 2023
PUBLISHED 09 January 2024

CITATION
Yao D, Zhang B, Li X, Zhan X, Zhan X and Zhang B (2024), Applying negative sample denoising and multi-view feature for lncRNA-disease association prediction.
Front. Genet. 14:1332273.
doi: 10.3389/fgene.2023.1332273

Applying negative sample denoising and multi-view feature for lncRNA-disease association prediction

Dengju Yao[1*†], Bo Zhang[1†], Xiangkui Li[1], Xiaojuan Zhan[2], Xiaorong Zhan[3] and Binbin Zhang[1]

[1]School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China, [2]College of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin, China, [3]Department of Endocrinology and Metabolism, Hospital of South University of Science and Technology, Shenzhen, China

Increasing evidence indicates that mutations and dysregulation of long non-coding RNA (lncRNA) play a crucial role in the pathogenesis and prognosis of complex human diseases. Computational methods for predicting the association between lncRNAs and diseases have gained increasing attention. However, these methods face two key challenges: obtaining reliable negative samples and incorporating lncRNA-disease association (LDA) information from multiple perspectives. This paper proposes a method called NDMLDA, which combines multi-view feature extraction, unsupervised negative sample denoising, and stacking ensemble classifier. Firstly, an unsupervised method (K-means) is used to design a negative sample denoising module to alleviate the imbalance of samples and the impact of potential noise in the negative samples on model performance. Secondly, graph attention networks are employed to extract multi-view features of both lncRNAs and diseases, thereby enhancing the learning of association information between them. Finally, lncRNA-disease association prediction is implemented through a stacking ensemble classifier. Existing research datasets are integrated to evaluate performance, and 5-fold cross-validation is conducted on this dataset. Experimental results demonstrate that NDMLDA achieves an AUC of 0.9907and an AUPR of 0.9927, with a 5-fold cross-validation variance of less than 0.1%. These results outperform the baseline methods. Additionally, case studies further illustrate the model's potential in cancer diagnosis and precision medicine implementation.

KEYWORDS

lncRNA-disease association, negative sample denoising, multi-view feature, stacking ensemble learning, graph attention networks

1 Introduction

Non-coding transcripts, particularly lncRNAs that do not encode proteins, constitute the majority of the genome (Maher, 2012). Typically, lncRNAs are transcripts that exceed 200 nucleotides in length. Noteworthy examples of lncRNAs such as H19 (Brannan et al., 1990) and Xist (Brockdorff et al., 1992) were first implicated in epigenetic regulation in the early 1990s. Numerous functional examples have also demonstrated the involvement of lncRNAs in various human physiological processes, including embryonic stem cell pluripotency, cell cycle regulation, and complex diseases (Rinn and Chang, 2012).

Therefore, exploring the relationship between lncRNAs and complex human diseases will contribute to a better understanding of disease pathogenesis and the development of lncRNA-based pharmacology.

In the past decade, extensive studies have identified many types of lncRNAs that can serve as promising biomarkers for cancer diagnosis and targeted therapy. For instance, LINC01608 has been identified as a promising prognostic biomarker for hepatocellular carcinoma (Liu et al., 2022), NALT1 promotes the targeting of PEG10 via sponge microRNA-574-5p to advance colorectal cancer progression (Ye et al., 2022), and RNA demethylase ALKBH5 promotes lung cancer progression (Shen et al., 2022). However, traditional biological experiments used to identify the association between lncRNA and diseases, such as PCR (Heid et al., 1996) and microarray analysis (Zhai et al., 2015), have always been limited by high costs and lack of specificity in exploring and understanding lncRNA.

With advances in computer technology and its ability to handle vast amounts of data, computational method has been explored to validate LDA and has yielded promising results. The first LDA prediction model (called LRLSLDA) was proposed by Chen et al. (Chen and Yan, 2013), utilizing the Laplace regularized least square method to predict LDA. This model is built on the hypothesis that similar diseases are associated with similar lncRNAs (Chen and Yan, 2013). Chen et al. (Chen et al., 2015) enhanced LRLSLDA by introducing a fusion method for lncRNA functional similarity. Although these methods did not achieve excellent prediction performance, they sparked further interest in studying the association between lncRNAs and diseases.

To capture comprehensive association information between lncRNAs and diseases, several LDA prediction methods based on similarity network feature fusion have been proposed. For example, Wei et al. proposed the iLncRNAdis-FB model for data fusion through feature blocks (Wei et al., 2021), Chen et al. proposed the iLDMSF model based on KNN for nonlinear multi-similarity fusion (Chen et al., 2021a), and Fan et al. proposed the GCRFLDA framework that integrates the conditional random field layer and the attention mechanism to fuse various similarities between lncRNAs and diseases in a linear manner as auxiliary features of nodes (Fan et al., 2022).

Moreover, Data sets in Bioinformatics usually present a high level of noise (Miranda et al., 2009). The noisy training data set increases the training time and complexity of the model. Consequently, identifying noisy instances and then eliminating or correcting them are useful techniques in data mining research (Nematzadeh et al., 2020). Chen et al. found that the presence of noisy samples can significantly impact the predictive performance of the LDA model (Chen et al., 2021b) Some papers (Yao et al., 2020; Wei et al., 2021; Kang et al., 2022; Lu and Xie, 2023) have used random sampling to create balanced datasets by including an equal number of unknown and positive samples in an attempt to mitigate the impact of unbalanced datasets. However, this approach may introduce potentially noisy data into the negative sample set. Lan et al. proposed an LDA prediction model based on an improved graph convolution network with Top-K negative sampling (Lan et al., 2021). Another method by Peng et al. involved screening reliable negative samples through a graph autoencoder (Peng et al., 2022). He et al. proposed two similarity-based negative sampling methods, one based on the Euclidean distance calculation between unlabeled samples and positive samples, and the other by reducing the number of unlabeled samples based on the functional similarity between lncRNAs (He et al., 2023).

Although existing methods have achieved good performance in predicting LDA, there still needs to be more potential in utilizing the association information between diseases and lncRNAs. Additionally, constructing the negative sample set may introduce latent LDA as noise, leading to reduced predictive accuracy of the model. This paper proposes a predictive model to construct a more accurate LDA model that combines multi-view feature extraction, an unsupervised negative sample denoising module, and a stacking ensemble classifier to uncover the associations between lncRNAs and diseases. The main contributions of this paper are as follows:

1. To mitigate the impact of sample imbalance and potential noise in negative samples on the model's performance, a negative sample denoising module is designed using an unsupervised method (K-means (Hartigan and Wong, 1979)). By simultaneously clustering positive and negative samples using K-means, this module not only improves the model's performance but also provides potential solutions for mitigating sample imbalance and achieving negative sample denoising in LDA.
2. To construct a more precise LDA model, we use graph attention networks (Veličković et al., 2017) to obtain multi-view features. These features are then combined with an unsupervised negative sample denoising module and a stacked ensemble classifier. Experimental results consistently demonstrate the outstanding performance of the proposed LDA prediction model. This model has potential applications in cancer diagnosis and can contribute to the advancement of precision medicine.

2 Materials and methods

The research flowchart of this paper can be divided into three steps, as illustrated in Figure 1 (I) data preprocessing (II) construction of the NDMLDA model by incorporating multi-view feature extraction, an unsupervised negative sample denoising module, and a stacking ensemble classifier, and (III) utilization of the NDMLDA model to make predictions regarding the association between unknown lncRNAs and diseases. Furthermore, in Figure 1 section (I), DSS (disease semantic similarity network), DCS (disease cosine similarity network), DGS (disease gaussian interaction profile kernel similarity network), LSES (lncRNA sequence similarity network), LGS (lncRNA gaussian interaction profile kernel similarity network), LFS (LncRNA functional Similarity network) represent six similarity networks, respectively.

