scientific reports

Check for updates

OPEN

Predicting lncRNA and disease associations with graph autoencoder and noise robust gradient boosting

Lili Tang[1], Liangliang Huang[2]✉ & Yi Yuan[1]✉

lncRNAs are densely related to many human diseases. Identifying new lncRNA-disease associations (LDAs) conduces to better deciphering mechanisms of diseases, finding new biomarkers, and further promoting their diagnosis and treatment. In this manuscript, we devise an LDA prediction framework called LDA-GARB. LDA-GARB first combines nonnegative matrix factorization to extract linear features of lncRNAs and diseases. Next, it computes lncRNA similarity and disease similarity and adopts a graph autoencoder to extract nonlinear features of lncRNAs and diseases. Subsequently, the extracted features are concatenated as a vector. Finally, it takes the obtained vector as inputs and designs a noise-robust gradient boosting model to uncover potential associations from unknown lncRNA-disease pairs. To investigate the LDA-GARB performance, we used precision, recall, accuracy, F1-score, AUC, and AUPR as measurement metrics and performed multiple comparison experiments. First, it was benchmarked with four representative LDA prediction methods, i.e., SDLDA, LDNFSGB, LDAenDL, and LDA-VGHB, under 5-fold cross validations on lncRNAs, diseases, and lncRNA-disease pairs. Next, it was compared with four representative boosting models, i.e., XGBoost, AdaBoost, CatBoost, and LightGBM, under the above three different cross validations. Subsequently, the performance of LDA-GARB against LDA-LNSUBRW, GAMCLDA, LDA-VGHB, LDAGM, and GANLDA on imbalanced data was evaluated. We also performed parameter sensitivity analysis and ablation experiments. The results demonstrated that LDA-GARB improved LDA prediction. Finally, LDA-GARB was applied to predict potential associated lncRNAs for colorectal cancer and breast cancer. CCDC26 and HAR1A have been inferred to have an association with the two cancers, respectively. As a useful LDA identification tool, LDA-GARB is freely available at https://github.com/smiling199/LDA-GARB.

Long non-coding RNAs (lncRNAs) are nucleotide sequences with length greater than 200[1,2]. lncRNAs are involved in many key physiological processes, for example, tissue development, tumorigenesis, and immune regulation. Furthermore, various human diseases have close associations with the dysregulation and mutation of lncRNAs[3].

Particularly, lncRNAs demonstrate differential roles in the progress and development of cancers[4]. lncRNAs have been taken as potential therapeutic molecular targets and offered new opportunities for cancer targeted therapy[5]. Many lncRNAs have been validated to be able to modulate chemotherapy resistance in cancers[6,7]. Consequently, the discovery of potential relationships between diseases and lncRNAs facilitates understanding molecular mechanisms of human complex diseases from the aspect of lncRNAs, detecting disease biomarkers, assisting in their diagnosis and treatment, and further promoting the development of personalized medicine[8].

However, potential lncRNA-disease association (LDA) identification is a huge challenge for biologists due to high cost and labor, and low success rate of *in vivo* experiments although biological experiments have yielded some LDAs[9]. Thus, computational techniques have been increasingly applied in association prediction tasks including LDA prediction[3,10,11]. Chen et al.[12] constructed an LDA database and provided experimentally validated LDAs for 166 diseases. Depending on the database, Chen et al.[13] inferred potential LDAs by combining lncRNA expressions. Following this work, many computational tools, mainly including network-based methods and machine learning-based methods, have been devised to uncover new LDAs[14].

[1]School of Computer Science, Hunan University of Technology, Zhuzhou 412007, China. [2]School of Information Technology and Administration, Hunan University of Finance and Economics, Changsha 410125, China. ✉email: hll20217@163.com; yuanyi@hut.edu.cn

Network-based methods first compute lncRNA similarity and disease similarity matrices according to their biological information, and then evaluate association score between each lncRNA-disease pair through network algorithms. These methods include Laplacian regularized least squares[13], KATZ measure[15], heterogeneous network model[16], network consistency projection[17], local random walk[18], Laplacian normalized random walk with restart[19] and bidirectional linear neighborhood label propagation[20].

Machine learning has been broadly utilized in various linkage prediction fields including LDA identification[21–23]. They first learn features of lncRNAs and diseases and then classify unknown lncRNA-disease pairs. Traditional machine learning-based LDA prediction methods include rotation forest[24], random forest[25], multi-label learning[26], matrix factorization[27], inductive matrix completion[28], weighted matrix factorization[29], matrix decomposition[30], collaborative filtering[31], bipartite local model[32], and heterogeneous Newton boosting machine[2]. Recently, deep learning algorithms have been gradually adopted to discover new LDAs due to the powerful representation learning ability, for example, collaborative deep learning[33], deep belief network[34], generative adversarial network[35], graph contrastive learning[36], deep neural network[37–39], heterogeneous graph learning[40], capsule network[41], dual-net neural network[42], graph convolutional autoencoder[43], graph attention network[44], and residual graph convolutional network with attention mechanism[8].

Machine learning has promoted LDA prediction. However, LDA datasets are imbalanced and contain noises. Machine learning-based LDA prediction algorithms, especially traditional boosting models, remain limitations in label noise, imbalanced datasets, and LDA feature extraction. To address the above problems, in this manuscript, we develop a computational model called LDA-GARB to interpret potential LDAs by combining Nonnegative Matrix Factorization (NMF), Graph Autoencoder (GAE), and noise-Robust gradient Boosting. This work mainly has three contributions:

- To solve the limitations of label noise and data imbalance in LDA classification task, we present a noise-robust gradient boosting model to classify unobserved lncRNA-disease pairs by integrating Gradient Boosting Decision Trees (GBDT) and robust loss.
- To obtain abundant LDA features, we leverage NMF for extracting linear features and GAE for extracting nonlinear features.
- We predict that CCDC26 and HAR1A could have an association with colorectal cancer (CRC) and breast cancer, respectively.

Results

In this manuscript, as shown in Fig. 1, we proposed an LDA prediction method, LDA-GARB, by incorporating LDA feature extraction through NMF and GAE[45] and LDA classification via the noise-Robust gradient Boosting model. Finally, we predicted associated lncRNAs for CRC and breast cancer through LDA-GARB.

Data preparation

Two human LDA datasets, Dataset 1 and Dataset 2[2,42], were used to evaluate the model and achieve predictions. lncRNAs, diseases, and experimentally confirmed LDAs in the two datasets were obtained from the lncRNADisease v2.0[46] and MNDR v2.0[47] databases, respectively. We removed diseases which have no MESH information or regular name, and lncRNAs which are lack of sequence data. After preprocessing, Dataset 1 includes 92 lncRNAs, 157 diseases, and 605 LDAs. Dataset 2 includes 89 lncRNAs, 190 diseases, and 1,529 LDAs. The preprocessed datasets are illustrated in Table 1. The association network was represented as a matrix $\boldsymbol{Y} \in \Re^{n \times m}$ where y_{ij} is 1 when there is an association between the i-th lncRNA and the j-th disease, y_{ij} is 0 otherwise.

Experimental settings

Peng et al.[2] designed multiple 5-fold Cross Validation (CV) ways and provided insights into evaluating the performance of linkage prediction models. Inspired by CVs proposed by Peng et al.[2], we used three distinct 5-fold CVs to test the model performance. The three CVs include 5-fold CV on lncRNAs (CV_1), 5-fold CV on diseases (CV_2), and 5-fold CV on lncRNA-disease pairs (CV_3). In each round, they run experiments as follows:

- CV_1: 20% of lncRNAs in $\boldsymbol{Y}$ were randomly hidden for test and the rest for training.
- CV_2: 20% of diseases in $\boldsymbol{Y}$ were randomly hidden for test and the rest for training.
- CV_3: 20% of lncRNA-disease pairs in $\boldsymbol{Y}$ were randomly hidden for test and the rest for training.

To assess the prediction accuracy of LDA-GARB, we used six machine learning indictors. The six evaluation metrics were precision, recall, accuracy, F1-score, Area Under receiver operating characteristic (ROC) Curve (AUC), and Area Under Precision-Recall (PR) curve (AUPR), which were provided by Refs.[2,48], respectively. All experiments were run on the Ubuntu system with 12th Gen Intel(R) Core (TM) i7-12650H, NVIDIA GeForce RTX 4060 Laptop GPU, and RAM of 32.0 GB. In addition, the version numbers of related softwares were Python version 3.8, Numpy version 1.23.2, Pandas version 2.1.4, scikit-learn version 1.3.0, and XGBoost version 2.0.0, respectively. The parameters of LDA-GARB and four competing LDA prediction methods (i.e., SDLDA, LDNFSGB, LDAenDL, and LDA-VGHB) were shown in Table 2.

Baselines

We compared the performance of LDA-GARB and four state-of-the-art LDA prediction models on two datasets. The four comparison methods were SDLDA[49], LDNFSGB[50], LDAenDL[37], and LDA-VGHB[2], respectively. For one lncRNA-disease pair, SDLDA[49] first learned its features by integrating singular value decomposition (SVD) and deep learning and then determined that it was associated or not through a full connection layer. LDNFSGB[50]

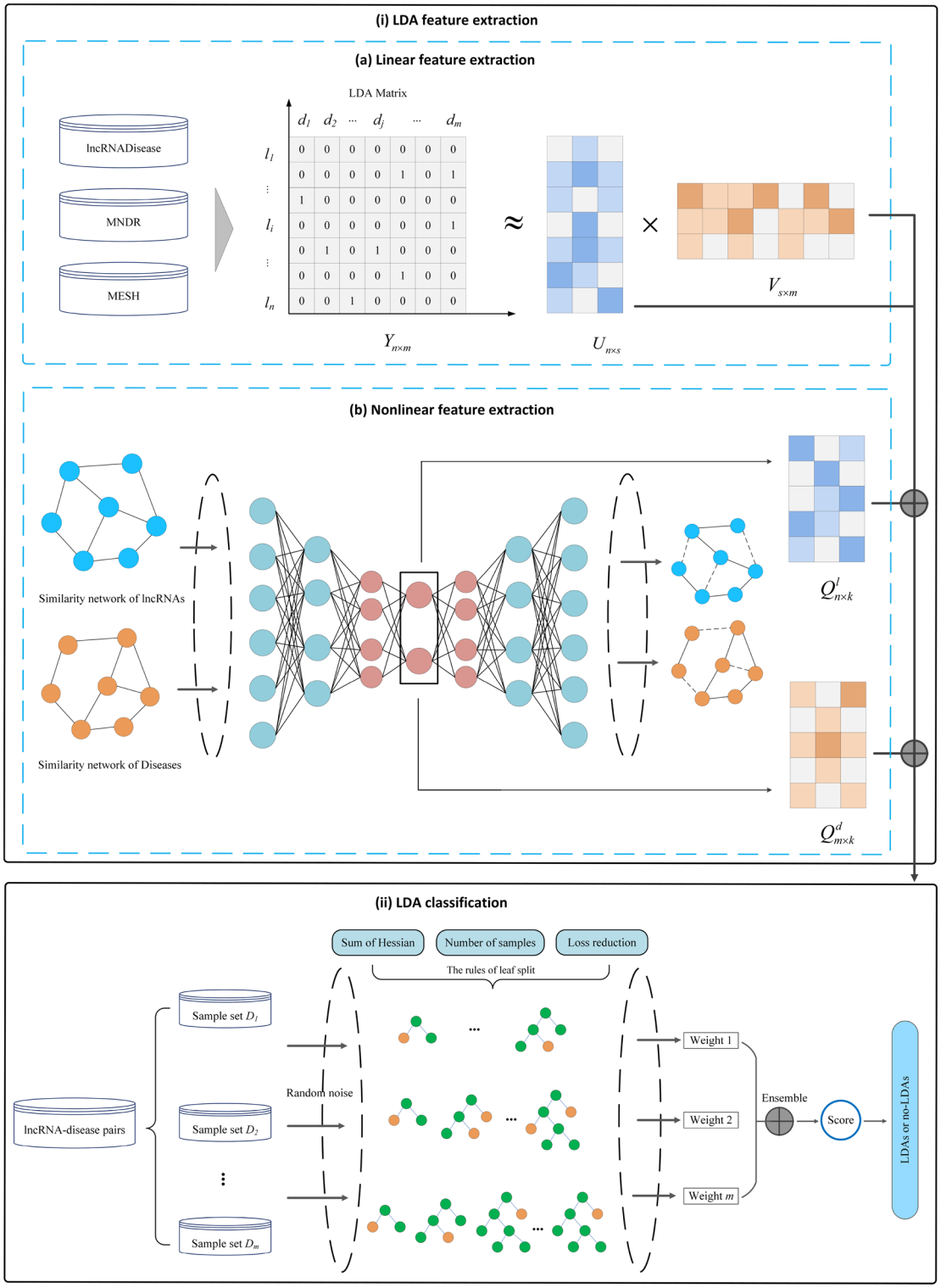

Fig. 1. The pipeline for LDA prediction with LDA-GARB. (i) Feature extraction. Linear and nonlinear features of lncRNAs and diseases are extracted by NMF and GAE. And each LDA is depicted as a vector through concatenating the learned linear and nonlinear features. (ii) LDA classification. The noise-robust gradient boosting model is designed to classify unobserved LDAs based on the extracted LDA features..

Dataset	lncRNA	Disease	Association
Dataset 1	92	157	605
Dataset 2	89	190	1529

Table 1. Introduction of two LDA datasets.

Method	Parameter settings
SDLDA	Pca dim = 64, pca maxiter = 200, learning rate = 0.001, batch size = 800, R layer1 num = 48, epoch num = 200, L layer1 num = 48, L layer2 num = 32, R layer2 num = 32
LDNFSGB	Encoding dim = 128, optimizer = 'adadelta', loss = 'binary crossentropy', epochs = 100, batch size = 128, activation = 'sigmoid', n estimators = 1200
LDAenDL	LightGBM: num leaves = 31, max depth = – 1, learning rate = 0.1, n estimators = 100, min split gain = 0, min child weight = 1e–3, min child samples = 20, subsample = 1 DNN: number of layers = 3, dropout = 0.2, activation = 'relu', epochs = 450, optimizer = Adam(), learning rate = 0.001
LDA-VGHB	Objective = 'logloss', num round = 1000, use gpu =True, gpu id = 1, learning rate = 0.001, min max depth = 1, max max depth = 25, subsample = 0.8
LDA-GARB	NMF: s = 64, $\lambda_1 = \lambda_2 = 0.01$ GAE: epochs = 120, aggregator = 'GraphSAGE', embedding size = 128, layers = 1, dropout = 0.4, slope = 0.1, learning rate = 0.001 Noise-Robust Gradient Boosting: max depth = (2, 12), reg alpha = (1e-4, 1.0), reg lambda = (1e-4, 5.0), learning rate = (1e–3, 1.0), n estimators = (10, 1200, 10), r = [0.0, 0.5, 1.0], q = [0.0, 0.1, 0.3, 0.5]

Table 2. Parameter settings in SDLDA, LDNFSGB, LDAenDL, LDA-VGHB, and LDA-GARB.

first reduced its feature dimension via an autoencoder and classified it through a gradient boosting algorithm. LDAenDL[37] extracted its biological features by integrating graph convolutional network, convolutional neural network, and graph attention network, and then inferred its association through deep neural network and LightGBM. LDA-VGHB[2] incorporated SVD and variational graph autoencoder for learning its features and heterogeneous Newton boosting machine for obtaining its class.

Additionally, similar to work[2], LDA-GARB was compared with four representative boosting models under CV_1, CV_2, and CV_3. These four boosting models include XGBoost[51], AdaBoost[52], CatBoost[53], and LightGBM[54]. They correspond to Extreme Gradient Boosting, weak learning, categorical boosting algorithm, and boosting with one-side sampling along with exclusive feature bundling. The parameters in the four boosting algorithms were set to defaults.