2.1 Materials

2.1.1 Data source

The data utilized in this research were obtained from five databases: Lnc2Cancer 3.0 (Gao et al., 2021), LncRNADisease

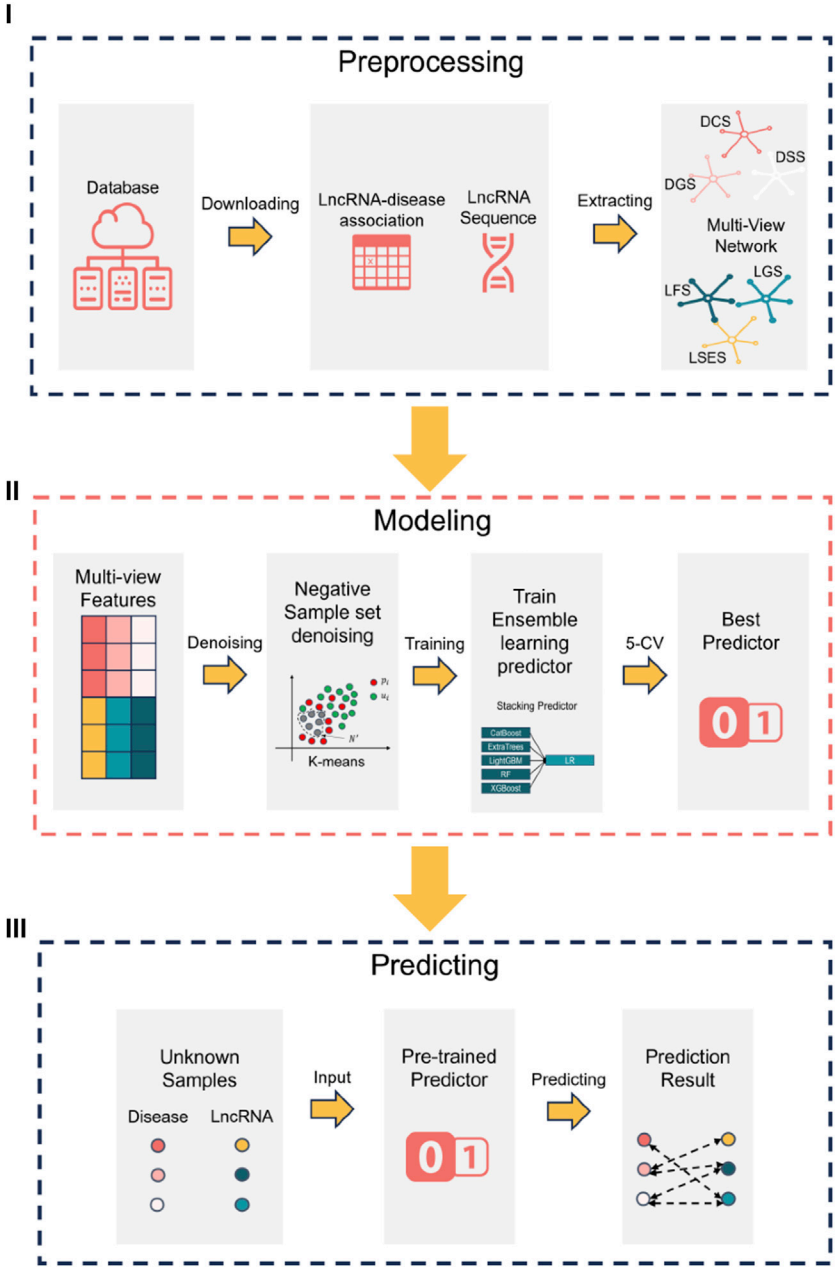

FIGURE 1
Research flowchart.

v2.0 (Bao et al., 2019), RNADisease v4.0 (Chen et al., 2022), NONCODE v6.0 (Zhao et al., 2021) and lncTarD 2.0 (Zhao et al., 2023). The Lnc2Cancer 3.0 database comprises 9,254 associations between lncRNAs and diseases, involving 2,659 lncRNAs and 216 diseases (Gao et al., 2021). LncRNADisease v2.0 collects 205,959 lncRNA-disease associations, encompassing 19,166 lncRNAs and 529 diseases (Bao et al., 2019). RNADisease v4.0 compiles 11,525 experimentally validated lncRNA-disease associations, encompassing 11,490 lncRNAs and 1,002 diseases (Chen et al., 2022). The NONCODE database includes a total of 96,411 pieces of information regarding non-coding RNA sequences (Zhao et al., 2021). The lncTarD database recruits 8,360 key lncRNA-target regulations associations with 419 disease subtypes, 1,355 lncRNAs, 506 miRNAs, 1,743 protein-coding genes and 286 biological functions.

To gain a more comprehensive understanding of the correlation between lncRNAs and diseases, we merged and manually curated the LDA data from three databases: Lnc2Cancer, LncRNADisease, and RNADisease (see supplementary for details). As a result, we obtained a total of 8,334 lncRNA-disease associations involving 629 lncRNAs and 511 diseases, which were stored in matrix A. Subsequently, we retrieved the sequence information of all lncRNAs in matrix A from the NONCODE database. Additionally, we applied the same preprocessing method to process the data from lncTarD, resulting in 504 lncRNA-disease associations between 103 diseases

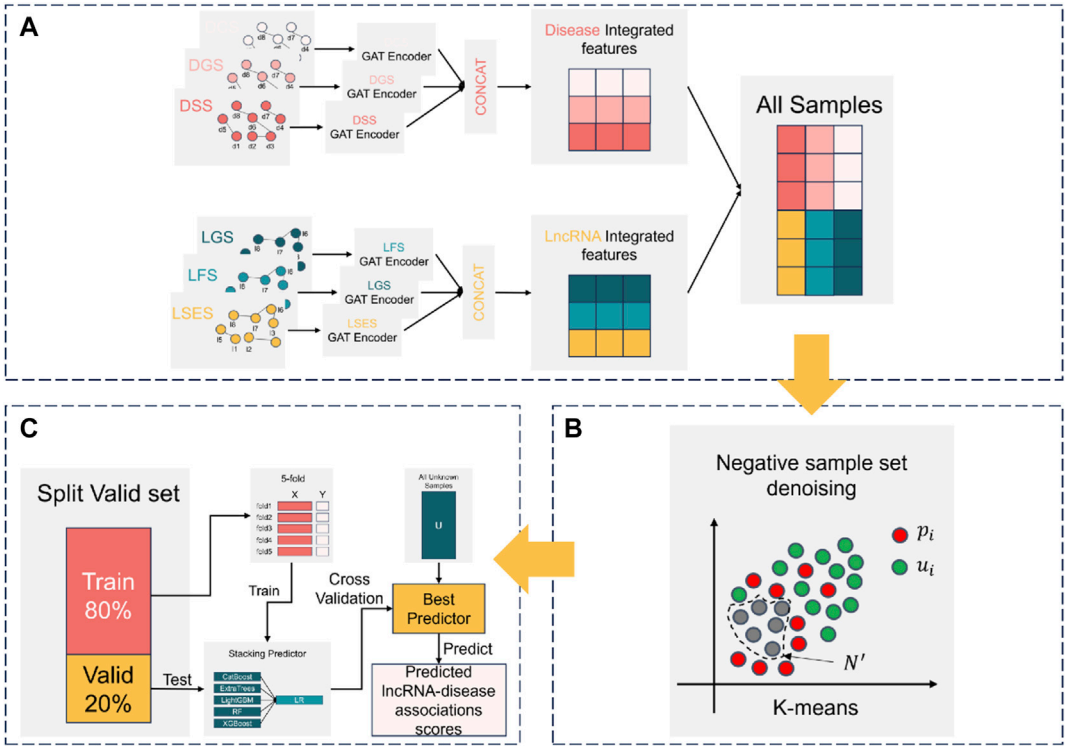

FIGURE 2
The construction process of NDMLDA comprises three parts: **(A)** multi-view feature extraction; **(B)** negative sample set denoising; **(C)** training and prediction of the stacking ensemble classifier.

and 212 lncRNAs. No further data manipulation was performed besides this.