To evaluate the ability of LDA-SCGB on imbalanced data, we compared it with five LDA prediction models, LDA-LNSUBRW[55], GAMCLDA[56], LDA-VGHB[2], LDAGM[43], and GANLDA[44], where their parameters were set to defaults. LDA-LNSUBRW[55] used an unbalanced bi-random walk for negative LDA selection. GAMCLDA[56] employed a cost-sensitive neural network to handle the problem of imbalance issue between positive LDAs and negative LDAs. LDA-VGHB is recent superior LDA identification model.

Performance comparison under CV_1

To evaluate the LDA-GARB performance when inferring diseases related to a target lncRNA under CV_1, we randomly selected 80% of lncRNAs for training and the rest for test. Figure 2A and D demonstrate the classification accuracy of LDA-GARB, SDLDA, LDNFSGB, LDAenDL, and LDA-VGHB under CV_1. Figure 3A and B delineated their ROC and PR curves on Dataset 1. Figure 3G and H delineated the ROC and PR curves on Dataset 2.

Table 3 shows their precision, recall, accuracy, F1-score, AUC, and AUPR under CV_1. From the results in Table 3, we found that LDA-GARB outperformed other four methods. It calculated the highest recall, accuracy, F1-score, AUC, and AUPR, with AUCs (0.9180 and 0.9716) better 3.99% and 1.80% than LDA-VGHB on Datasets 1 and 2, respectively, and AUPRs (0.9160 and 0.9723) better 2.30% and 1.09% than LDA-VGHB, respectively. In summary, it accurately inferred potential diseases related to a target lncRNA.

Performance comparison under CV_2

To assess the LDA-GARB performance when inferring lncRNAs related to a target disease under CV_2, we randomly selected 80% of diseases for training and the rest for test. Figure 2B and E demonstrate the classification accuracy of LDA-GARB, SDLDA, LDNFSGB, LDAenDL, and LDA-VGHB on two datasets under CV_2, respectively. Figure 3C and D delineated their ROC and PR curves on Dataset 1 under CV_2. Figure 3I and J delineated the ROC and PR curves on Dataset 2 under CV_2.

Table 4 shows the performance of LDA-GARB and the above four baselines under CV_2. From the results, we observed that LDA-GARB exceeded the four baselines. It calculated the highest recall, accuracy, F1-score, and AUC, with AUCs (0.9493 and 0.9817) better 3.99% and 1.80% than LDA-VGHB on Datasets 1 and 2, respectively, and AUPR with 0.9757 better 0.30 % than LDA-VGHB on Dataset 2. Although AUPR computed by LDA-GARB was slightly smaller than LDA-VGHB (0.9415 vs. 0.9429) on Dataset 1, their difference was very

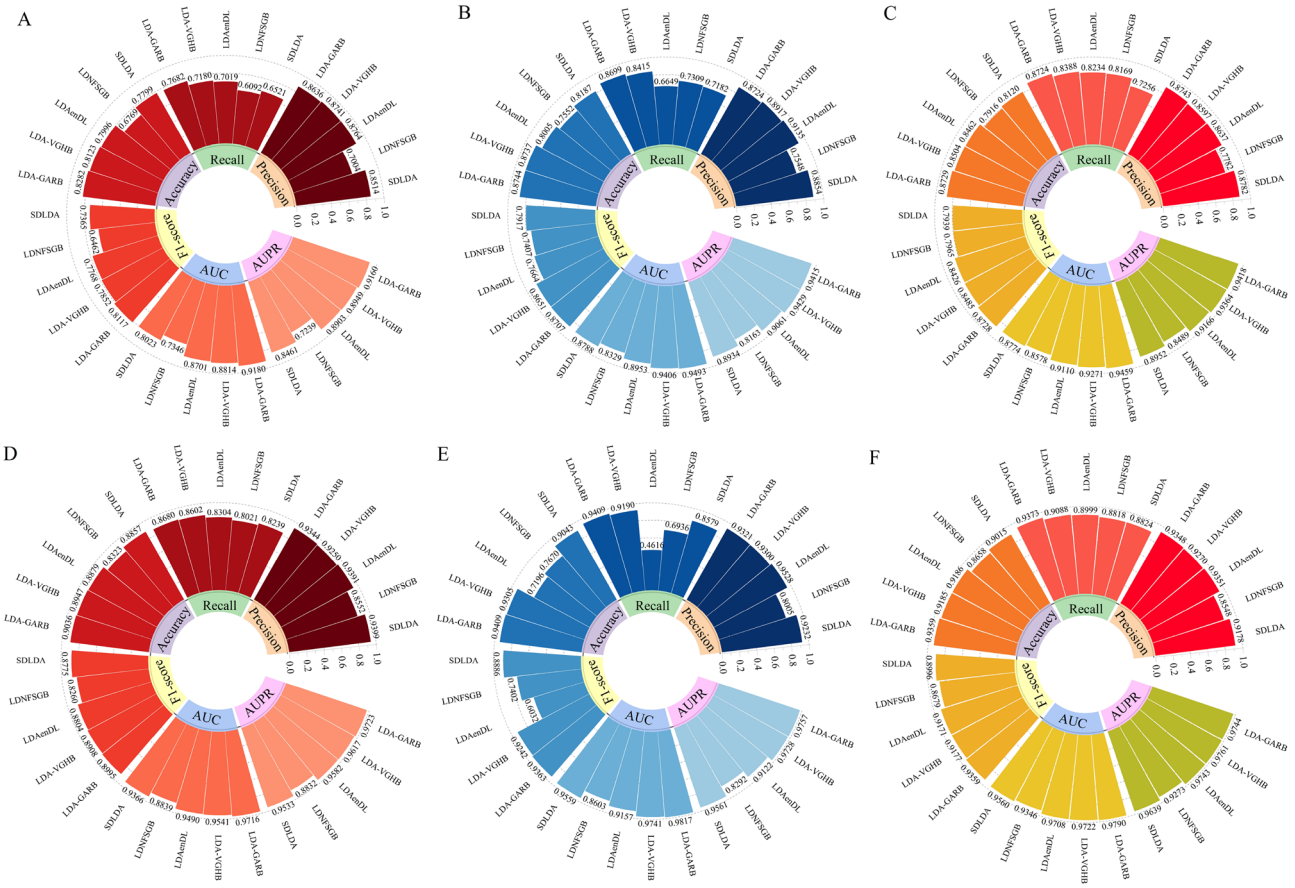

Fig. 2. Performance of LDA-GARB and other four methods. (**A**–**C**) on Dataset 1 under CV_1, CV_2, and CV_3, respectively. (**D**–**F**) on Dataset 2 under CV_1, CV_2, and CV_3, respectively.

small. In summary, it relatively accurately predicted lncRNAs that may associate with a disease without known lncRNA data.

Performance comparison under CV_3

To measure the performance of LDA-GARB when inferring potential LDAs under CV_3, we randomly selected 80% of lncRNA-disease pairs for training and the rest for test. Figure 2C and F demonstrate the classification accuracy of LDA-GARB, SDLDA, LDNFSGB, LDAenDL, and LDA-VGHB on two LDA datasets under CV_3, respectively. Figure 3E and F delineated their ROC and PR curves on Dataset 1 under CV_3. Figure 3K and L delineated their ROC and PR curves on Dataset 2 under CV_2.

Table 5 shows the values corresponding to the six indictors under CV_3. The results demonstrated that LDA-GARB surpassed other four methods. It calculated the highest recall, accuracy, F1-score, and AUC, with AUCs (0.9459 and 0.9790) better 1.99% and 0.69% than LDA-VGHB on Datasets 1 and 2, respectively, and AUPR with 0.9418 better 0.57% than LDA-VGHB on Dataset 1. Similar to CV_2, although LDA-GARB computed slightly lower AUPR than LDA-GARB on Dataset 2, the difference was tiny. Thus, LDA-GARB could effectively capture new associations from unknown lncRNA-disease pairs.

Performance under different boosting algorithms

LDA-GARB used a noise-robust gradient boosting model for classifying unobserved lncRNA-disease pairs. To validate the LDA classification ability of the robust boosting model, we compared LDA-GARB with XGBoost, AdaBoost, CatBoost, and LightGBM under three different CVs. The results are shown in Tables 6, 7, 8 and Fig. 4. On two datasets, LDA-GARB computed the highest performance in most cases under all three CVs, greatly outperforming other four boosting models. As a result, the noise-robust gradient boosting model can perform effective predictions.

Performance on imbalanced data

In the LDA matrix $\boldsymbol{Y}$, known associations are very few and negative associations are very difficult to obtain. Consequently, most computational LDA prediction models randomly selected negative associations from unobserved lncRNA-disease pairs. However, these unlabeled pairs may contain positive samples, thereby severely affecting the model predictions. Thus, researchers have explored machine learning algorithms such as positive-unlabeled learning to select reliable negative associations, or devised a model with more robustness

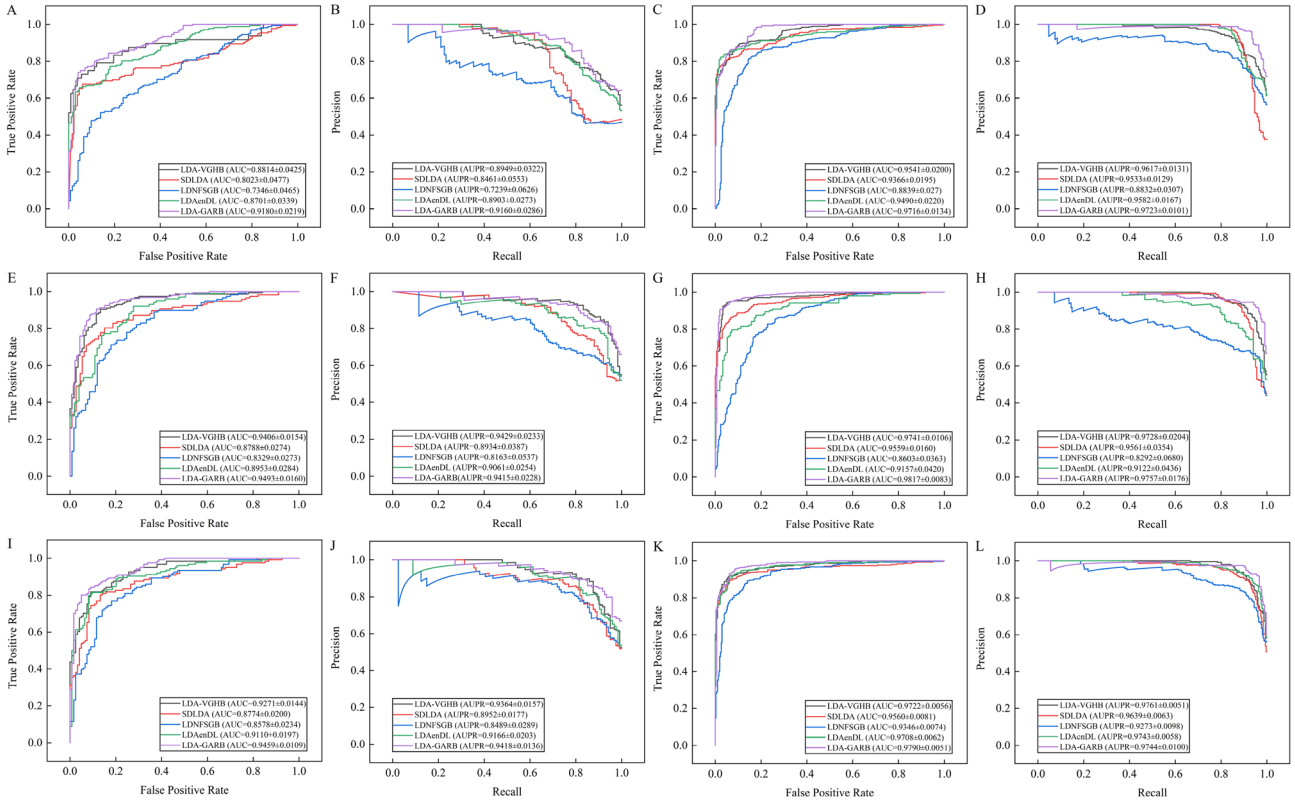

Fig. 3. The ROC and PR curves of LDA-GARB and other four methods. A-B, C-D, and E-F denote their ROC and PR curves under CV_1, CV_2, and CV_3 on Dataset 1, respectively. G-H, I-J and K-L denote their ROC and PR curves under CV under CV_1, CV_2, and CV_3 on Dataset 2, respectively.

	Dataset	SDLDA	LDNFSGB	LDAenDL	LDA-VGHB	LDA-GARB
Precision	Dataset 1	0.8514 ± 0.0509	0.7004±0.0639	**0.8764 ± 0.0493**	0.8741 ± 0.0484	0.8636 ± 0.0450
	Dataset 2	**0.9399 ± 0.0154**	0.8552 ± 0.0393	0.9391 ± 0.0290	0.9250 ± 0.0201	0.9344 ± 0.0147
Recall	Dataset 1	0.6521 ± 0.0732	0.6092 ± 0.0790	0.7019 ± 0.0639	0.7180 ± 0.0713	**0.7682 ± 0.0452**
	Dataset 2	0.8239 ± 0.0437	0.8021 ± 0.0498	0.8304 ± 0.0523	0.8602 ± 0.0395	**0.8680 ± 0.0429**
Accuracy	Dataset 1	0.7799 ± 0.0341	0.6769 ± 0.0423	0.7996 ± 0.0312	0.8123 ± 0.0384	**0.8282 ± 0.0338**
	Dataset 2	0.8857 ± 0.0283	0.8323 ± 0.0230	0.8879 ± 0.0289	0.8947 ± 0.0258	**0.9036 ± 0.0281**
F1-score	Dataset 1	0.7365 ± 0.0563	0.6462 ± 0.0451	0.7768 ± 0.0399	0.7852 ± 0.0412	**0.8117 ± 0.0312**
	Dataset 2	0.8775 ± 0.0278	0.8260 ± 0.0230	0.8804 ± 0.0334	0.8908 ± 0.0227	**0.8995 ± 0.0266**
AUC	Dataset 1	0.8023 ± 0.0477	0.7346 ± 0.0465	0.8701 ± 0.0339	0.8814 ± 0.0425	**0.9180 ± 0.0219**
	Dataset 2	0.9366 ± 0.0195	0.8839 ± 0.0270	0.9490 ± 0.0220	0.9541 ± 0.0200	**0.9716 ± 0.0134**
AUPR	Dataset 1	0.8461 ± 0.0553	0.7239 ± 0.0626	0.8903 ± 0.0273	0.8949 ± 0.0322	**0.9160 ± 0.0286**
	Dataset 2	0.9533 ± 0.0129	0.8832 ± 0.0307	0.9582 ± 0.0167	0.9617 ± 0.0131	**0.9723 ± 0.0101**

Table 3. Performance comparison under CV_1. The best performance is denoted as bold.

to solve the data imbalance issue. In this section, we adopted a noise-robust gradient boosting model to run predictions on imbalanced LDA datasets.