2.1.2 Disease semantics similarity

We use the method proposed by Wang et al. (2010) to calculate the semantic similarity of diseases which is given by the following formula:

$$DSS(d_i, d_j) = \frac{\sum_{d \in D_i \cap D_j} (SC_{d_i}(d) + SC_{d_j}(d))}{SV_{d_i} + SV_{d_j}}$$

Where, d represents disease; D represents ancestors' nodes of d; $\cap$ represents intersection; SC and SV represent the semantic contribution value and semantic value of disease, respectively.

2.1.3 Disease cosine similarity

The cosine similarity between two diseases can be calculated using the following formula:

$$DCS(d_i, d_j) = \frac{A(i,:) \cdot A(j,:)}{\|A(i,:)\| \times \|A(j,:)\|}$$

Where, vector $A(i,:)$ represents the set of elements in the ith row in matrix A. The length of this vector is denoted as $\|A(i,:)\|$.

2.1.4 Disease (lncRNA) Gaussian interaction profile kernel similarity

We utilize the algorithm presented by van Laarhoven et al. (2011) to calculate the similarity of gaussian interaction profile kernel similarity for disease (lncRNA), which is given by the following formulas:

$$DGS = exp(-\gamma_d \|A(i,:) - A(j,:)\|^2)$$
$$LGS = exp(-\gamma_l \|A(:,i) - A(:,j)\|^2)$$

Where, DGS and LGS represents disease (lncRNA) gaussian interaction profile kernel similarity; γ represents the normalized kernel bandwidth.

2.1.5 LncRNA functional similarity

We adopt the method proposed by Sun et al. (2014) to calculate the functional similarity of lncRNAs(LFS). The formula is as follows:

$$LFS(l_i, l_j) = \frac{\sum_{1 \le i \le nD_i} SS(d_i, D_i) + \sum_{1 \le i \le nD_j} SS(d_j, D_j)}{nD_i + nD_j}$$

Where d represents a disease associated with a lncRNA l; D represents a group of diseases associated with l and nD represents the total number of diseases in this group; $SS(d, D)$ represents the maximum semantic similarity between d and D.

2.1.6 LncRNA sequence similarity

We are inspired by Li et al. (2020) to introduce lncRNA sequence similarity (LSES), which is calculated using the following formula:

$$LSES(l_i, l_j) = \frac{cost(l_i, l_j)}{len(l_i) + len(l_j)}$$

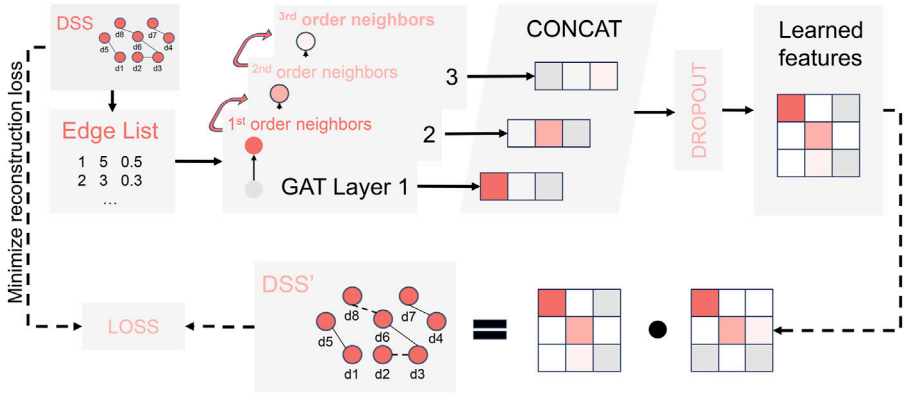

FIGURE 3
The encoding process of a specific input network.

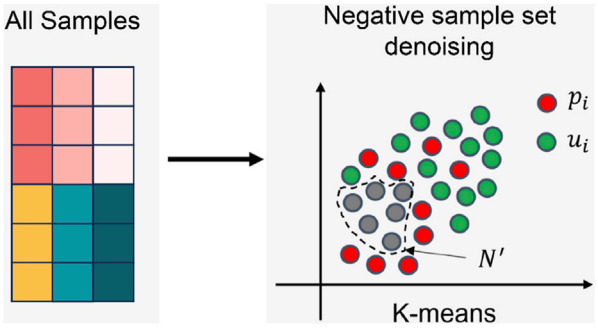

FIGURE 4
The denoising process of negative sample set.

Where $len(l)$ represents the length of the sequence l; $cost(l_i, l_j)$ is used to measure the minimum cost required to transform the sequence of l_i into the sequence of l_j by performing three types of operations: insertion, deletion, and replacement (with insertion or deletion cost being 1, and replacement cost being 2).

2.2 Methods

The construction process of NDMLDA is shown in Figure 2, which mainly consists of three steps (A) multi-view feature extraction; (B) negative sample set denoising; (C) training and prediction of the stacking ensemble classifier.

2.2.1 Multi-view feature extraction

The Graph Attention Network (GAT) has demonstrated significant potential in predicting LDA as a primary approach for multi-view feature extraction (Shi et al., 2021; Liang et al., 2022; Zhao et al., 2022) Following the guidance of previous literature (Forster et al., 2022), we developed a GAT-based module for multi-view feature extraction, as depicted in Figure 3. To begin with, we transformed the similarity networks of various views (including DSS, DCS, DGS, LFS, LGS, and LSES) into edge list format, where each row represents the source, target, and weight. Subsequently, for each input network, we constructed an encoder by concatenating three GAT layers, as illustrated in Figure 3. This encoding process enables the learning of high-order neighborhood features for the specific input networks using dedicated encoders. Next, by employing feature aggregation and random loss processing, a unified disease (or lncRNA) feature H is generated. Finally, all node features were arranged in a matrix F. Through multiple experiments, we fixed the length of this feature to 64 (see supplementary material for details).

$$GAT(A, H) = \sigma(\alpha H W^T)$$

Where α represents the attention coefficient, H represents the features of nodes in A, W represents the trainable weight parameters, T represents the transpose operation, and σ represents the non-linear activation function LeakyReLU (Maas et al., 2013).

To enhance the quality of feature extraction, we decode and reconstruct the unified feature matrix F, which has learned the lncRNA (or disease). Our objective is to minimize the discrepancy between the reconstructed network F^T and the original input network. The process of network reconstruction after decoding is exemplified below.

$$A = F \cdot F^T$$

The loss function in this process can be defined as follows:

$$Loss = \frac{1}{n^2}\sum_{j}^{N}\left\| b_j \odot \left(A - A_j\right) \odot b_j^T \right\|_F^2$$

Where, n represents the total number of nodes in the input network, b_j represents the node mask in input network j, A_j represents the adjacency matrix corresponding to input network j, $\odot$ represents the inner product and $\|\cdot\|_F$ represents the F-norm.

2.2.2 Negative sample set denoising

The process of the negative sample set denoising is shown in Figure 4 Firstly, the positions of elements with values 1 and 0 in the LDA matrix are recorded separately. Then, the unified feature vectors of corresponding diseases and lncRNAs are retrieved based on these positions. These two feature vectors are directly

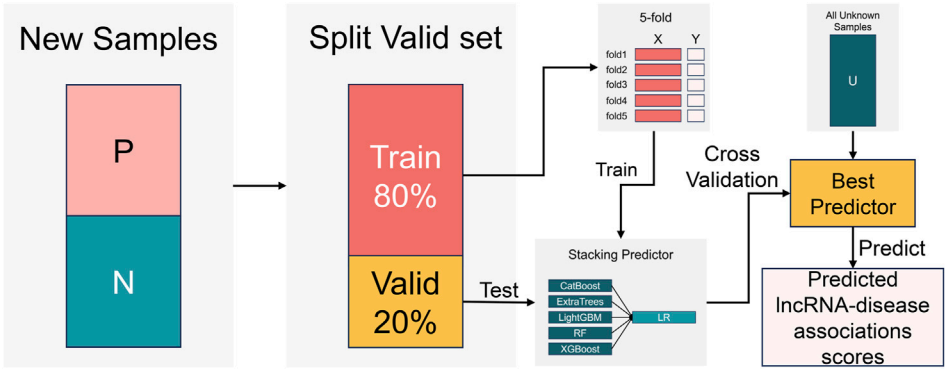

FIGURE 5
The training process of a stacking ensemble classifier.

concatenated, with diseases preceding lncRNAs, to form a sample. The complete sample set (All Samples) is obtained by concatenating the features of all positions. Next, the number of clusters K is determined by calculating the silhouette coefficient. The silhouette coefficient (SC), which ranges from −1 to 1, is a commonly used indicator in previous studies for evaluating the effectiveness of clustering algorithms (Rousseeuw, 1987).

SC usually follows the trend of K-value changes. When the silhouette coefficient approaches 1, the K-value also approaches the ideal value. SC can be calculated as follows:

$$SC_i = \frac{C_b(i) - C_a(i)}{maximum(C_a(i), C_b(i))}$$

Where, $C_a(i)$ represents the average distance between sample i and the other samples in its cluster, while $C_b(i)$ represents the minimum average distance between sample i and the samples in different clusters. In this study, we set K as 3.

We used the K-means algorithm (Hartigan and Wong, 1979) to perform 10 rounds of clustering on the entire sample set. The complete description of the negative sample denoising process is as follows:

Let P represent the known positive sample set, $P = \{p_1, p_2, \ldots, p_m\}$, where each sample p_i represents a known lncRNA-disease association. Let U represent the unknown sample set, $U = \{u_1, u_2, \ldots, u_{n-m}\}$. Assuming that the samples in U that are similar to P are noise samples, we take the following steps to denoise U:

First, we cluster the entire sample set using the K-means algorithm, which results in cluster divisions $C = \{C_1, C_2, \ldots, C_k\}$, where each cluster C_i is a set. For each cluster C_i, we calculate the proportion of positive samples and denote it as $r(C_i)$.

Then, we repeat the following steps 10 times:

1. Cluster the sample set using the K-means algorithm to obtain cluster divisions $C' = \{C_1', C_2', \ldots, C_k'\}$.
2. For each cluster C_i', calculate the proportion of positive samples and denote it as $r'(C_i')$.
3. Find the cluster C_i' with the highest $r'(C_i')$ and denote its unknown sample set as U'.
4. Save U'.

Finally, we take the intersection of the noise sample sets obtained from these 10 clustering iterations, $U_{noise} = U_1' \cap U_2' \cap \ldots \cap U_{10}'$ and remove these samples from U. The final denoised unknown sample set is represented as $U_{reliable} = U - U_{noise}$. The unknown samples in $U_{reliable}$ represent the denoised negative samples.

2.2.3 Training stacking ensemble classifier

To overcome the limited predictive capabilities of individual classifier, we draw inspiration from previous research (Li et al., 2021; Liang et al., 2022). The training process of the stacking ensemble classifier is illustrated in Figure 5. Five decision tree-based classifiers, including CatBoost (Dorogush et al., 2018), ExtraTrees (Geurts et al., 2006), LightGBM (Ke et al., 2017), RandomForest (Breiman, 2001), and XGBoost (Chen and Guestrin, 2016), are employed as base classifier, with LogisticsRegression (Cramer, 2002) serving as the meta-classifier. This framework creates a stacked ensemble LDA prediction model (refer to the supplementary material for the training process of the ensemble classifier). We conduct a five-fold cross-validation on 80% of the samples from the reconstructed new dataset (details can be found in the supplementary material), while the remaining 20% of samples are used as an independent dataset to evaluate the trained classifiers. Finally, we select the classifier with the best performance for the final LDA prediction.

3 Results

3.1 Experimental settings

The performance evaluation of NDMLDA is conducted using five performance metrics: accuracy (ACC), Matthew's correlation coefficient (MCC) (Harald, 1946), F1-score, area under the receiver operating characteristic curve (AUC), and area under the precision-recall curve (AUPR). The calculation formulas for these metrics are as follows:

$$ACC = \frac{TN + TP}{TN + TP + FN + FP}$$

TABLE 1 Comparison of the performance of NDMLDA with other LDA prediction methods.

Model	AUC	AUPR	MCC	F1	ACC
NDMLDA	0.9907 ± 5.2e-8	0.9927 ± 2.2e-8	0.9249 ± 9.6e-6	0.9631 ± 2.3e-6	0.9624 ± 2.4e-6
NDMLDA*[1]	0.9683 ± 2.9e-7	0.9718 ± 2.7e-7	0.8229 ± 2.2e-5	0.9135 ± 4.9e-6	0.9114 ± 5.3e-6
MAGCNSE	0.9665 ± 4.8e-6	0.9773 ± 1.26e-5	0.8729 ± 9.4e-5	0.9462 ± 1.4e-4	0.9357 ± 1.2e-7
VGAELDA	0.9212 ± 3.1e-4	0.7469 ± 1.2e-4	0.6872 ± 5.3e-4	0.6514 ± 8.9e-4	0.9806 ± 1.4e-6
CapsNet-LDA	0.9634 ± 1.5e-5	0.7452 ± 1.2e-5	0.6764 ± 9.9e-5	0.6843 ± 8.2e-5	0.9836 ± 9.5e-7
LDAformer	0.9452 ± 9.3e-4	0.2439 ± 1.2e-2	0.1844 ± 5.5e-3	0.1034 ± 1.2e-3	0.9403 ± 7.4e-4
SSMF-BLNP	0.8251 ± 1.1e-5	0.1535 ± 3e-5	0.1815 ± 2e-5	0.4339 ± 2e-5	0.9255 ± 1.4e-5

[1]NDMLDA*
Stands for negative sample denoising not being executed.

TABLE 2 Comparison of the performance of NDMLDA with other LDA prediction methods on lncTarD dataset.