To investigate the LDA classification accuracy of LDA-SCGB on imbalanced dataset, we compared it with five representative association prediction methods, i.e., LDA-LNSUBRW[55], GAMCLDA[56], LDA-VGHB[2], LDAGM[43], and GANLDA[44]. LDA-LNSUBRW[55] adopted unbalanced bi-random walk for potential LDA inference. GAMCLDA[56] employed graph autoencoder matrix completion to identify new LDAs. LDA-VGHB[2] classified unknown lncRNA-disease pairs through heterogeneous Newton boosting machine. LDAGM[43] learned deep topological features bases on linkages between lncRNA, diseases, and miRNA, and then devised a multi-view heterogeneous network to infer LDAs by combining graph convolutional autoencoder. GANLDA[44] presented a graph attention network to compute LDA score matrix. The five methods are representative LDA prediction models. Table 9 shows their AUCs and AUPRs on two LDA datasets under CV_3. Figure 5 depicts

	Dataset	SDLDA	LDNFSGB	LDAenDL	LDA-VGHB	LDA-GARB
Precision	Dataset 1	0.8854 ± 0.0377	0.7548 ± 0.0639	**0.9135 ± 0.0317**	0.8917 ± 0.0316	0.8724 ± 0.0365
	Dataset 2	0.9232 ± 0.0331	0.8005 ± 0.0625	**0.9528 ± 0.0225**	0.9300 ± 0.0251	0.9321 ± 0.0277
Recall	Dataset 1	0.7182 ± 0.0694	0.7309 ± 0.0646	0.6649 ± 0.0814	0.8415 ± 0.0449	**0.8699 ± 0.0377**
	Dataset 2	0.8579 ± 0.0655	0.6936 ± 0.0794	0.4616 ± 0.1702	0.9190 ± 0.0397	**0.9409 ± 0.0262**
Accuracy	Dataset 1	0.8187 ± 0.0282	0.7552 ± 0.0291	0.8005 ± 0.0381	0.8737 ± 0.0177	**0.8744 ± 0.0255**
	Dataset 2	0.9043 ± 0.0174	0.7670 ± 0.0432	0.7196 ± 0.0821	0.9305 ± 0.0153	**0.9409 ± 0.0158**
F1-score	Dataset 1	0.7917 ± 0.0519	0.7407 ± 0.0526	0.7664 ± 0.0593	0.8651 ± 0.0304	**0.8707 ± 0.0316**
	Dataset 2	0.8886 ± 0.0475	0.7402 ± 0.0577	0.6032 ± 0.1612	0.9242 ± 0.0298	**0.9363 ± 0.0243**
AUC	Dataset 1	0.8788 ± 0.0274	0.8329 ± 0.0273	0.8953 ± 0.0284	0.9406 ± 0.0154	**0.9493 ± 0.0160**
	Dataset 2	0.9559 ± 0.0160	0.8603 ± 0.0363	0.9157 ± 0.0420	0.9741 ± 0.0106	**0.9817 ± 0.0083**
AUPR	Dataset 1	0.8934 ± 0.0387	0.8163 ± 0.0537	0.9061 ± 0.0254	**0.9429 ± 0.0233**	0.9415 ± 0.0228
	Dataset 2	0.9561 ± 0.0354	0.8292 ± 0.0680	0.9122 ± 0.0436	0.9728 ± 0.0204	**0.9757 ± 0.0176**

Table 4. Performance comparison under CV_2. The best performance is denoted as bold.

	Dataset	SDLDA	LDNFSGB	LDAenDL	LDA-VGHB	LDA-GARB
Precision	Dataset 1	**0.8782 ± 0.0306**	0.7782 ± 0.0270	0.8637 ± 0.0312	0.8597 ± 0.0269	0.8743 ± 0.0284
	Dataset 2	0.9178 ± 0.0154	0.8548 ± 0.0156	**0.9351 ± 0.0157**	0.9270 ± 0.0143	0.9348 ± 0.0130
Recall	Dataset 1	0.7256 ± 0.0376	0.8169 ± 0.0408	0.8234 ± 0.0314	0.8388 ± 0.0332	**0.8724 ± 0.0305**
	Dataset 2	0.8824 ± 0.0198	0.8818 ± 0.0204	0.8999 ± 0.0179	0.9088 ± 0.0169	**0.9373 ± 0.0137**
Accuracy	Dataset 1	0.8120 ± 0.0216	0.7916 ± 0.0256	0.8462 ± 0.0229	0.8504 ± 0.0189	**0.8729 ± 0.0204**
	Dataset 2	0.9015 ± 0.0114	0.8658 ± 0.0127	0.9186 ± 0.0126	0.9185 ± 0.0110	**0.9359 ± 0.0085**
F1-score	Dataset 1	0.7939 ± 0.0260	0.7965 ± 0.0262	0.8426 ± 0.0232	0.8485 ± 0.0198	**0.8728 ± 0.0204**
	Dataset 2	0.8996 ± 0.0119	0.8679 ± 0.0129	0.9171 ± 0.0130	0.9177 ± 0.0112	**0.9359 ± 0.0085**
AUC	Dataset 1	0.8774 ± 0.0200	0.8578 ± 0.0234	0.9110 ± 0.0197	0.9271 ± 0.0144	**0.9459 ± 0.0109**
	Dataset 2	0.9560 ± 0.0081	0.9346 ± 0.0074	0.9708 ± 0.0062	0.9722 ± 0.0056	**0.9790 ± 0.0051**
AUPR	Dataset 1	0.8952 ± 0.0177	0.8489 ± 0.0289	0.9166 ± 0.0203	0.9364 ± 0.0157	**0.9418 ± 0.0136**
	Dataset 2	0.9639 ± 0.0063	0.9273 ± 0.0098	0.9743 ± 0.0058	**0.9761 ± 0.0051**	0.9744 ± 0.0100

Table 5. Performance comparison under CV_3. The best performance is denoted as bold.

	Dataset	XGBoost	AdaBoost	CatBoost	LightGBM	LDA-GARB
Precision	Dataset 1	0.8285 ± 0.0450	0.8014 ± 0.0467	0.8436 ± 0.0386	0.8359 ± 0.0537	**0.8636±0.0450**
	Dataset 2	0.9094 ± 0.0203	0.8945 ± 0.0250	0.9200 ± 0.0224	0.9007 ± 0.0240	**0.9344±0.0147**
Recall	Dataset 1	0.7627 ± 0.0615	0.7407 ± 0.0855	**0.7772±0.0541**	0.7624 ± 0.0668	0.7682 ± 0.0452
	Dataset 2	0.8739 ± 0.0305	0.8605 ± 0.0432	**0.8761±0.0427**	0.8755 ± 0.0346	0.8680 ± 0.0429
Accuracy	Dataset 1	0.8083 ± 0.0325	0.7946 ± 0.0312	0.8191 ± 0.0322	0.8190 ± 0.0268	**0.8282±0.0338**
	Dataset 2	0.8935 ± 0.0241	0.8795 ± 0.0263	0.8997 ± 0.0288	0.8895 ± 0.0293	**0.9036±0.0281**
F1-score	Dataset 1	0.7925 ± 0.0410	0.7681 ± 0.0626	0.8077 ± 0.0352	0.7964 ± 0.0550	**0.8117±0.0312**
	Dataset 2	0.8910 ± 0.0205	0.8764 ± 0.0248	0.8969 ± 0.0266	0.8876 ± 0.0251	**0.8995±0.0266**
AUC	Dataset 1	0.8938 ± 0.0263	0.8528 ± 0.0386	0.9099 ± 0.0214	0.8988 ± 0.0277	**0.9180±0.0219**
	Dataset 2	0.9629 ± 0.0125	0.9421 ± 0.0265	0.9660 ± 0.0133	0.9574 ± 0.0212	**0.9716±0.0134**
AUPR	Dataset 1	0.8799 ± 0.0421	0.8558 ± 0.0591	0.9053 ± 0.0275	0.8939 ± 0.0582	**0.9160±0.0286**
	Dataset 2	0.9660 ± 0.0097	0.9495 ± 0.0189	0.9674 ± 0.0111	0.9628 ± 0.0135	**0.9723±0.0101**

Table 6. Performance of different boosting algorithms under CV_1. The best performance is denoted as bold.

the corresponding ROC and PR curves. The results elucidated that LDA-GARB obviously surpassed LDA-LNSUBRW, GAMCLDA, LDA-VGHB, LDAGM, and GANLDA, demonstrating its powerful ability to solve with imbalanced datasets.

Sensitivity of parameters

LDA-GARB used GAE to extract nonlinear features of lncRNAs and diseases. Consequently, the embeddings of lncRNAs and diseases are particularly important to LDA prediction performance. Thus, we analyzed the impact

	Dataset	XGBoost	AdaBoost	CatBoost	LightGBM	LDA-GARB
Precision	Dataset 1	0.8609 ± 0.0409	0.8172 ± 0.0457	0.8621 ± 0.0391	0.8565 ± 0.0393	**0.8724±0.0365**
	Dataset 2	0.8966 ± 0.0316	0.8805 ± 0.0367	0.9126 ± 0.0271	0.9052 ± 0.0341	**0.9321±0.0277**
Recall	Dataset 1	0.8230 ± 0.0422	0.8271 ± 0.0581	0.8342 ± 0.0444	0.8358 ± 0.0503	**0.8699±0.0377**
	Dataset 2	0.9026 ± 0.0345	0.8569 ± 0.0597	0.9126 ± 0.0271	0.9063 ± 0.0354	**0.9409±0.0262**
Accuracy	Dataset 1	0.8486 ± 0.0239	0.8283 ± 0.0245	0.8533 ± 0.0251	0.8515 ± 0.0252	**0.8744±0.0255**
	Dataset 2	0.9055 ± 0.0161	0.8815 ± 0.0161	0.9177 ± 0.0116	0.9122 ± 0.0143	**0.9409±0.0158**
F1-score	Dataset 1	0.8406 ± 0.0315	0.8214 ± 0.0468	0.8473 ± 0.0350	0.8449 ± 0.0345	**0.8707±0.0316**
	Dataset 2	0.8994 ± 0.0306	0.8681 ± 0.0461	0.9094 ± 0.0305	0.9056 ± 0.0326	**0.9363±0.0243**
AUC	Dataset 1	0.9291 ± 0.0168	0.8926 ± 0.0246	0.9424 ± 0.0162	0.9325 ± 0.0171	**0.9493±0.0160**
	Dataset 2	0.9666 ± 0.0098	0.9521 ± 0.0116	0.9731 ± 0.0076	0.9702 ± 0.0093	**0.9817±0.0083**
AUPR	Dataset 1	0.9258 ± 0.0274	0.8859 ± 0.0477	**0.9420±0.0222**	0.9242 ± 0.0282	0.9415 ± 0.0228
	Dataset 2	0.9603 ± 0.0233	0.9463 ± 0.0296	0.9687 ± 0.0198	0.9649 ± 0.0310	**0.9757±0.0176**

Table 7. Performance of different boosting algorithms under CV_2. The best performance is denoted as bold.

	Dataset	XGBoost	AdaBoost	CatBoost	LightGBM	LDA-GARB
Precision	Dataset 1	0.8624 ± 0.0222	0.8085 ± 0.0251	0.8737 ± 0.0285	0.8531 ± 0.0267	**0.8743±0.0284**
	Dataset 2	0.8983 ± 0.0156	0.9041 ± 0.0155	0.9121 ± 0.0173	0.9070 ± 0.0130	**0.9348±0.0130**
Recall	Dataset 1	0.8450 ± 0.0334	0.8147 ± 0.0338	0.8590 ± 0.0320	0.8245 ± 0.0303	**0.8724±0.0305**
	Dataset 2	0.9052 ± 0.0144	0.8596 ± 0.0193	0.9073 ± 0.0144	0.9115 ± 0.0148	**0.9373±0.0137**
Accuracy	Dataset 1	0.8547 ± 0.0178	0.8103 ± 0.0197	0.8668 ± 0.0189	0.8408 ± 0.0201	**0.8729±0.0204**
	Dataset 2	0.9012 ± 0.0118	0.8841 ± 0.0129	0.9097 ± 0.0104	0.9089 ± 0.0089	**0.9359±0.0085**
F1-score	Dataset 1	0.8531 ± 0.0192	0.8110 ± 0.0207	0.8657 ± 0.0192	0.8381 ± 0.0208	**0.8728±0.0204**
	Dataset 2	0.9016 ± 0.0117	0.8811 ± 0.0135	0.9096 ± 0.0102	0.9091 ± 0.0090	**0.9359±0.0085**
AUC	Dataset 1	0.9260 ± 0.0131	0.8788 ± 0.0168	0.9427 ± 0.0109	0.9163 ± 0.0141	**0.9459±0.0109**
	2	0.9679 ± 0.0050	0.9580 ± 0.0065	0.9723 ± 0.0045	0.9689 ± 0.0053	**0.9790±0.0051**
AUPR	Dataset 1	0.9240 ± 0.0145	0.8750 ± 0.0194	0.9411 ± 0.0108	0.9135 ± 0.0174	**0.9418±0.0136**
	Dataset 2	0.9690 ± 0.0048	0.9601 ± 0.0065	0.9732 ± 0.0046	0.9696 ± 0.0061	**0.9744±0.0100**

Table 8. Performance of different boosting algorithms under CV_3. The best performance is denoted as bold.

of different embedding dimensions k of lncRNAs and diseases and the number of different encoder layers N on the model performance.

Tables 10, 11, 12, 13, 14, 15, 16, 17 and 18 show the performance of LDA-GARB when k was set to 64, 128, and 256 and N was set to 1, 2, 3, 4, and 5 under CV_1, CV_2, and CV_3, respectively. We comprehensively considered the performance of LDA-GARB under different embedding dimensions and different encoder layer number, and found that LDA-GARB obtained relatively good performance when $k = 64$ and $N = 1$. Therefore, we set $k = 64$ and $N = 1$.

In addition, the noise-robust gradient boosting model fully utilized robust focal loss and thus had high robustness to effectively address the issues of label noise and data imbalance. Moreover, during training, the model automatically optimized parameters based on the proportion of noises and used the optimized parameters for testing. Therefore, we didn't additionally analyze the impact of the model parameters on the model performance. The related parameters were shown in Table 2.

Ablation study

The proposed LDA-GARB method extracted LDA linear features through NMF and nonlinear features through GAE. To measure the effect of different feature selection methods on the LDA prediction performance, we conducted ablation experiments. Tables 19, 20, 21 and Fig. 6 give the performance of LDA-GARB with linear features, nonlinear features, their combination under CV_1, CV_2, and CV_3, respectively. As shown in Tables 19, 20, 21 and Fig. 6, under most conditions, LDA-GARB with the two types of features outperformed LDA-GARB only with linear features and LDA-GARB only with nonlinear features. Thus, the combination of the two types of features assists in improving LDA prediction.

Case study

CRC and breast cancer are two cancers which severely affect human health. Identifying potential lncRNAs for them helps their diagnosis and therapy. We have validated the LDA-GARB performance after multiple experiments. Subsequently, we adopted LDA-GARB to infer associated lncRNAs for CRC and breast cancer.