Model	AUC	AUPR	MCC	F1	ACC
NDMLDA	0.9479 ± 1.5e-4	0.9635 ± 1.6e-4	0.7852±1e-3	0.8917 ± 4.7e-4	0.8928 ± 2.4e-4
MAGCNSE	0.9492 ± 1.9e-5	0.7314 ± 9.36e-4	0.6439 ± 8.6e-4	0.6279 ± 1e-3	0.9857 ± 9.7e-7
VGAELDA	0.9337 ± 3.9e-4	0.7779 ± 1.5e-3	0.7825 ± 9.6e-4	0.7637 ± 1.3e-3	0.9903 ± 1.5e-6
CapsNet-LDA	0.9058 ± 3e-4	0.7367 ± 1.2e-3	0.7312 ± 1e-3	0.7262 ± 1e-3	0.9896 ± 9.5e-7
LDAformer	0.8115 ± 1.1e-3	0.2574 ± 5e-4	0.1937 ± 7.5e-5	0.0415 ± 2.9e-5	0.8527 ± 3e-3
SSMF-BLNP	0.9303 ± 1.1e-5	0.5738 ± 1e-4	0.3791 ± 3.1e-5	0.7936 ± 5.3e-5	0.943 ± 2.4e-4

TABLE 3 The performance comparison between individual classifiers and stacked ensemble classifiers.

Classifier	AUC	AUPR	MCC	F1	ACC
Stacking	0.9907 ± 5.2e-8	0.9927 ± 2.2e-8	0.9249 ± 9.6e-6	0.9631 ± 2.3e-6	0.9624 ± 2.4e-6
LogisticsRegression	0.9825 ± 9.6e-9	0.9845 ± 1.2e-8	0.8721 ± 1.3e-	0.9372 ± 3.9e-7	0.9360 ± 3.5e-7
RandomForest	0.9890 ± 3.3e-7	0.9914 ± 1.1e-7	0.9178 ± 5.9e-6	0.9597 ± 1.4e-6	0.9590 ± 1.5e-6
ExtraTrees	0.9891 ± 3.3e-7	0.9912 ± 1.9e-7	0.9251 ± 6.9e-6	0.9579 ± 1.4e-6	0.9625 ± 1.7e-6
XGB	0.9904 ± 2.4e-7	0.9924 ± 1.2e-7	0.9143 ± 9.7e-6	0.9582 ± 2.2e-6	0.9571 ± 2.4e-6
LGBM	0.9907 ± 1.4e-7	0.9924 ± 5.4e-8	0.9096 ± 1.2e-5	0.9559 ± 2.9e-6	0.9548 ± 3.1e-6
MLP	0.9887 ± 1.5e-6	0.9907 ± 1.5e-6	0.9006 ± 9.9e-5	0.9536 ± 1.8e-5	0.9503 ± 2.4e-5
SVM	0.9855 ± 5.4e-8	0.9882 ± 1.8e-8	0.8914 ± 6.2e-7	0.9465 ± 1.7e-7	0.9456 ± 1.7e-7

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

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

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

$$Recall = \frac{TP}{TP + FN}$$

In the context of the confusion matrix, TP, TN, FP, and FN are variables that represent the four different types of prediction situations.

3.2 Comparison results with other methods

We conducted a comparative analysis of several LDA prediction methods, including MAGCNSE (Liang et al., 2022), MCHNLDA (Zhao et al., 2022), VGAELDA (Shi et al., 2021), CapsNet-LDA (Zhang et al., 2022), LDAformer (Zhou et al., 2022), and SSMF-BLNP (Xie et al., 2023).

MAGCNSE (Liang et al., 2022) employs a two-step approach, first utilizing GCN to extract the multi-view representation of lncRNA and diseases, and then employing CNN to obtain the final representation. The integrated classifier is then used for prediction (Zhao et al., 2022).

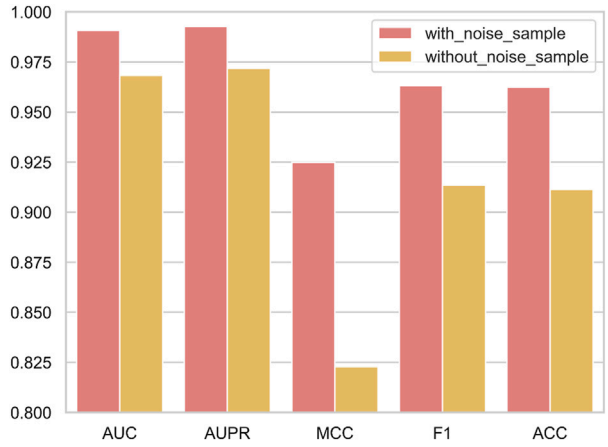

FIGURE 6
The influence of negative sample set denoising module on the predictive performance of NDMLDA.

VGAELDA (Shi et al., 2021) proposes a LDA prediction method that combines variational inference and graph autoencoders.

CapsNet-LDA (Zhang et al., 2022) presents a prediction method that leverages capsule networks and stacked autoencoders.

LDAformer (Zhou et al., 2022) introduces a LDA prediction method based on topological feature extraction and Transformer encoding.

As shown in Table 1, NDMLDA achieved higher AUC and AUPR by 2.5% and 1.6%, respectively, compared to the second-best MAGCNSE. Furthermore, the overall performance of NDMLDA (with all metrics above 92%) is superior to other comparative methods.

MAGCNSE and CapsNet-LDA mitigate the impact of sparse features on the model through a multi-view approach, achieving good performance (overall performance higher than 0.8). However, they are affected by negative sample noise, resulting in suboptimal performance. Additionally, as shown in Table 1, our model, despite having a decrease in performance in five evaluation metrics without sample reconstruction, still outperforms methods such as SSMF-BLNP and CapsNet-LDA. This indicates that our negative sample denoising module is effective in mitigating the impact of negative sample noise on the model.

LDAformer proposed a method for LDA prediction based on topological feature extraction and Transformer encoder. By enhancing feature extraction, the performance of complex models is improved. Compared to our method, without using the sample denoising module, we obtain multi-view features through GAT and achieve better overall performance in LDA using a simple stacking model. This indicates that our multi-view feature processing method is effective.

Meanwhile, to further demonstrate the generalization ability of our method, we conducted comparative experiments on an independent dataset lncTarD. The experimental results are shown in Table 2. It can be observed that our proposed method still outperforms the comparative methods in four main indicators, indicating the robustness of NDMLDA.

3.3 Ablation studies

3.3.1 The influence of negative sample set denoising on the predictive performance of NDMLDA

When the negative sample set denoising module is integrated into NDMLDA (as depicted in Figure 6), all five performance measures exhibit superior results compared to the state without the module. Notably, the addition of the module improves the AUC by 2.3%, AUPR by 2.2%, MCC by 12.4%, F1-score by 5.4% and ACC by 5.6%. These findings suggest that incorporating the negative sample set denoising module enhances the prediction performance of NDMLDA. We visualized the distribution of samples before and after denoising using t-SNE (Maaten and Hinton, 2008). Figure 7 shows the visualization results. Comparing Figures 7A,C, it can be observed that our proposed method for denoising the negative sample set successfully removes the noisy samples.