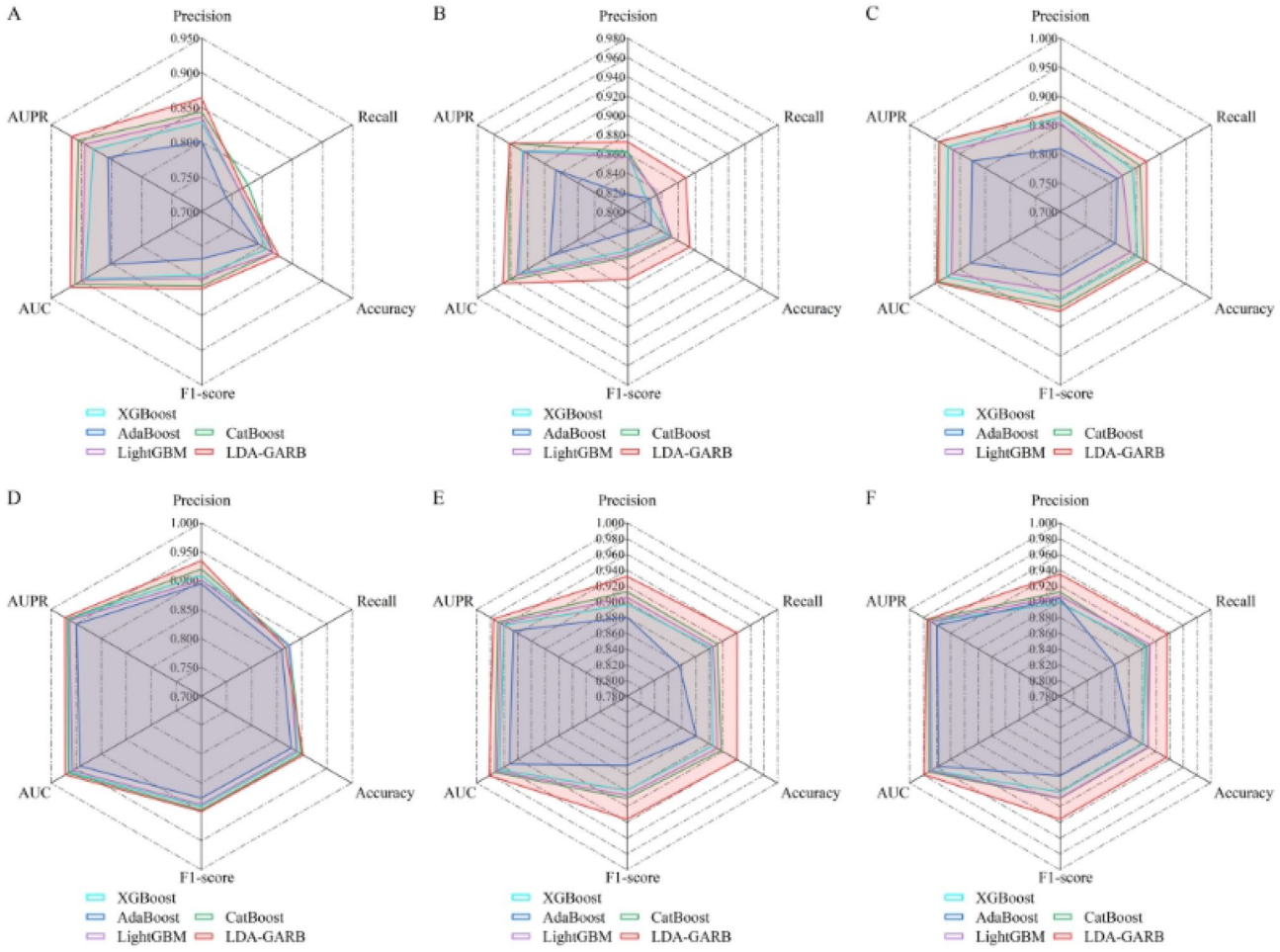

Fig. 4. Performance comparison of other four boosting algorithms under CV_1, CV_2, and CV_3. (A-C) Dataset 1. (D-F) Dataset 2.

	Dataset	LDA-LNSUBRW	GAMCLDA	LDA-VGHB	LDAGM	GANLDA	LDA-GARB
AUC	Dataset 1	0.9456 ± 0.0017	0.9495 ± 0.0042	0.9281 ± 0.0031	0.9084 ± 0.0096	0.8751 ± 0.0047	**0.9788±0.0016**
	Dataset 2	0.8794 ± 0.0039	0.9721 ± 0.0022	0.9757 ± 0.0008	0.9466 ± 0.0093	0.9425 ± 0.0033	**0.9892±0.0005**
AUPR	Dataset 1	**0.8262 ± 0.0031**	0.7111 ± 0.0146	0.6242 ± 0.0074	0.5755 ± 0.0661	0.4656 ± 0.0185	0.7743±0.0085
	Dataset 2	0.4891 ± 0.0064	0.8389 ± 0.0145	0.9116 ± 0.0021	0.8213 ± 0.0113	0.7804 ± 0.0138	**0.9336±0.0022**

Table 9. Performance comparison of different methods on imbalanced datasets under CV_3. The best performance is denoted as bold.

Predicting associated lncRNAs for CRC

CRC is one of the most frequent cancers worldwide. Recently, CRC incidences have increased rapidly in patients with age less than 50 years[57]. Thus, we want to predict potential lncRNAs for CRC. As shown in Table 22 and Fig. 7a, we predicted the top 20 lncRNAs that could associate with CRC on Dataset 1. Among the top 20 lncRNAs, 13 lncRNAs have been verified to associate with CRC. Particularly, we predicted that lncRNA CCDC26 could have an association with CRC. CCDC26 is a novel biomarker[58] and can inhibit myeloid leukemia cell[59]. Its silencing suppresses the growth and migration of glioma cells[60]. Its downregulation helps imatinib resistance in gastrointestinal stromal tumors[61].

Identifying new lncRNAs for breast cancer

Breast cancer[62] is the most frequent women cancer. It has been estimated to be 2.3 million new cases and more than 666,000 deaths in 2022[63]. During the past two decades, survival rates of breast cancer have been markedly improved, but its incidence have still risen. Thus, its effective therapy is making an essential problem.

As shown in Table 23 and Fig. 7b, we inferred the top 20 lncRNAs which could associate with breast cancer on Dataset 2. Among the 20 lncRNAs, 12 lncRNAs have been verified to have association with breast cancer.

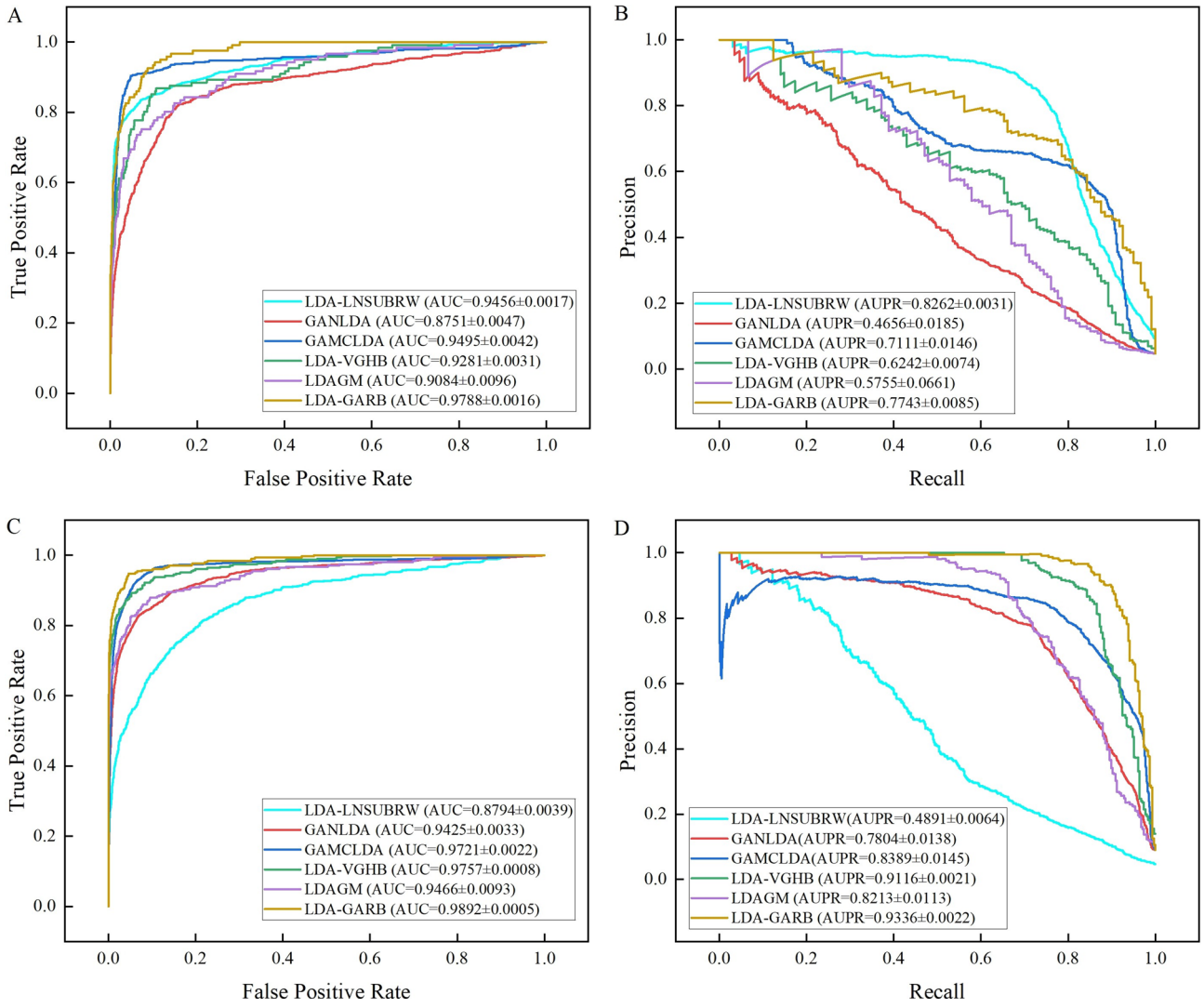

Fig. 5. Performance comparison of different methods on imbalanced datasets under CV_3. (**A**, **B**) Dataset 1, (**C**, **D**) Dataset 2..

	Dataset	$k = 64,\ N = 1$	$k = 64,\ N = 2$	$k = 64,\ N = 3$	$k = 64,\ N = 4$	$k = 64,\ N = 5$
Precision	Dataset 1	0.8636 ± 0.0450	0.8502 ± 0.0510	0.8513 ± 0.0391	0.8697 ± 0.0422	**0.8764±0.0419**
	Dataset 2	0.9344 ± 0.0147	0.9342 ± 0.0155	0.9277 ± 0.0186	0.9315 ± 0.0164	**0.9403±0.0183**
Recall	Dataset 1	0.7682 ± 0.0452	**0.7716±0.0501**	0.7690 ± 0.0570	0.7469 ± 0.0642	0.7552 ± 0.0677
	Dataset 2	0.8680 ± 0.0429	0.8633 ± 0.0439	**0.8698±0.0374**	0.8620 ± 0.0430	0.8459 ± 0.0498
Accuracy	Dataset 1	**0.8282±0.0338**	0.8241 ± 0.0298	0.8218 ± 0.0325	0.8234 ± 0.0394	0.8281 ± 0.0350
	Dataset 2	**0.9036±0.0281**	0.9010 ± 0.0264	0.9011 ± 0.0257	0.8990 ± 0.0259	0.8957 ± 0.0289
F1-score	Dataset 1	**0.8117±0.0312**	0.8077 ± 0.0394	0.8066 ± 0.0370	0.8017 ± 0.0424	0.8088 ± 0.0392
	Dataset 2	**0.8995±0.0266**	0.8967 ± 0.0244	0.8974 ± 0.0230	0.8948 ± 0.0251	0.8897 ± 0.0280
AUC	Dataset 1	**0.9180±0.0219**	0.9167 ± 0.0244	0.9144 ± 0.0249	0.8988 ± 0.0324	0.8998 ± 0.0269
	Dataset 2	**0.9716±0.0134**	0.9709 ± 0.0121	0.9714 ± 0.0123	0.9692 ± 0.0120	0.9669 ± 0.0110
AUPR	Dataset 1	**0.9160±0.0286**	0.9048 ± 0.0395	0.9119 ± 0.0303	0.9014 ± 0.0323	0.9007 ± 0.0324
	Dataset 2	0.9723 ± 0.0101	0.9706 ± 0.0096	0.9730 ± 0.0090	0.9701 ± 0.0095	0.9660 ± 0.0113

Table 10. Performance of LDA-GARB with $k = 64$ and different N on two datasets under CV_1. The best performance is denoted as bold. The second-best performance is denoted as underline.

	Dataset	$k = 64,\ N = 1$	$k = 64,\ N = 2$	$k = 64,\ N = 3$	$k = 64,\ N = 4$	$k = 64,\ N = 5$
Precision	Dataset 1	0.8724 ± 0.0365	0.8686 ± 0.0332	0.8897 ± 0.0272	0.8747 ± 0.0248	**0.8980±0.0278**
	Dataset 2	0.9321 ± 0.0277	0.9339 ± 0.0228	0.9264 ± 0.0269	**0.9381±0.0225**	0.9379 ± 0.0268
Recall	Dataset 1	0.8699 ± 0.0377	**0.8784±0.0393**	0.8764 ± 0.0317	0.8535 ± 0.0455	0.8630 ± 0.0422
	Dataset 2	0.9409 ± 0.0262	**0.9448±0.0233**	0.9411 ± 0.0237	0.9357 ± 0.0235	0.9343 ± 0.0251
Accuracy	Dataset 1	0.8744 ± 0.0255	0.8760 ± 0.0210	**0.8855±0.0221**	0.8689 ± 0.0243	0.8845 ± 0.0239
	Dataset 2	0.9409 ± 0.0158	0.9432 ± 0.0136	0.9371 ± 0.0149	**0.9409±0.0113**	0.9403 ± 0.0131
F1-score	Dataset 1	0.8707 ± 0.0316	0.8729 ± 0.0292	**0.8826±0.0227**	0.8634 ± 0.0298	0.8795 ± 0.0282
	Dataset 2	0.9363 ± 0.0243	**0.9392±0.0209**	0.9336 ± 0.0225	0.9368 ± 0.0203	0.9359 ± 0.0219
AUC	Dataset 1	0.9493 ± 0.0160	0.9516 ± 0.0137	**0.9558±0.0138**	0.9340 ± 0.0202	0.9474 ± 0.0164
	Dataset 2	0.9817 ± 0.0083	0.9824 ± 0.0063	0.9800 ± 0.0066	0.9817 ± 0.0066	**0.9828±0.0064**
AUPR	Dataset 1	0.9415 ± 0.0228	0.9468 ± 0.0219	**0.9548±0.0160**	0.9316 ± 0.0219	0.9475 ± 0.0194
	Dataset 2	0.9757 ± 0.0176	0.9774 ± 0.0146	0.9752 ± 0.0126	0.9754 ± 0.0202	**0.9793±0.0143**

Table 11. Performance of LDA-GARB with $k = 64$ and different *N* on two datasets under CV_2. The best performance is denoted as bold. The second-best performance is denoted as underline.

	Dataset	$k = 64,\ N = 1$	$k = 64,\ N = 2$	$k = 64,\ N = 3$	$k = 64,\ N = 4$	$k = 64,\ N = 5$
Precision	Dataset 1	0.8743 ± 0.0284	0.8714 ± 0.0289	0.8704 ± 0.0269	0.8711 ± 0.0274	**0.8835±0.0260**
	Dataset 2	0.9348 ± 0.0130	**0.9413±0.0132**	0.9336±0.0138	0.9327 ± 0.0125	0.9343 ± 0.0104
Recall	Dataset 1	0.8724 ± 0.0305	0.8689 ± 0.0309	0.8781 ± 0.0346	0.8767 ± 0.0292	**0.8806±0.0343**
	Dataset 2	0.9373 ± 0.0137	**0.9407±0.0132**	0.9335 ± 0.0134	0.9329 ± 0.0160	0.9365 ± 0.0147
Accuracy	Dataset 1	0.8729 ± 0.0204	0.8698 ± 0.0208	0.8733 ± 0.0225	0.8731 ± 0.0223	**0.8819±0.0225**
	Dataset 2	0.9359 ± 0.0085	**0.9409±0.0090**	0.9335 ± 0.0096	0.9327 ± 0.0094	0.9353 ± 0.0090
F1-score	Dataset 1	0.8728 ± 0.0204	0.8696 ± 0.0208	0.8737 ± 0.0231	0.8735 ± 0.0221	**0.8816±0.0231**
	Dataset 2	0.9359 ± 0.0085	**0.9409±0.0090**	0.9335 ± 0.0096	0.9327 ± 0.0095	0.9353 ± 0.0091
AUC	Dataset 1	0.9459 ± 0.0109	0.9462 ± 0.0133	**0.9479±0.0134**	0.9367 ± 0.0174	0.9409 ± 0.0174
	Dataset 2	0.9790 ± 0.0051	**0.9824±0.0034**	0.9801 ± 0.0046	0.9790 ± 0.0048	0.9798 ± 0.0050
AUPR	Dataset 1	0.9418 ± 0.0136	**0.9449±0.0161**	0.9432 ± 0.0192	0.9331 ± 0.0206	0.9371 ± 0.0214
	Dataset 2	0.9744 ± 0.0100	**0.9818±0.0040**	0.9787 ± 0.0061	0.9781 ± 0.0060	0.9790 ± 0.0065

Table 12. Performance of LDA-GARB with $k = 64$ and different *N* on two datasets under CV_3. The best performance is denoted as bold. The second-best performance is denoted as underline.

	Dataset	$k = 128,\ N = 1$	$k = 128,\ N = 2$	$k = 128,\ N = 3$	$k = 128,\ N = 4$	$k = 128,\ N = 5$
Precision	Dataset 1	0.8528 ± 0.0432	0.8543 ± 0.0434	0.8569 ± 0.0490	0.8432 ± 0.0453	**0.8635±0.0394**
	Dataset 2	0.9259 ± 0.0197	0.9200 ± 0.0246	0.9261 ± 0.0191	0.9188 ± 0.0225	**0.9325±0.0183**
Recall	Dataset 1	**0.7631±0.0585**	0.7520 ± 0.0673	0.7605 ± 0.0621	0.7385 ± 0.0714	0.7320 ± 0.0684
	Dataset 2	0.8579 ± 0.0434	**0.8624±0.0400**	0.8607 ± 0.0456	0.8527 ± 0.0475	0.8448 ± 0.0466
Accuracy	Dataset 1	0.8207 ± 0.0349	0.8180 ± 0.0400	**0.8234±0.0341**	0.8090 ± 0.0367	0.8122 ± 0.0374
	Dataset 2	0.8944 ± 0.0268	0.8933 ± 0.0269	**0.8959±0.0270**	0.8887 ± 0.0313	0.8915 ± 0.0270
F1-score	Dataset 1	**0.8041±0.0415**	0.7980 ± 0.0456	0.8038 ± 0.0412	0.7855 ± 0.0499	0.7902 ± 0.0444
	Dataset 2	0.8899 ± 0.0252	0.8896 ± 0.0236	**0.8914±0.0261**	0.8839 ± 0.0309	0.8856 ± 0.0258
AUC	Dataset 1	**0.9128±0.0271**	0.9090 ± 0.0317	0.9116 ± 0.0249	0.8947 ± 0.0268	0.8733 ± 0.0389
	Dataset 2	0.9635 ± 0.0164	0.9618 ± 0.0163	**0.9654±0.0158**	0.9587 ± 0.0178	0.9600 ± 0.0148
AUPR	Dataset 1	**0.9132±0.0362**	0.9093 ± 0.0395	0.9064 ± 0.0376	0.8877 ± 0.0469	0.8888 ± 0.0348
	Dataset 2	0.9647 ± 0.0128	0.9637 ± 0.0125	**0.9677±0.0116**	0.9603 ± 0.0152	0.9628 ± 0.0106

Table 13. Performance of LDA-GARB with $k = 128$ and different *N* on two datasets under CV_1. The best performance is denoted as bold. The second-best performance is denoted as underline.

	Dataset	$k = 128,\ N = 1$	$k = 128,\ N = 2$	$k = 128,\ N = 3$	$k = 128,\ N = 4$	$k = 128,\ N = 5$
Precision	Dataset 1	0.8729 ± 0.0380	0.8603 ± 0.0341	0.8514 ± 0.0324	0.8709 ± 0.0361	**0.8808±0.0320**
	Dataset 2	0.9270 ± 0.0262	0.9290 ± 0.0249	0.9329 ± 0.0266	0.9354 ± 0.0276	**0.9362±0.0272**
Recall	Dataset 1	0.8720 ± 0.0431	0.8617 ± 0.0349	0.8620 ± 0.0340	**0.8722±0.0385**	0.8537 ± 0.0414
	Dataset 2	0.9371 ± 0.0228	0.9384 ± 0.0281	**0.9398±0.0261**	0.9328 ± 0.028	0.9357 ± 0.0256
Accuracy	Dataset 1	**0.8766±0.0239**	0.8621 ± 0.0252	0.8575 ± 0.0231	0.8749 ± 0.0238	0.8708 ± 0.0236
	Dataset 2	0.9359 ± 0.0137	0.9379 ± 0.0163	**0.9402±0.0145**	0.9394 ± 0.0153	0.9397 ± 0.0159
F1-score	Dataset 1	**0.8720±0.0348**	0.8604 ± 0.0258	0.8561 ± 0.0253	0.8709 ± 0.0291	0.8663 ± 0.0263
	Dataset 2	0.9319 ± 0.0224	0.9335 ± 0.0239	**0.9362±0.0243**	0.9339 ± 0.0251	0.9358 ± 0.0233
AUC	Dataset 1	**0.9505±0.0159**	0.9423 ± 0.0153	0.9332 ± 0.0171	0.9436 ± 0.0161	0.9330 ± 0.0193
	Dataset 2	**0.9817±0.0067**	0.9813 ± 0.0074	0.9804 ± 0.0076	0.9815 ± 0.0070	0.9799 ± 0.0079
AUPR	Dataset 1	**0.9405±0.0299**	0.9363 ± 0.0199	0.9267 ± 0.0230	0.9346 ± 0.0241	0.9338 ± 0.0205
	Dataset 2	0.9776 ± 0.0146	0.9777 ± 0.0151	0.9746 ± 0.0200	**0.9783±0.0156**	0.9770 ± 0.0158

Table 14. Performance of LDA-GARB with $k = 128$ and different N on two datasets under CV_2. The best performance is denoted as bold. The second-best performance is denoted as underline.

	Dataset	$k = 128,\ N = 1$	$k = 128,\ N = 2$	$k = 128,\ N = 3$	$k = 128,\ N = 4$	$k = 128,\ N = 5$
Precision	Dataset 1	0.8631 ± 0.0282	0.8631 ± 0.0234	0.8626 ± 0.0267	0.8584 ± 0.0291	**0.8733±0.0285**
	Dataset 2	0.9329 ± 0.0117	0.9367 ± 0.0127	**0.9293±0.0129**	0.9310 ± 0.0130	0.9375 ± 0.0140
Recall	Dataset 1	0.8694 ± 0.0326	0.8653 ± 0.0309	0.8583 ± 0.0336	0.8620 ± 0.0309	**0.8750±0.0328**
	Dataset 2	0.9329 ± 0.0169	**0.9378±0.0131**	0.9283 ± 0.0135	0.9303 ± 0.0154	0.9357 ± 0.0147
Accuracy	Dataset 1	0.8653 ± 0.0228	0.8636 ± 0.0190	0.8603 ± 0.0212	0.8593 ± 0.0213	**0.8735±0.0213**
	Dataset 2	0.9328 ± 0.0101	**0.9371±0.0091**	0.9287 ± 0.0091	0.9306 ± 0.0093	0.9366 ± 0.0089
F1-score	Dataset 1	0.8658 ± 0.0229	0.8637 ± 0.0196	0.8599 ± 0.0218	0.8596 ± 0.0211	**0.8736±0.0216**
	Dataset 2	0.9328 ± 0.0103	**0.9372±0.0091**	0.9287 ± 0.0091	0.9306 ± 0.0094	0.9365 ± 0.0089
AUC	Dataset 1	0.9356 ± 0.0159	0.9370 ± 0.0134	0.9345 ± 0.0140	**0.9381±0.0146**	0.9317 ± 0.0152
	Dataset 2	0.9812 ± 0.0049	**0.9817±0.0038**	0.9789 ± 0.0042	0.9765 ± 0.0048	0.9790 ± 0.0047
AUPR	Dataset 1	0.9258 ± 0.0222	0.9355 ± 0.0146	0.9300 ± 0.0168	**0.9359±0.0184**	0.9333 ± 0.0172
	Dataset 2	0.9812 ± 0.0056	**0.9816±0.0041**	0.9796 ± 0.0045	0.9749 ± 0.0065	0.9799 ± 0.0050

Table 15. Performance of LDA-GARB with $k = 128$ and different N on two datasets under CV_3. The best performance is denoted as bold. The second-best performance is denoted as underline.

	Dataset	$k = 256,\ N = 1$	$k = 256,\ N = 2$	$k = 256,\ N = 3$	$k = 256,\ N = 4$	$k = 256,\ N = 5$
Precision	Dataset 1	0.8466 ± 0.0399	0.8486 ± 0.0543	0.8447 ± 0.0507	0.8307 ± 0.0485	**0.8657±0.0417**
	Dataset 2	0.9340 ± 0.0154	0.9323 ± 0.0173	0.9280 ± 0.0177	0.9273 ± 0.0202	**0.9364±0.0176**
Recall	Dataset 1	0.7521 ± 0.0575	**0.7587±0.0614**	0.7369 ± 0.0708	0.7404 ± 0.0609	0.7334 ± 0.0612
	Dataset 2	0.8648 ± 0.0421	**0.8672±0.0385**	0.8632 ± 0.0467	0.8449 ± 0.0559	0.8489 ± 0.0534
Accuracy	Dataset 1	0.8123 ± 0.0371	**0.8205±0.0307**	0.8087 ± 0.0394	0.8012 ± 0.0373	0.8165 ± 0.0382
	Dataset 2	0.9018 ± 0.0277	**0.9019±0.0269**	0.8980 ± 0.0301	0.8890 ± 0.0354	0.8952 ± 0.0295
F1-score	Dataset 1	0.7947 ± 0.0344	**0.7996±0.0484**	0.7850 ± 0.0499	0.7809 ± 0.0406	0.7924 ± 0.0416
	Dataset 2	0.8976 ± 0.0260	0.8982 ± 0.0244	0.8938 ± 0.0287	0.8832 ± 0.0342	**0.8894±0.0290**
AUC	Dataset 1	0.9012 ± 0.0307	**0.9163±0.0225**	0.8999 ± 0.0300	0.8931 ± 0.0284	0.8883 ± 0.0307
	Dataset 2	0.9648 ± 0.0172	0.9653 ± 0.0135	**0.9665±0.0157**	0.9566 ± 0.0237	0.9601 ± 0.0163
AUPR	Dataset 1	0.9019 ± 0.0244	**0.9117±0.0414**	0.8992 ± 0.0487	0.8949 ± 0.0365	0.8934 ± 0.0389
	Dataset 2	0.9673 ± 0.0130	0.9684 ± 0.0093	**0.9700±0.0109**	0.9630 ± 0.0154	0.9649 ± 0.0118

Table 16. Performance of LDA-GARB with $k = 256$ and different N on two datasets under CV_1. The best performance is denoted as bold. The second-best performance is denoted as underline.

	Dataset	$k = 256,\ N = 1$	$k = 256,\ N = 2$	$k = 256,\ N = 3$	$k = 256,\ N = 4$	$k = 256,\ N = 5$
Precision	Dataset 1	0.8708 ± 0.0298	0.8697 ± 0.0356	0.8805 ± 0.0341	0.8618 ± 0.0363	**0.8866±0.0332**
	Dataset 2	0.9344 ± 0.0260	0.9241 ± 0.0272	0.9234 ± 0.0329	0.9226 ± 0.0323	**0.9459±0.0269**
Recall	Dataset 1	0.8693 ± 0.0389	0.8560 ± 0.0354	0.8710 ± 0.0370	**0.8713±0.0350**	0.8586 ± 0.0408
	Dataset 2	**0.9460±0.0219**	0.9346 ± 0.0240	0.9347 ± 0.0330	0.9297 ± 0.0265	0.9337 ± 0.0300
Accuracy	Dataset 1	0.8729 ± 0.0239	0.8670 ± 0.0213	0.8778 ± 0.0229	0.8698 ± 0.0195	**0.8786±0.0199**
	Dataset 2	0.9434 ± 0.0145	0.9326 ± 0.0161	0.9352 ± 0.0162	0.9309 ± 0.0167	**0.9444±0.0147**
F1-score	Dataset 1	0.8696 ± 0.0283	0.8623 ± 0.0286	**0.8751±0.0261**	0.8660 ± 0.0284	0.8717 ± 0.0293
	Dataset 2	**0.9400±0.0213**	0.9292 ± 0.0235	0.9288 ± 0.0308	0.9260 ± 0.0269	0.9396 ± 0.0251
AUC	Dataset 1	0.9477 ± 0.0150	0.9479 ± 0.0139	**0.9484±0.0142**	0.9433 ± 0.0135	0.9350 ± 0.0149
	Dataset 2	**0.9838±0.0063**	0.9796 ± 0.0075	0.9802 ± 0.0078	0.9761 ± 0.0085	0.9798 ± 0.0077
AUPR	Dataset 1	0.9423 ± 0.0197	**0.9471±0.0199**	0.9432 ± 0.0193	0.9387 ± 0.0240	0.8934 ± 0.0389
	Dataset 2	**0.9807±0.0121**	0.9759 ± 0.0140	0.9745 ± 0.0208	0.9712 ± 0.0199	0.9766 ± 0.0165

Table 17. Performance of LDA-GARB with $k = 256$ and different N on two datasets under CV_2. The best performance is denoted as bold. The second-best performance is denoted as underline.

	Dataset	$k = 256,\ N = 1$	$k = 256,\ N = 2$	$k = 256,\ N = 3$	$k = 256,\ N = 4$	$k = 256,\ N = 5$
Precision	Dataset 1	0.8683 ± 0.0265	0.8687 ± 0.0229	0.8591 ± 0.027	0.8609 ± 0.0251	**0.8816±0.0254**
	Dataset 2	**0.9354±0.0121**	0.9238 ± 0.0134	0.9230 ± 0.0141	0.9213 ± 0.0138	0.9349 ± 0.0119
Recall	Dataset 1	0.8744 ± 0.0345	0.8593 ± 0.0336	0.8645 ± 0.0293	0.8687 ± 0.0316	**0.8868±0.0301**
	Dataset 2	0.9297 ± 0.0148	0.9271 ± 0.0146	0.9262 ± 0.0166	0.9196 ± 0.0155	**0.9321±0.0156**
Accuracy	Dataset 1	0.8704 ± 0.0211	0.8643 ± 0.0186	0.8610 ± 0.0211	0.8638 ± 0.0210	**0.8834±0.0190**
	Dataset 2	0.9327 ± 0.0103	0.9252 ± 0.0097	0.9244 ± 0.0113	0.9204 ± 0.0103	**0.9335±0.0092**
F1-score	Dataset 1	0.8708 ± 0.0217	0.8634 ± 0.0198	0.8614 ± 0.0211	0.8643 ± 0.0214	**0.8837±0.0192**
	Dataset 2	0.9324 ± 0.0105	0.9254 ± 0.0097	0.9245 ± 0.0114	0.9203 ± 0.0104	**0.9334±0.0094**
AUC	Dataset 1	0.9410 ± 0.0128	0.9392 ± 0.0128	0.9390 ± 0.0140	0.9374 ± 0.0133	**0.9413±0.0151**
	Dataset 2	**0.9810±0.0042**	0.9774 ± 0.0044	0.9768 ± 0.0054	0.9735 ± 0.0048	0.9781 ± 0.0054
AUPR	Dataset 1	0.9316 ± 0.0180	0.9335 ± 0.0162	**0.9402±0.0143**	0.9374 ± 0.0159	0.9387 ± 0.0195
	Dataset 2	**0.9811±0.0049**	0.9776 ± 0.0047	0.9776 ± 0.0056	0.9743 ± 0.0053	0.9799 ± 0.0049

Table 18. Performance of LDA-GARB with $k = 256$ and different N on two datasets under CV_3. The best performance is denoted as bold. The second-best performance is denoted as underline.