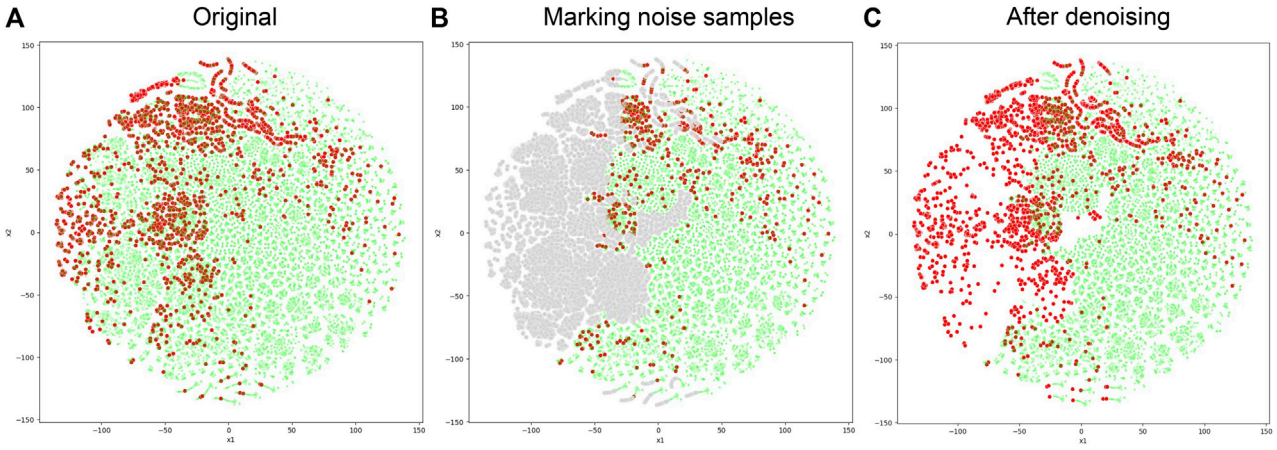

FIGURE 7
Comparison of sample distribution before and after negative sample denoising. **(A)** represents the original distribution of samples, with red dots indicating positive samples and green dots indicating unknown samples; **(B)** uses gray dots to represent noisy samples in the unknown samples; **(C)** represents the distribution of samples after denoising.

TABLE 4 Top 30 lncRNAs related to breast cancer predicted by NDMLDA.

No.	LncRNA	Evidence	No.	LncRNA	Evidence
1	TUNAR	P&R&D	16	SNHG3	P*&R&C&D
2	SLC26A4-AS1	P&R&D	17	DLX6-AS1	P*&R&C&D
3	LINC00665	P*&R&C&D	18	LINC00319	P&R
4	HAR1B	P&R&D	19	NEAT1	P*&R&C&D
5	SNHG15	P*&R&C&D	20	MCM3AP-AS1	P&R&D
6	KCNQ1OT1	P*&R&C&D	21	DLEU1	P&R&C&D
7	DPP10-AS1	P&R&D	22	HMMR-AS1	P&R&C&D
8	LINC00461	P*&R&C&D	23	SNHG7	P*&R&C&D
9	FEZF1-AS1	P*&R&C	24	LINC00339	P&R&D
10	HOXA11-AS	P*&R&C&D	25	MIR7-3HG	P&R&D
11	SNHG4	P&R&D	26	ZFAS1	P*&R&C&D
12	FAS-AS1	P*&R&C	27	OIP5-AS1	P*&R&C&D
13	FENDRR	P&R&C&D	28	GNG12-AS1	P&R&C&D
14	ST8SIA6-AS1	P*&R&C&D	29	LINC01234	P*&R&D
15	FOXC2-AS1	P&R&C&D	30	PAX8-AS1	P&R&D

We systematically validated the top 30 lncRNAs, associated with each specific type of cancer by cross-referencing three important databases: LncRNADisease v2.0, Lnc2Cancer v3.0, and RNADisease v4.0 (Bao et al., 2019; Gao et al., 2021; Chen et al., 2022), as well as consulting relevant literature records.

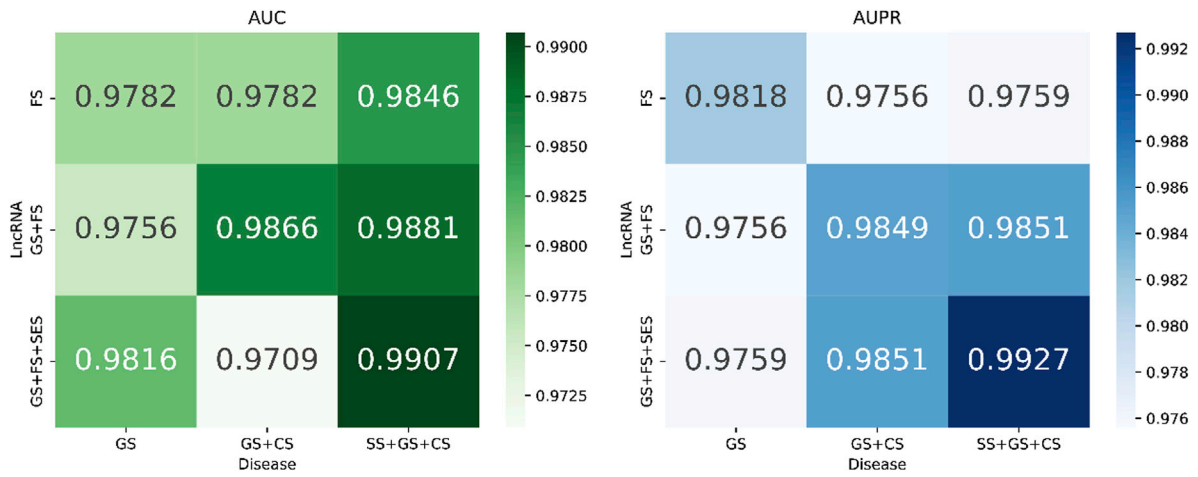

FIGURE 8
AUC and AUPR corresponding to different combinations of views.

3.3.2 Classifier selection

Table 3 demonstrates that among the five metrics, the stacked ensemble classifier attained optimal results for three of them. While the stacked ensemble classifier's performance in terms of MCC and ACC is slightly lower than that of ExtraTrees (with a maximum difference of 0.02%), it surpasses ExtraTrees in the more significant evaluation metrics of AUC and AUPR (with improvements of 0.16% and 0.15% respectively). These results indicate that the inclusion of the stacked ensemble classifier can enhance the predictive performance of NDMLDA.

3.3.3 Combination of different views

According to Figure 8, the performance of the model is influenced by the combination of different views (AUC: 0.9709–0.9907; AUPR: 0.9818–0.9927). Furthermore, increasing the number of combined views leads to an improvement in the model's performance. To construct a more precise LDA prediction model, we have chosen to utilize fusion features from lncRNA, which include lncRNA gaussian interaction profile kernel similarity (LGS), lncRNA functional similarity (LFS), and lncRNA sequence similarity (LSES), as well as fusion features from diseases, which

TABLE 5 Top 30 lncRNAs related to cervical cancer predicted by NDMLDA.

No.	LncRNA	Evidence	No.	LncRNA	Evidence
1	SNHG8	P&R&C&D	16	PCGEM1	P&R&C&D
2	LINC00665	P&R	17	LINC01503	P*&R&C&D
3	TDRG1	P*&R&C&D	18	ZFAS1	P*&R&C&D
4	GAS5-AS1	P&R&C&D	19	OIP5-AS1	P*&R&C&D
5	FEZF1-AS1	P*&R&C	20	PAX8-AS1	P*&R&D
6	HOXA11-AS	P*&R&C&D	21	BDNF-AS	P&R
7	SNHG4	P&R&C&D	22	LINC01139	P&R&C&D
8	FENDRR	P&R&D	23	MIR22HG	P*&R&C&D
9	SNHG3	P&R&C&D	24	TUSC8	P*&R&C& D
10	DLG1-AS1	P&R&C&D	25	FOXD2-AS1	P*&R&C&D
11	DLX6-AS1	P* & R & C &D	26	CYTOR	P* & R & D
12	LINC00319	P&R&C&D	27	MALAT1	P*&R&C&D
13	NEAT1	P*&R&C&D	28	PCAT6	P*&R&C&D
14	DLEU1	P&R&C&D	29	SOX2-OT	P*&R&D
15	SNHG7	P*&R&C&D	30	PVT1	P*&R&C&D

include disease semantic similarity (DSS), disease gaussian interaction profile kernel similarity (DGS), and disease cosine similarity (DCS).