	Dataset	Linear feature	Nonlinear feature	LDA-GARB
Precision	Dataset 1	**0.8708 ± 0.0479**	0.8174 ± 0.0598	0.8636 ± 0.0450
	Dataset 2	0.9321 ± 0.0182	0.9144 ± 0.0180	**0.9344 ± 0.0147**
Recall	Dataset 1	0.7507 ± 0.0583	**0.8004 ± 0.0581**	0.7682 ± 0.0452
	Dataset 2	0.8430 ± 0.0566	**0.8703 ± 0.0347**	0.8680 ± 0.0429
Accuracy	Dataset 1	0.8248 ± 0.0387	0.8223 ± 0.0319	**0.8282 ± 0.0338**
	Dataset 2	0.8906 ± 0.0290	0.8946 ± 0.0232	**0.9036 ± 0.0281**
F1-score	Dataset 1	0.8046 ± 0.0413	0.8070 ± 0.0463	**0.8117 ± 0.0312**
	Dataset 2	0.8841 ± 0.0320	0.8914 ± 0.0212	**0.8995 ± 0.0266**
AUC	Dataset 1	0.8960 ± 0.0327	0.9084 ± 0.0240	**0.9180 ± 0.0219**
	Dataset 2	0.9576 ± 0.0169	0.9637 ± 0.0147	**0.9716 ± 0.0134**
AUPR	Dataset 1	0.9005 ± 0.0350	0.8768 ± 0.0579	**0.9160 ± 0.0286**
	Dataset 2	0.9569 ± 0.0150	0.9664 ± 0.0106	**0.9723 ± 0.0101**

Table 19. Performance when using different feature selection methods under CV_1. The best performance is denoted as bold.

	Dataset	Linear feature	Nonlinear feature	LDA-GARB
Precision	Dataset 1	**0.9040 ± 0.0261**	0.8464 ± 0.0409	0.8724 ± 0.0365
	Dataset 2	0.9251 ± 0.0318	0.8944 ± 0.0473	**0.9321 ± 0.0277**
Recall	Dataset 1	0.8621 ± 0.0411	0.8633 ± 0.0444	**0.8699 ± 0.0377**
	Dataset 2	0.9351 ± 0.0343	0.8929 ± 0.0537	**0.9409 ± 0.0262**
Accuracy	Dataset 1	**0.8876 ± 0.0214**	0.8564 ± 0.0260	0.8744 ± 0.0255
	Dataset 2	0.9343 ± 0.0164	0.9044 ± 0.0209	**0.9409 ± 0.0158**
F1-score	Dataset 1	**0.8820 ± 0.0278**	0.8541 ± 0.0347	0.8707 ± 0.0316
	Dataset 2	0.9297 ± 0.0288	0.8934 ± 0.0489	**0.9363 ± 0.0243**
AUC	Dataset 1	0.9459 ± 0.0139	0.9266 ± 0.0193	**0.9493 ± 0.0160**
	Dataset 2	0.9745 ± 0.0108	0.9679 ± 0.0138	**0.9817 ± 0.0083**
AUPR	Dataset 1	**0.9467 ± 0.0205**	0.9047 ± 0.0370	0.9415 ± 0.0228
	Dataset 2	0.9646 ± 0.0244	0.9578 ± 0.0419	**0.9757 ± 0.0176**

Table 20. Performance when using different feature selection methods under CV_2. The best performance is denoted as bold.

	Dataset	Linear feature	Nonlinear feature	LDA-GARB
Precision	Dataset 1	0.8714 ± 0.0283	0.8361 ± 0.0271	**0.8743 ± 0.0284**
	Dataset 2	0.9328 ± 0.0119	0.9082 ± 0.0177	**0.9348 ± 0.0130**
Recall	Dataset 1	**0.8731 ± 0.0301**	0.8555 ± 0.0286	0.8724 ± 0.0305
	Dataset 2	0.9330 ± 0.0149	0.8994 ± 0.0180	**0.9373 ± 0.0137**
Accuracy	Dataset 1	0.8717 ± 0.0227	0.8433 ± 0.0197	**0.8729 ± 0.0204**
	Dataset 2	0.9328 ± 0.0095	0.9041 ± 0.0121	**0.9359 ± 0.0085**
F1-score	Dataset 1	0.8719 ± 0.0227	0.8452 ± 0.0193	**0.8728 ± 0.0204**
	Dataset 2	0.9328 ± 0.0097	0.9036 ± 0.0122	**0.9359 ± 0.0085**
AUC	Dataset 1	0.9315 ± 0.0159	0.9256 ± 0.0143	**0.9459 ± 0.0109**
	Dataset 2	0.9782 ± 0.0049	0.9710 ± 0.0057	**0.9790 ± 0.0051**
AUPR	Dataset 1	0.9257 ± 0.0224	0.9186 ± 0.0204	**0.9418 ± 0.0136**
	Dataset 2	**0.9762 ± 0.0063**	0.9712 ± 0.0061	0.9744 ± 0.0100

Table 21. Performance when using different feature selection methods under CV_3. The best performance is denoted as bold.

Based on the rankings in Table 23 and Fig. 7, we predicted that HAR1A may be associated with breast cancer. HAR1A can inhibit non-small cell lung cancer progression[64], regulate oral cancer development[65], and affect brain development[66]. The association between breast cancer and HAR1A needs further validation.

Discussion

lncRNAs are closely associated with many important physiological processes and have been regarded as potential biomarkers of cancers. Identifying potential LDAs promotes us to better understand complex molecular mechanisms of human diseases, find new biomarkers, and further facilitate disease diagnosis and therapy.

In this manuscript, we proposed a computational framework called LDA-GARB for LDA prediction. LDA-GARB first calculated disease similarity based on their semantic features and GAPK, and lncRNA similarity based on their functional information and GAPK. Subsequently, LDA-GARB extracted linear features through NMF and their nonlinear features via similarity matrices and GAE for lncRNAs and diseases. Finally, LDA-GARB took the extracted features as inputs and designed a noise-robust gradient boosting model to decipher potential associations from unknown lncRNA-disease pairs.

To ascertain the LDA-GARB performance, we conducted multiple comparison experiments. First, LDA-GARB was compared with four representative LDA prediction methods under three distinct CVs. These four methods include SDLDA, LDNFSGB, LDAenDL, and LDA-VGHB. LDNFSGB and LDA-VGHB are two boosting-based LDA classification models. SDLDA and LDAenDL are two deep learning-based LDA inference algorithms. LDA-GARB obviously outperformed the two boosting-based models and the two deep learning-based methods, demonstrating its better LDA inference accuracy and feature learning ability.

Next, LDA-GARB was compared with four boosting models under the three CVs, i.e., XGBoost, AdaBoost, CatBoost, and LightGBM. XGBoost is a scalable end-to-end extreme gradient boosting system. AdaBoost performs highly accurate prediction by integrating multiple weak rules. CatBoost is a unbiased model for categorical feature learning. LightGBM is a highly efficient GBDT. The four methods are classical and wide-used boosting algorithms. LDA-GARB surpassed the four models, elucidating the powerful LDA classification ability of the noise-robust gradient boosting model.

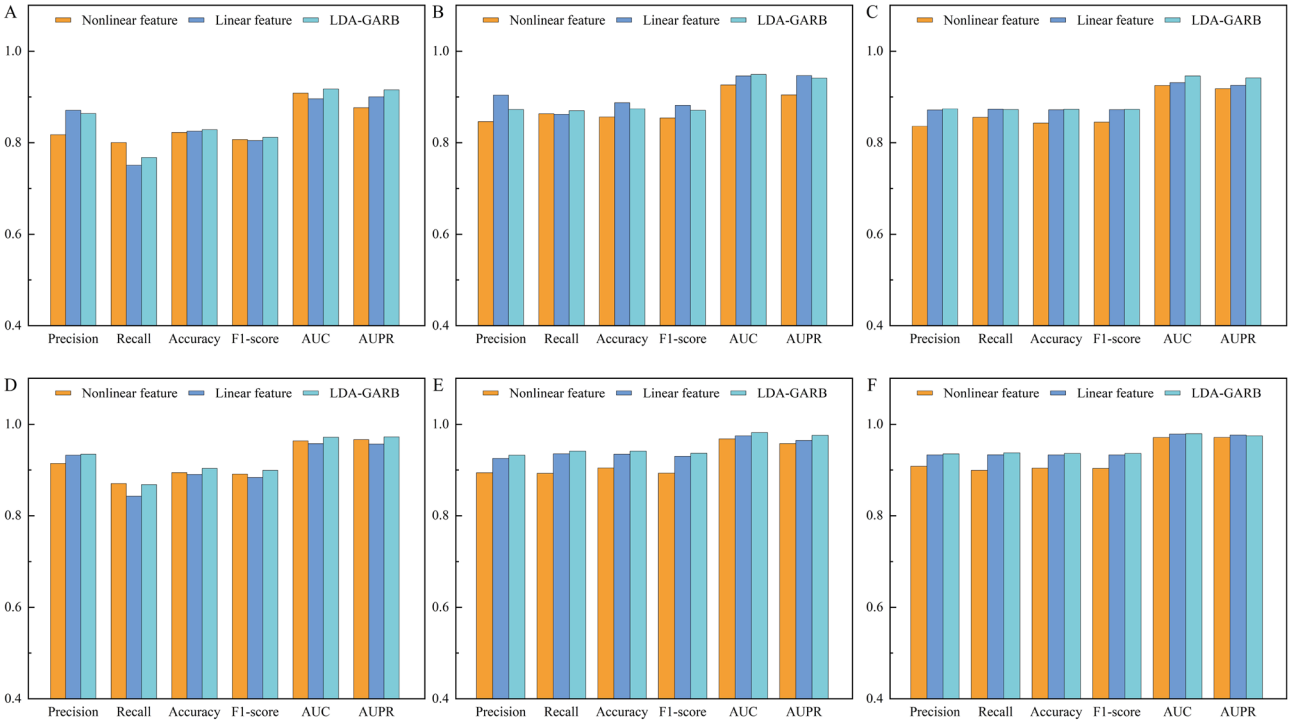

Fig. 6. Performance when using different feature selection methods under CV_1, CV_2, and CV_3. (**A**–**C**) on Dataset 1. (**D**–**F**) on Dataset 2.

Rank	lncRNA	Evidence
1	BCYRN1	Lnc2Cancer3.0, RNADisease, LncRNADisease v3.0
2	CCAT2	Lnc2Cancer3.0, RNADisease, LncRNADisease v3.0
3	HIF1A-AS1	Lnc2Cancer3.0, RNADisease
4	KCNQ1DN	Lnc2Cancer3.0, RNADisease
5	IGF2-AS	RNADisease
6	CCDC26	Unknown
7	ADAMTS9-AS2	Lnc2Cancer3.0, RNADisease, LncRNADisease v3.0
8	MINA	Unknown
9	PTENP1	RNADisease
10	HCP5	Lnc2Cancer3.0, RNADisease, LncRNADisease v3.0
11	WT1-AS	Lnc2Cancer3.0, RNADisease
12	HIF1A-AS2	Lnc2Cancer3.0, RNADisease
13	DNM3OS	Unknown
14	MIR31HG	Lnc2Cancer3.0, RNADisease, LncRNADisease v3.0
15	LINC00032	Unknown
16	TCL6	Unknown
17	MIR155HG	Lnc2Cancer3.0, RNADisease, LncRNADisease v3.0
18	IFNG-AS1	Unknown
19	CBR3-AS1	Lnc2Cancer3.0, RNADisease, LncRNADisease v3.0
20	PCA3	Unknown

Table 22. The predicted top 20 lncRNAs associated with CRC on Dataset 1.

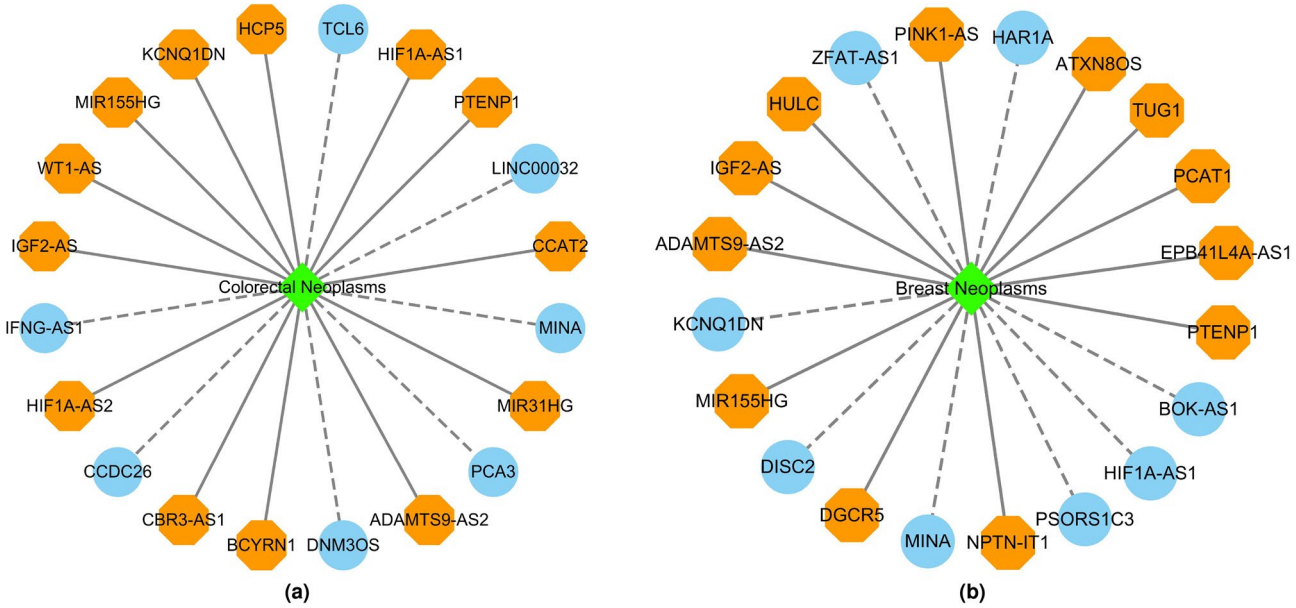

Fig. 7. (**a**) The top potential 20 lncRNAs for CRC on Dataset 1. (**b**) The top 20 potential lncRNAs for breast cancer on Dataset 2. Solid line denotes an LDA that can be predicted and validated. Dashed line denote an LDA that can be predicted but not validated, respectively..

Input: LDA dataset $D = \{(\boldsymbol{x}_i, \mathbf{y}_i')\}_{i=1}^{N}$, $\boldsymbol{x}_i$ and $\mathbf{y}_i'$ denote the i-th training sample and its label.

Output: lncRNA-disease association probability $\mathbf{y}_i' = f_T(\boldsymbol{x}_i)$

1: **Initializing** $f_0(\boldsymbol{x}_i) = 0$

2: **For** $t = 1, ..., T$ do:

3: Calculating gradient g_i^t and Hessian h_i^t

4: Calculating the optimal leaf weight $w_j^* = \frac{-\sum_{i \in I_j} g_i^t}{\sum_{i \in I_j} h_i^t + \lambda}$ and objective $\widetilde{\mathscr{L}_j^*} = -\frac{1}{2} \frac{(\sum_{i \in I_j} g_i^t)^2}{\sum_{i \in I_j} h_i^t + \lambda}$ of decision tree

5: Calculating $gain = \frac{1}{2}\left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda}\right]$ to obtain the best split feature and corresponding split value of the tree

6: Updating $f_t(\boldsymbol{x}_i) = f_{t-1}(\boldsymbol{x}_i) + \alpha f_t(\boldsymbol{x}_i)$

7: **End For**

Algorithm 1 LDA classification algorithm.