3.4 Case studies

To further validate the performance of NDMLDA in predicting the association between specific diseases and lncRNA, we conducted case studies on six prevalent cancers: breast cancer, cervical cancer, colon cancer, esophageal cancer, lung cancer, and stomach cancer. In each case study, we utilized all samples related to cancer as the testing set, while the remaining samples served as the training set. Subsequently, we trained NDMLDA on the training set and employed it to evaluate the samples in the testing set.

The validated lncRNAs related to breast cancer and cervical cancer are summarized in Table 4 and Table 5, respectively. In the evidence column, "C" denotes candidate lncRNAs corroborated by the Lnc2Cancer database. "D" denotes candidate lncRNAs supported by the LncRNADisease database. "P" denotes candidate lncRNAs supported by a single literature source. "R" denotes candidate lncRNAs corroborated by the RNADisease database. "P*" denotes candidate lncRNAs supported by multiple published literature sources. Further details regarding the predictions of NDMLDA for lncRNAs associated with four other cancers can be found in the supplementary materials.

4 Discussion

The NDMLDA method utilizes the negative sample denoising module to obtain negative sample data that closely approximates the real distribution. Instead of introducing a new clustering method, our approach focuses on integrating the multi-view similarity network with the negative sample denoising technique. To achieve this, we adopt the K-means algorithm, a well-established clustering algorithm, as the core algorithm for negative sample denoising.

The NDMLDA model demonstrates good performance by combining stacked classifiers. However, we have also noticed that several single classifiers used for comparison have AUC and AUPR values around 0.99. On one hand, this is because we balanced the positive and negative samples during classifier evaluation. On the other hand, it is due to the relatively small number of known lncRNA-disease associations, which results in an insufficient number of samples for performance evaluation. However, considering the increasing complexity of data in future model applications, we have chosen the stacked ensemble classifier as our final classifier to ensure the competitiveness of our model.

However, our proposed model (NDMLDA) still has some limitations. Although we obtained a large number of known LDAs (8,334) by merging multiple databases, the comparison with the huge number of unknown samples (313,085) is still very sparse. At the same time, the dataset only includes a limited number of lncRNA-disease pairs, which is only a small fraction of the real-world scenarios. Therefore, in the future, we will attempt to further expand the number of LDAs in the dataset to address the constantly changing real situations. We also recognize that there is still a possibility that some reliable negative samples may be discarded in the process. To mitigate this, we plan to conduct further research and improvements in our future work.

LncRNAs have been established as pivotal regulators of gene expression, playing a significant role in a wide range of biological functions and disease processes, including cancer. This study presents a model known as NDMLDA, which integrates multi-view feature extraction, unsupervised negative sample denoising, and stacked ensemble classifier. The experimental results demonstrate that the proposed prediction method achieves exceptional performance across five metrics (including AUC, AUPR, MCC, F1-score and ACC). Additionally, the accuracy and reliability of NDMLDA in the prediction process for LDA are further substantiated through six case studies (involving breast cancer, cervical cancer, colon cancer, esophageal cancer, lung cancer, and gastric cancer).

5 Conclusion

This article introduces an LDA prediction model (NDMLDA) that combines negative sample denoising and multi-view network feature extraction. The experimental results demonstrate that our method outperforms the six recent base models, achieving excellent performance in five metrics (including AUC, AUPR, and MCC). Additionally, the results of six case studies (breast cancer, cervical cancer, colon cancer, esophageal cancer, lung cancer, and gastric cancer) further validate the accuracy and reliability of NDMLDA in LDA prediction tasks.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

DY: Supervision, Writing–review and editing. BoZ: Writing–original draft. XL: Data curation, Writing–review and editing. XZ: Investigation, Writing–review and editing. XZ: Investigation, Writing–review and editing. BiZ: Data curation, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the National Natural Science Foundation of China (No. 62172128).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., and Dong, D. (2019). LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res.* 47, D1034–D1037. doi:10.1093/nar/gky905

Brannan, C. I., Dees, E. C., Ingram, R. S., and Tilghman, S. M. (1990). The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.* 10, 28–36. doi:10.1128/mcb.10.1.28-36.1990

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/A:1010933404324

Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., et al. (1992). The product of the mouse xist gene is a 15 Kb inactive X-specific transcript containing No conserved ORF and located in the nucleus. *Cell* 71, 515–526. doi:10.1016/0092-8674(92)90519-i

Chen, J., Lin, J., Hu, Y., Ye, M., Yao, L., Wu, L., et al. (2022). RNADisease v4.0: an updated resource of RNA-associated diseases, providing RNA-disease analysis, enrichment and prediction. *Nucleic Acids Res.* 51, D1397–D1404. doi:10.1093/nar/gkac814

Chen, Q., Lai, D., Lan, W., Wu, X., Chen, B., Liu, J., et al. (2021a). ILDMSF: inferring associations between long non-coding RNA and disease based on multi-similarity fusion. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18, 1106–1112. doi:10.1109/TCBB.2019.2936476

Chen, T., and Guestrin, C. (2016). "XG boost: a scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 16, Association for Computing Machinery (ACM), New York, NY, USA, August 13 2016, 785–794.

Chen, X., Yan, C. C., Luo, C., Ji, W., Zhang, Y., and Dai, Q. (2015). Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci. Rep.* 5, 11338. doi:10.1038/srep11338

Chen, X., and Yan, G.-Y. (2013). Novel human lncrna–disease association inference based on lncrna expression profiles. *Bioinformatics* 29, 2617–2624. doi:10.1093/bioinformatics/btt426

Chen, X.-J., Hua, X.-Y., and Jiang, Z.-R. (2021b). ANMDA: anti-noise based computational model for predicting potential miRNA-disease associations. *BMC Bioinforma.* 22, 358. doi:10.1186/s12859-021-04266-6

Cramer, J. S. (2002). *The origins of logistic regression by J.S. Cramer: ssrn*. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=360300 (Accessed October 29, 2023).

Dorogush, A. V., Ershov, V., and Gulin, A. (2018). *CatBoost: gradient boosting with categorical features support*.

Fan, Y., Chen, M., and Pan, X. (2022). GCRFLDA: scoring lncrna-disease associations using graph convolution matrix completion with conditional random field. *Brief. Bioinform.* 23, bbab361. doi:10.1093/bib/bbab361

Forster, D. T., Li, S. C., Yashiroda, Y., Yoshimura, M., Li, Z., Isuhuaylas, L. A. V., et al. (2022). BIONIC: biological network integration using convolutions. *Nat. Methods* 19, 1250–1261. doi:10.1038/s41592-022-01616-x

Gao, Y., Shang, S., Guo, S., Li, X., Zhou, H., Liu, H., et al. (2021). 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 Acids Res.* 49, D1251–D1258. doi:10.1093/nar/gkaa1006

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.* 63, 3–42. doi:10.1007/s10994-006-6226-1

Harald, L. (1946). *Mathematical methods of statistics*.

Hartigan, J. A., and Wong, M. A. (1979). Algorithm as 136: a K-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* 28, 100–108. doi:10.2307/2346830

He, J., Li, M., Qiu, J., Pu, X., and Guo, Y. (2023). HOPEXGB: a consensual model for predicting miRNA/lncRNA-Disease associations using a heterogeneous disease-miRNA-lncRNA information network. *J. Chem. Inf. Model.* doi:10.1021/acs.jcim.3c00856

Heid, C. A., Stevens, J., Livak, K. J., and Williams, P. M. (1996). Real time quantitative PCR. *Genome Res.* 6, 986–994. doi:10.1101/gr.6.10.986

Kang, C., Zhang, H., Liu, Z., Huang, S., and Yin, Y. (2022). LR-GNN: a graph neural network based on link representation for predicting molecular associations. *Brief. Bioinform.* 23, bbab513. doi:10.1093/bib/bbab513

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). *LightGBM: a highly efficient gradient boosting decision tree*.