After that, to analyze the performance of LDA-GARB on imbalanced data, it was compared with LDA-LNSUBRW, GAMCLDA, LDA-VGHB, LDAGM, and GANLDA. The five methods utilized unbalanced bi-random walk, graph autoencoder matrix completion, heterogeneous Newton boosting machine, multi-view heterogeneous network, and graph attention network. LDA-GARB outperformed the five model and exhibited better classification ability on imbalanced datasets. Finally, to discern the effect of the proposed feature extraction techniques on predictions, LDA-GARB also conducted ablation experiments. The outcomes indicated that the combination of NMF-based linear feature extraction and GAE-based nonlinear feature extraction improved LDA prediction.

CRC and breast cancer are two of the most frequent cancers worldwide. After determining the performance, LDA-GARB was applied to predict possible lncRNAs for CRC and breast cancer. The results showed that CRC could have an association with lncRNA CCDC26 and breast cancer may be associated with HAR1A. The above results provided new potential biomarker for CRC and breast cancer.

LDA-GARB demonstrated two main advantages when deciphering possible LDAs. (i) It could effectively reduce the effect of label noises on predictions. Machine learning requires negative samples when predicting LDAs. However, current data resources do not provide negative samples due to experiment limitations. Thus, most machine learning-based LDA prediction models had to obtain negative associations from unlabeled

Rank	lncRNA	Evidence
1	DGCR5	RNADisease, LncRNADisease v3.0
2	HAR1A	Unknown
3	TUG1	Lnc2Cancer3.0, RNADisease, LncRNADisease v3.0
4	PINK1-AS	RNADisease
5	IGF2-AS	RNADisease, LncRNADisease v3.0
6	MIR155HG	RNADisease
7	NPTN-IT1	Lnc2Cancer3.0, RNADisease
8	BOK-AS1	Unknown
9	PCAT1	Lnc2Cancer3.0, RNADisease, LncRNADisease v3.0
10	PTENP1	Lnc2Cancer3.0, RNADisease, LncRNADisease v3.0
11	HULC	Lnc2Cancer3.0, RNADisease, LncRNADisease v3.0
12	KCNQ1DN	Unknown
13	MINA	Unknown
14	HIF1A-AS1	Unknown
15	EPB41L4A-AS1	Lnc2Cancer3.0, RNADisease
16	ATXN8OS	Lnc2Cancer3.0, RNADisease, LncRNADisease v3.0
17	ZFAT-AS1	Unknown
18	ADAMTS9-AS2	Lnc2Cancer3.0, RNADisease, LncRNADisease v3.0
19	DISC2	Unknown
20	PSORS1C3	Unknown

Table 23. The predicted top 20 lncRNAs associated with breast cancer in the dataset 2.

lncRNA-disease pairs through random selection. However, these unlabeled lncRNA-disease pairs may contain a handle of positive LDAs, which causes label noises and severely affects the model performance. To solve the issue, LDA-GARB adopted a noise-robust gradient boosting algorithm to alleviate the effect of label noises on LDA prediction. (ii) It was more appropriate to solve imbalanced LDA datasets. Current LDA datasets are imbalanced while existing boosting models are limited to imbalanced datasets. To address this issue, LDA-GARB used non-convex loss function and exhibited the powerful adaptability on imbalanced LDA datasets.

Although LDA-GARB calculated better predictions, it remains limitations. During LDA prediction, we need learn feature vectors of lncRNAs and diseases from their biological information. However, several diseases have no directed acyclic graphs, resulting in that we can't compute their biological similarity based on their MESH descriptors. So, we had to use GAPK to measure disease similarity and extract their feature vectors from GAPK similarity matrix, which have been widely applied to various disease-related association prediction. But GAPK similarity was computed based on association information, which may cause data leakage during test. The data leakage is a common issue in disease-related association tasks and is urgent to solve. Text mining techniques can effectively capture information hidden in unstructured text data. As a result, in the future, we will design a text mining algorithm to obtain semantic features for all diseases especially diseases without directed acyclic graph from diverse health and medical literatures. By doing so, we can effectively avoid data leakage during LDA prediction and boost the performance of various association prediction models.

Conclusion

In this manuscript, we presented a computational model called LDA-GARB for identifying potential LDAs by integrating NMF, GAE, and the noise-robust gradient boosting model. Compared to four state-of-the-art LDA identification methods, four classical boosting models, and five imbalanced data solution algorithms, LDA-GARB computed better predictions under three distinct CVs (i.e., CV_1, CV_2, and CV_3). Moreover, LDA-GARB inferred that lncRNAs CCDC 26 and HAR1A could separately associate with CRC and breast cancer and may be their biomarkers, which provided clues of treatment for the two cancers. As a useful computational tool for identifying potential lncRNAs for human diseases, we anticipate that LDA-GARB can help to find new biomarkers for various complex diseases and further promote their diagnosis and therapy.

Materials and methods

LDA-GARB mainly contain two procedures: (i) LDA feature extraction. First, LDA-GARB employs NMF and LDA information to extract linear features of each lncRNA and disease. Next, LDA-GARB computes disease similarity based on semantic features and Gaussian association profile kernel (GAPK) similarity, and lncRNA similarity based on functional information and GAPK similarity. By combining disease similarity and lncRNA similarity, LDA-GARB proposes a GAE model to extract nonlinear features of lncRNAs and diseases. And the

extracted linear and nonlinear features of lncRNA are concatenated as a vector to depict the lncRNA. Similarly, the learned linear and nonlinear features of one disease are concatenated as a vector to depict the disease. And the concatenation of lncRNA features and disease features is used to character each lncRNA-disease pair. (ii) LDA classification. LDA-GARB takes the obtained feature vector as input and devises a noise-robust gradient boosting to perform predictions.

Linear feature extraction based on NMF

NMF can effectively reduce feature dimensionality by combining the non-negativity constraint[42,67]. Here, we adopt NMF to learn linear representations of lncRNAs and diseases. First, we decompose an LDA matrix $\boldsymbol{Y}$ into two low-rank matrices $\boldsymbol{U} = R^{n\times s}$ and $\boldsymbol{V} = R^{s\times m}$. Next, to make $\boldsymbol{U}$ and $\boldsymbol{V}$ more smooth, we add a weighted matrix $\boldsymbol{W} \in R^{n\times m}$ and perform L_2 regularization. Thus, we build an objective function with regularization parameters λ_1 and λ_2 to learn lncRNA linear features $\boldsymbol{U}$ and disease linear features $\boldsymbol{V}$ by Eq. (1):

$$\min_{\boldsymbol{U}\geq 0,\boldsymbol{V}\geq 0} \|\boldsymbol{W}\odot(\boldsymbol{Y}-\boldsymbol{U}V)\|_F^2+\lambda_1\|\boldsymbol{U}\|_F^2+\lambda_2\|\boldsymbol{V}\|_F^2 \tag{1}$$

where $\odot$ is the Hadamard product, $\boldsymbol{U}\geq 0$, and $\boldsymbol{V}\geq 0$.

Nonlinear feature extraction based on GAE and similarity computation

Similarity computation

To extract nonlinear features of lncRNAs and diseases based on GAE, we need to compute disease similarity and lncRNA similarity. First, we compute disease semantic similarity $\boldsymbol{S}_d^{sem}$ using the IDSSIM method[68] based on MeSH descriptors. Since the MeSH database does not provide directed acyclic graph for some diseases, we can't measure their similarity according to their directed acyclic graphs. Thus, we adopt GAPK to compute their similarity. Specially, for diseases d_i and d_j, their Gaussian association profiles are represented as $\boldsymbol{Y}_{.i}$(the i-th column of $\boldsymbol{Y}$) and $\boldsymbol{Y}_{.j}$(the j-th column of $\boldsymbol{Y}$), respectively. And their GAPK similarity is defined by Eq. (2):

$$\begin{aligned}\boldsymbol{G}_d(i,j)&=\exp\left(-\theta_d\left\|\boldsymbol{Y}_{.i}-\boldsymbol{Y}_{.j}\right\|^2\right)\\ \theta_d&=\frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{Y}_{.i}\|^2\end{aligned} \tag{2}$$

Consequently, disease similarity matrix $\boldsymbol{D}$ is built by integrating their semantic similarity and GAPK similarity by Eq. (3):

$$\mathrm{D}(i,j)=\begin{cases}\frac{\mathrm{S}_d^{sem}(i,j)+\mathrm{G}_d(i,j)}{2} & \text{if } \mathrm{S}_d^{sem}(i,j)\neq 0\\ \mathrm{G}_d(i,j) & \text{otherwise}\end{cases} \tag{3}$$

lncRNA similarity is computed based on their functional similarity and GAPK similarity[2,42]. lncRNA functional similarity $\boldsymbol{S}_l^{fun}$ is measured through the IDSSIM method[68] based on $\boldsymbol{S}_d^{sem}$. lncRNA GAPK similarity matrix $\boldsymbol{G}_l$ is calculated by Eq. (4):

$$\begin{aligned}\boldsymbol{G}_l(i,j)&=\exp\left(-\theta_l\left\|\boldsymbol{Y}_{i.}-\boldsymbol{Y}_{j.}\right\|^2\right)\\ \theta_l&=\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{Y}_{i.}\|^2\end{aligned} \tag{4}$$

where $\boldsymbol{Y}_{i.}$ and $\boldsymbol{Y}_{j.}$ denote the i-th and j-th rows of $\boldsymbol{Y}$, respectively.

Consequently, lncRNA similarity matrix matrix $\boldsymbol{L}$ is built by incorporating their functional similarity and GAPK similarity by Eq. (5):

$$\boldsymbol{L}(i,j)=\begin{cases}\frac{\mathrm{S}_l^{fun}(i,j)+\mathrm{G}_l(i,j)}{2} & \text{if } \mathrm{S}_l^{fun}(i,j)\neq 0\\ \mathbf{G_1}(i,j) & \text{otherwise.}\end{cases} \tag{5}$$

Nonlinear feature extraction

GAE is a novel graph neural network model and can effectively learn graph embedding features[69]. Here, we extract nonlinear features of lncRNAs and diseases based on the following five steps:

Step 1 Bipartite graph construction

First, a bipartite graph is constructed based on known LDA matrix $\boldsymbol{Y}$. The constructed bipartite graph contains two types of nodes, i.e., lncRNAs and diseases. Features of nodes can be represented based on their similarity matrix. And edge represents relationship between each lncRNA and disease.

Step 2 Feature projection

Both lncRNAs and diseases are projected to the space of vector with the same dimension based on a linear transformation matrix $\mathrm{Q}_{\varnothing}$. Taken lncRNA nodes as an example, lncRNAs are projected to a k-dimensional vector space by Eq. (6):

$$\mathrm{Q}_l = \mathrm{Q}_{\varnothing_l} \cdot \mathrm{T}_l \tag{6}$$

where Q_l, T_l, and $\mathrm{Q}_{\varnothing_l}$ denote lncRNA projected features, lncRNA similarity matrix, and linear transformation matrix which can be solved by minimizing the loss function, respectively.

Similarly, diseases are still projected into a *k*-dimensional vector space.

Step 3 Feature aggregation

An encoder is employed to yield the embeddings of lncRNAs and diseases by combining neighborhood node information. Given an lncRNA l_i, the aggregation $Q_{l_i}^c$ of features related to its direct neighbors $\{d_1, d_2, ...\}$ is depicted by an aggregate function $f(\cdot)$ defined by Eq. (7):

$$Q_{l_i}^c = \frac{1}{D_{l_i}} f\left(Q_{d_1}, Q_{d_2}, \dots\right) \tag{7}$$

where we usually use $sum(\cdot)$ as aggregator. D_{l_i} indicates the degree of l_i.

Step 4 Feature concatenation

The aggregated features $Q_{l_i}^c$ in Eq. (7) and the projected features Q_{l_i} in Eq. (6) are concatenated to update features Q'_{l_i} of l_i by a multi-layer perceptron by Eq. (8):

$$Q'_{l_i} = \mathrm{LeakyReLU}\left(g\left(Q_{l_i} \bigoplus Q_{l_i}^c\right)\right) \tag{8}$$

where $\bigoplus$ denotes the concatenation operation, $g(\cdot)$ is a multi-layer perceptron layer with LeakyReLU$(\cdot)$ and *k* outputs.

Similarly, the aggregated features Q'_{d_i} of diseases are updated.

Step 5 lncRNA and disease embedding learning

To incorporate abundant neighbor features and boost the model classification ability, we use an encoder based on stacking graph neural network with *N* layers to achieve the final embeddings (Q_l^N and Q_d^N) of lncRNAs and diseases.

Subsequently, a bilinear decoder is used to decode the input graph based on the association score $\hat{y}_{ij}$ between l_i and d_j by Eq. (9):

$$\hat{y}_{ij} = \mathrm{sigmoid}\left(Q_{d_j}^N H \left(Q_{l_i}^N\right)^T\right). \tag{9}$$

where H is a $k \times k$ parameter matrix. Consequently, we obtain nonlinear features Q_l^N of lncRNAs and Q_d^N of diseases.

During nonlinear feature learning, a cross-entropy loss *Loss* is used to evaluate whether the model effectively encodes LDA features and accurately reconstructs the input graph by Eq. (10):

$$Loss = -\sum_{i,j \in \mathscr{Y}^+ \cup \mathscr{Y}^-} \left(y_{ij} \log \hat{y}_{ij} + (1 - y_{ij}) \log\left(1 - \hat{y}_{ij}\right)\right). \tag{10}$$

where y_{ij} denotes the known relationship between l_i and d_j in a dataset, $\mathscr{Y}^+$ and $\mathscr{Y}^-$ indicate positive LDAs and negative LDAs, respectively. By minimizing the loss function defined by Eq. (10), we can solve the linear transformation matrix in Eq. (6).

LDA prediction

The gradient boosting decision tree models, such as XGBoost and LightGBM, effectively combine powerful learners and optimization methods, and thus improve the model classification accuracy, accelerate the model training, and enhance the model ability to handle intricate datasets. These models overcome the computational efficiency limitations that inhibit current boosting models. Thus, they have been taken as the most efficient classification tools and the most preferred choices to solve practical problems. However, when using the cross entropy loss, these models have nonsymmetric and unbounded features. As a result, they are sensitive to label noises, making the effect of noise be amplified[70]. To address the above problems, we devise a noise-robust gradient boosting model for LDA classification.

For an LDA dataset $D = \{(\boldsymbol{x}_i, \boldsymbol{y}'_i)\}_{i=1}^N$, suppose that $\boldsymbol{x}_i$ and y'_i denote the *i*-th training sample (i.e., lncRNA-disease pair) and its label. $\boldsymbol{x}_i$ is represented a feature vector by concatenating linear and nonlinear features of lncRNA and linear and nonlinear features of disease. $\boldsymbol{y}'_i = 1$ when the *i*-th pair has a link, otherwise $\boldsymbol{y}'_i = 0$. As shown in Algorithm 1, we perform predictions through the following five steps:

Step 1 Model initialization

Let f_{t+1} denotes a new decision tree, $z_i^t = z_i^0 + \alpha \sum_{j=1}^t f_j(x_i)$ denotes the model's raw prediction with the initial prediction z_i^0, and $p_i^{t+1} = S(z_i^{t+1}) = 1/(1 + e^{-z_i^{t+1}})$ is computed based on the Sigmoid function. Particularly, z_i^0 is set to zero (for all $i = 1, 2, \cdots, n$) to reduce impact of the model on the final prediction.