Lan, W., Dong, Y., Chen, Q., Liu, J., Wang, J., Chen, Y.-P. P., et al. (2021). IGNSCDA: predicting CircRNA-disease associations based on improved graph convolutional network and negative sampling. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 19, 3530–3538. doi:10.1109/TCBB.2021.3111607

Li, J., Zeng, X., Dou, Y., Xia, F., and Peng, S. (2021). "LADstackING: stacking ensemble learning-based computational model for predicting potential LncRNA-disease associations," in Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 177–182.

Li, M., Liu, M., Bin, Y., and Xia, J. (2020). Prediction of circRNA-disease associations based on inductive matrix completion. *BMC Med. Genomics* 13, 42. doi:10.1186/s12920-020-0679-0

Liang, Y., Zhang, Z.-Q., Liu, N.-N., Wu, Y.-N., Gu, C.-L., and Wang, Y.-L. (2022). MAGCNSE: predicting lncRNA-disease associations using multi-view attention graph convolutional network and stacking ensemble model. *BMC Bioinforma.* 23, 189. doi:10.1186/s12859-022-04715-w

Liu, J., Jiang, M., Guan, J., Wang, Y., Yu, W., Hu, Y., et al. (2022). Lncrna Kcnq1ot1 enhances the radioresistance of lung squamous cell carcinoma by targeting the mir-491-5p/tpx2-rnf2 Axis. *J. Thorac. Dis.* 14, 4081–4095. doi:10.21037/jtd-22-1261

Lu, C., and Xie, M. (2023). LDAEXC: LncRNA-disease associations prediction with deep autoencoder and XGBoost classifier. *Interdiscip. Sci. Comput. Life Sci.* 15, 439–451. doi:10.1007/s12539-023-00573-z

Maas, A. L., Hannun, A. Y., and Ng, A. Y. (2013). *Rectifier nonlinearities improve neural network acoustic models*.

Maaten, L. V. D., and Hinton, G. (2008). Visualizing data using T-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Maher, B. (2012). ENCODE: the human encyclopaedia. *Nature* 489, 46–48. doi:10.1038/489046a

Miranda, A. L. B., Garcia, L. P. F., Carvalho, A. C. P. L. F., and Lorena, A. C. (2009). "Use of classification algorithms in noise detection and elimination," in *Proceedings of the hybrid artificial intelligence systems*. Editors E. Corchado, X. Wu, E. Oja, Á. Herrero, and B. Baruque (Berlin, Heidelberg: Springer), 417–424.

Nematzadeh, Z., Ibrahim, R., and Selamat, A. (2020). A hybrid model for class noise detection using K-means and classification filtering algorithms. *SN Appl. Sci.* 2, 1303. doi:10.1007/s42452-020-3129-x

Peng, L., Huang, L., Lu, Y., Liu, G., Chen, M., and Han, G. (2022). "Identifying possible lncRNA-disease associations based on deep learning and positive-unlabeled learning," in Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 168–173.

Rinn, J. L., and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* 81, 145–166. doi:10.1146/annurev-biochem-051410-092902

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi:10.1016/0377-0427(87)90125-7

Shen, W., Pu, J., Zuo, Z., Gu, S., Sun, J., Tan, B., et al. (2022). The rna demethylase Alkbh5 promotes the progression and angiogenesis of lung cancer by regulating the stability of the lncrna Pvt1. *Cancer Cell Int.* 22, 353. doi:10.1186/s12935-022-02770-0

Shi, Z., Zhang, H., Jin, C., Quan, X., and Yin, Y. (2021). A representation learning model based on variational inference and graph autoencoder for predicting lncRNA-disease associations. *BMC Bioinforma.* 22, 136. doi:10.1186/s12859-021-04073-z

Sun, J., Shi, H., Wang, Z., Zhang, C., Liu, L., Wang, L., et al. (2014). Inferring novel lncrna–disease associations based on a random walk model of a lncrna functional similarity network. *Mol. Biosyst.* 10, 2074–2081. doi:10.1039/C3MB70608G

van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 27, 3036–3043. doi:10.1093/bioinformatics/btr500

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2017). *Graph attention networks*.

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26, 1644–1650. doi:10.1093/bioinformatics/btq241

Wei, H., Liao, Q., and Liu, B. (2021). iLncRNAdis-FB: identify lncRNA-disease associations by fusing biological feature blocks through deep neural network. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18, 1946–1957. doi:10.1109/TCBB.2020.2964221

Xie, G.-B., Chen, R.-B., Lin, Z.-Y., Gu, G.-S., Yu, J.-R., Liu, Z.-G., et al. (2023). Predicting lncrna-disease associations based on combining selective similarity matrix fusion and bidirectional linear neighborhood label propagation. *Brief. Bioinform.* 24, bbac595. doi:10.1093/bib/bbac595

Yao, D., Zhan, X., Zhan, X., Kwoh, C. K., Li, P., and Wang, J. (2020). A random forest based computational model for predicting novel lncRNA-disease associations. *BMC Bioinforma.* 21, 126. doi:10.1186/s12859-020-3458-1

Ye, M., Zhao, L., Zhang, L., Wu, S., Li, Z., Qin, Y., et al. (2022). Lncrna Nalt1 promotes colorectal cancer progression via targeting peg10 by sponging microrna-574-5p. *Cell Death Dis.* 13, 960. doi:10.1038/s41419-022-05404-5

Zhai, W., Li, X., Wu, S., Zhang, Y., Pang, H., and Chen, W. (2015). Microarray expression profile of lncRNAs and the upregulated ASLNC04080 lncRNA in human endometrial carcinoma. *Int. J. Oncol.* 46, 2125–2137. doi:10.3892/ijo.2015.2897

Zhang, Z., Xu, J., Wu, Y., Liu, N., Wang, Y., and Liang, Y. (2022). CapsNet-LDA: predicting lncRNA-disease associations using attention mechanism and capsule network based on multi-view data. *Brief. Bioinform.* 24, bbac531. doi:10.1093/bib/bbac531

Zhao, H., Yin, X., Xu, H., Liu, K., Liu, W., Wang, L., et al. (2023). LncTarD 2.0: an updated comprehensive database for experimentally-supported functional lncRNA-target regulations in human diseases. *Nucleic Acids Res.* 51, D199–D207. doi:10.1093/nar/gkac984

Zhao, L., Wang, J., Li, Y., Song, T., Wu, Y., Fang, S., et al. (2021). NONCODEV6: an updated database dedicated to long non-coding rna annotation in both animals and plants. *Nucleic Acids Res.* 49, D165–D171. doi:10.1093/nar/gkaa1046

Zhao, X., Wu, J., Zhao, X., and Yin, M. (2022). Multi-view contrastive heterogeneous graph attention network for lncrna-disease association prediction. *Brief. Bioinform.* 24, bbac548. doi:10.1093/bib/bbac548

Zhou, Y., Wang, X., Yao, L., and Zhu, M. (2022). LDAformer: predicting lncRNA-disease associations based on topological feature extraction and Transformer encoder. *Brief. Bioinform.* 23, bbac370. doi:10.1093/bib/bbac370