Step 2 Residual calculation

Let $l(y'_i, p)$ indicate the loss function for LDA classification, where *p* is the probability that the *i*-th lncRNA-disease pair is labeled as positive class. Consequently, $\lim_{p \to 0} l(0, p) = 0$ and $\lim_{p \to 1} l(1, p) = 1$. At the $(t+1)$-th iteration ($t \geq 0$), the objective function with the learning rate α is defined as Eq. (11):

$$\mathscr{L}^{t+1} = \sum_{i=1}^{n} l(y_i, p_i^{t+1}) = \sum_{i=1}^{n} l(y_i', S(z_i^{t+1})) = \sum_{i=1}^{n} l(y_i, S(z_i^t + \alpha f_{t+1}(\mathrm{x}_i))) \tag{11}$$

For convenience, we define the probability $\hat{p}$ for the ground-truth class as

$$\hat{p} = \begin{cases} p, & y = 1 \\ 1 - p, & y = 0 \end{cases} \tag{12}$$

Consequently, $l(y, p)$ is written as $l(1, \hat{p})$($l(\hat{p})$ for simplicity) and $\lim_{p \to 1} l(\hat{p}) = 0, \forall y \in \{0, 1\}$.

Step 3 Calculating gradient and Hessian

The Newton's method is adopted to optimize the regularized objective (13) with regularization parameter $\Omega(f_{t+1})$ for the boosting model:

$$\widetilde{\mathscr{L}}^{t+1} = \sum_{i=1}^{n} [g_i^t f_{t+1}(\mathrm{x}_i) + \frac{1}{2} h_i^t f_{t+1}(\mathrm{x}_i)^2] + \Omega(f_{t+1}) \tag{13}$$

where g_i^t and h_i^t denote the gradient and the Hessian, respectively. They are defined by Eq. (14):

$$\begin{aligned} g_i^t &= \frac{\partial l}{\partial z_i^t} = \frac{\partial l}{\partial \hat{p}_i^t} \frac{\partial \hat{p}_i^t}{\partial p_i^t} \frac{\partial p_i^t}{\partial z_i^t} = \frac{\partial l}{\partial \hat{p}_i^t} (2y_i^{'} - 1)(\hat{p}_i^t (1 - \hat{p}_i^t)) \\ h_i^t &= \frac{\partial^2 l}{\partial (z_i^t)^2} = \frac{\partial^2 l}{\partial (\hat{p}_i^t)^2} (\hat{p}_i^t (1 - \hat{p}_i^t))^2 + \frac{\partial l}{\partial \hat{p}_i^t} (\hat{p}_i^t (1 - \hat{p}_i^t)(1 - 2\hat{p}_i^t)) \end{aligned} \tag{14}$$

and

$$\frac{\partial p_i^t}{\partial z_i^t} = p_i^t (1 - p_i^t) = \hat{p}_i^t (1 - \hat{p}_i^t) \tag{15}$$

Step 4 Calculating the optimal weight of the leaves

For a decision tree with fixed structure, $f_{t+1}(\mathrm{x})$ is written as $f_{t+1}(\mathrm{x}) = \sum_{j=1}^{J} w_j I_j$, where I_j denotes an instance set involved to leaf j. Suppose that $\Omega(f_{t+1}) = \frac{1}{2} \lambda \sum_{j=1}^{J} w_j^2, \lambda \geq 0$, the optimal objective is rewritten as Eq. (16):

$$\widetilde{\mathscr{L}}^{t+1} = \sum_{i=1}^{n} [g_i^t f_{t+1}(\mathrm{x}_i) + \frac{1}{2} h_i^t f_{t+1}(\mathrm{x}_i)^2] + \Omega(f_{t+1}) = \sum_{j=1}^{J} [(\sum_{i \in I_j} g_i^t) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i^t + \lambda) w_j^2] \tag{16}$$

$f(x)$ is separable for each leaf, thus, the optimal weight w_j^* of leaf j can be computed by Eq. (17):

$$w_j^* = \frac{-\sum_{i \in I_j} g_i^t}{\sum_{i \in I_j} h_i^t + \lambda} \tag{17}$$

and the corresponding optimal objective is defined by Eq. (18):

$$\widetilde{\mathscr{L}}_j^* = -\frac{1}{2} \frac{(\sum_{i \in I_j} g_i^t)^2}{\sum_{i \in I_j} h_i^t + \lambda} \tag{18}$$

Step 5 Finding the best tree structure

We use the information gain to assess whether the tree will be grown and to identify the split feature as well as split value by Eq. (19):

$$gain = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] \tag{19}$$

where I_L and I_R denote the left node instance set and the right node instance set after the split, respectively. $I = I_L \cup I_R$ denotes the sample set related to their father nodes. I_L, I_R and I are used to limit the leaf split and alleviate the overfitting risk from the following three situations:

- If the number of samples for a leaf is not enough, the leaf split will be stopped.
- If the sum of hessian $\sum_{i \in I_j} h_i^t$ within a leaf is less than a small threshold ϵ, i.e., $\sum_{i \in I_j} h_i^t < \epsilon$, the growth of the tree will be stopped.

- If both the above two situations are satisfied, but $gain < \delta$, the growth process will be also stopped to prevent overfitting.

Additionally, we adopt a novel loss function called the Robust Focal Loss l_{RFL} to increase the model robustness by Eq. (20):

$$l_{RFL} = (1 - \hat{p})^r \frac{1 - \hat{p}^q}{q} \tag{20}$$

where hyperparameter $r \geq 0$, $q \in (0, 1)$.

The randomly selected negative LDAs may contain a small number of positive associations. Thus, the model optimizes its parameters by dealing with different levels of noises during training. When the optimal parameters are determined, LDA prediction is achieved.

Data availability

The datasets and codes for this study are available on GitHub at https://github.com/smiling199/LDA-GARB.

Received: 8 July 2024; Accepted: 19 May 2025
Published online: 31 May 2025

References

1. Zhang, X. et al. Fusion of multi-source relationships and topology to infer lncrna-protein interactions. *Mol. Therapy-Nucleic Acids* **2024**, 56 (2024).
2. Peng, L. et al. Lda-vghb: identifying potential lncrna-disease associations with singular value decomposition, variational graph auto-encoder and heterogeneous newton boosting machine. *Brief. Bioinform.* **25**, bbad466 (2024).
3. Yan, J., Wang, R. & Tan, J. Recent advances in predicting lncrna-disease associations based on computational methods. *Drug Discov. Today* **28**, 103432 (2023).
4. Chen, X. et al. Computational models for lncrna function prediction and functional similarity calculation. *Brief. Funct. Genom.* **18**, 58–82 (2019).
5. Rezakhani, L., Salmani, S., Eliyasi Dashtaki, M. & Ghasemi, S. Resveratrol: targeting cancer stem cells and ncrnas to overcome cancer drug resistance. *Curr. Mol. Med.* **24**, 951–961 (2024).
6. Wang, Y., Wang, X., Sun, H., Zhang, Z. & Gu, J. Lncrna mcm3ap-as1 promotes chemoresistance in triple-negative breast cancer through the mir-524-5p/rbm39 axis. *Mol. Cell. Biochem.* **2024**, 1–14 (2024).
7. Aswathy, R. & Sumathi, S. Defining new biomarkers for overcoming therapeutical resistance in cervical cancer using lncrna. *Mol. Biol. Rep.* **50**, 10445–10460 (2023).
8. Wang, S., Qiao, J. & Feng, S. Prediction of lncrna and disease associations based on residual graph convolutional networks with attention mechanism. *Sci. Rep.* **14**, 5185 (2024).
9. Sheng, N. et al. Data resources and computational methods for lncrna-disease association prediction. *Comput. Biol. Med.* **153**, 106527 (2023).
10. Peng, L. et al. Dti-mvsca: an anti-over-smoothing multi-view framework with negative sample selection for predicting drug-target interactions. *IEEE J. Biomed. Health Inform.* **29**, 711–723 (2025).
11. Peng, L. et al. Do-gma: an end-to-end drug-target interaction identification framework with a depthwise overparameterized convolutional network and the gated multihead attention mechanism. *J. Chem. Inf. Model.* **65**, 1318–1337 (2025).
12. Chen, G. et al. Lncrnadisease: a database for long-non-coding rna-associated diseases. *Nucleic Acids Res.* **41**, D983–D986 (2012).
13. Chen, X. & Yan, G.-Y. Novel human lncrna-disease association inference based on lncrna expression profiles. *Bioinformatics* **29**, 2617–2624 (2013).
14. Chen, X., Yan, C. C., Zhang, X. & You, Z.-H. Long non-coding rnas and complex diseases: from experimental results to computational models. *Brief. Bioinform.* **18**, 558–576 (2017).
15. Chen, X. Katzlda: Katz measure for the lncrna-disease association prediction. *Sci. Rep.* **5**, 16840 (2015).
16. Ganegoda, G. U., Li, M., Wang, W. & Feng, Q. Heterogeneous network model to infer human disease-long intergenic non-coding rna associations. *IEEE Trans. Nanobiosci.* **14**, 175–183 (2015).
17. Li, G. et al. Prediction of lncrna-disease associations based on network consistency projection. *Ieee Access* **7**, 58849–58856 (2019).
18. Li, J. et al. A novel approach for potential human lncrna-disease association prediction based on local random walk. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **18**, 1049–1059 (2019).
19. Chen, X. & Huang, L. Lrsslmda: Laplacian regularized sparse subspace learning for mirna-disease association prediction. *PLoS Comput. Biol.* **13**, e1005912 (2017).
20. Xie, G.-B. et al. Predicting lncrna-disease associations based on combining selective similarity matrix fusion and bidirectional linear neighborhood label propagation. *Brief. Bioinform.* **24**, bbac595 (2023).
21. Khalid, R., Naveed, H. & Khalid, Z. Computational prediction of disease related lncrnas using machine learning. *Sci. Rep.* **13**, 806 (2023).
22. Peng, L. et al. Predicting cell-cell communication by combining heterogeneous ensemble deep learning and weighted geometric mean. *Appl. Soft Comput.* **2025**, 112839 (2025).
23. Peng, L. et al. Bindti: a bi-directional intention network for drug-target interaction identification based on attention mechanisms. *IEEE J. Biomed. Health Inform.* **29**, 1602–1612 (2025).
24. 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. *IScience* **19**, 786–795 (2019).
25. Yao, D. et al. A random forest based computational model for predicting novel lncrna-disease associations. *BMC Bioinform.* **21**, 1–18 (2020).
26. Zhang, J., Sun, Q. & Liang, C. Prediction of lncrna-disease associations based on robust multi-label learning. *Curr. Bioinform.* **16**, 1179–1189 (2021).
27. Ha, J. Lncrna expression profile-based matrix factorization for identifying lncrna-disease associations. *IEEE Access* **2024**, 56 (2024).
28. Lu, C. et al. Prediction of lncrna-disease associations based on inductive matrix completion. *Bioinformatics* **34**, 3357–3364 (2018).
29. Yu, G. *et al.* Weighted matrix factorization based data fusion for predicting lncrna-disease associations. In *2018 IEEE international Conference on Bioinformatics and Biomedicine (BIBM)* 572–577 (IEEE, 2018).
30. Wang, B., Liu, R., Zheng, X., Du, X. & Wang, Z. lncrna-disease association prediction based on matrix decomposition of elastic network and collaborative filtering. *Sci. Rep.* **12**, 12700 (2022).

31. Yu, J., Xuan, Z., Feng, X., Zou, Q. & Wang, L. A novel collaborative filtering model for lncrna-disease association prediction based on the naïve bayesian classifier. *BMC Bioinform.* **20**, 1–13 (2019).
32. Cui, Z., Liu, J.-X., Gao, Y.-L., Zhu, R. & Yuan, S.-S. Lncrna-disease associations prediction using bipartite local model with nearest profile-based association inferring. *IEEE J. Biomed. Health Inform.* **24**, 1519–1527 (2019).
33. Lan, W. et al. Ldicdl: Lncrna-disease association identification based on collaborative deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **19**, 1715–1723 (2020).
34. Madhavan, M. & Gopakumar, G. Dbnlda: deep belief network based representation learning for lncrna-disease association prediction. *Appl. Intell.* **52**, 5342–5352 (2022).
35. Du, B., Tang, L., Liu, L. & Zhou, W. Predicting lncrna-disease association based on generative adversarial network. *Curr. Gene Ther.* **22**, 144–151 (2022).
36. Chen, Q., Qiu, J., Lan, W. & Cao, J. Similarity-guided graph contrastive learning for lncrna-disease association prediction. *J. Mol. Biol.* **2024**, 168609 (2024).
37. Su, Z., Lu, H., Wu, Y., Li, Z. & Duan, L. Predicting potential lncrna biomarkers for lung cancer and neuroblastoma based on an ensemble of a deep neural network and lightgbm. *Front. Genet.* **14**, 569 (2023).
38. Meng, Y. et al. Drug repositioning based on weighted local information augmented graph neural network. *Brief. Bioinform.* **25**, bbad431 (2024).
39. Tang, X. et al. Enhancing drug repositioning through local interactive learning with bilinear attention networks. *IEEE J. Biomed. Health Inf.* **2023**, 569 (2023).
40. Cao, Y. et al. X-lda: an interpretable and knowledge-informed heterogeneous graph learning framework for lncrna-disease association prediction. *Comput. Biol. Med.* **167**, 107634 (2023).
41. Zhang, Z. et al. Capsnet-lda: predicting lncrna-disease associations using attention mechanism and capsule network based on multi-view data. *Brief. Bioinform.* **24**, bbac531 (2023).
42. Peng, L., Ren, M., Huang, L. & Chen, M. Genddn: an lncrna-disease association identification framework based on dual-net neural architecture and deep neural network. *Interdiscipl. Sci. Comput. Life Sci.* **2024**, 1–21 (2024).
43. Zhang, B. et al. Ldagm: prediction lncrna-disease asociations by graph convolutional auto-encoder and multilayer perceptron based on multi-view heterogeneous networks. *BMC Bioinform.* **25**, 332 (2024).
44. Lan, W. et al. Ganlda: graph attention network for lncrna-disease associations prediction. *Neurocomputing* **469**, 384–393 (2022).
45. Xu, J. et al. Graph embedding and gaussian mixture variational autoencoder network for end-to-end analysis of single-cell rna sequencing data. *Cell Rep. Methods* **3**, 100382 (2023).
46. Bao, Z. et al. Lncrnadisease 2.0: an updated database of long non-coding rna-associated diseases. *Nucleic Acids Res.* **47**, D1034–D1037 (2019).
47. Cui, T. et al. Mndr v2. 0: an updated resource of ncrna-disease associations in mammals. *Nucleic Acids Res.* **46**, D371–D374 (2018).
48. Peng, L., Xiong, W., Han, C., Li, Z. & Chen, X. Celldialog: a computational framework for ligand-receptor-mediated cell-cell communication analysis. *IEEE J. Biomed. Health Inf.* **28**, 145 (2024).
49. Zeng, M. et al. Sdlda: lncrna-disease association prediction based on singular value decomposition and deep learning. *Methods* **179**, 73–80 (2020).
50. 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 Bioinform.* **21**, 1–27 (2020).
51. Chen, T. & Guestrin, C. Xgboost: a scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd International Conference on Knowledge Discovery and Data Mining* 785–794 (2016).
52. Zhou, L. et al. Lpi-hyadbs: a hybrid framework for lncrna-protein interaction prediction integrating feature selection and classification. *BMC Bioinform.* **22**, 1–31 (2021).
53. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. Catboost: unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **31**, 693 (2018).
54. Ke, G. et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30**, 896 (2017).
55. Xie, G., Jiang, J. & Sun, Y. Lda-lnsubrw: lncrna-disease association prediction based on linear neighborhood similarity and unbalanced bi-random walk. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **19**, 989–997 (2020).
56. Wu, X. et al. Inferring lncrna-disease associations based on graph autoencoder matrix completion. *Comput. Biol. Chem.* **87**, 107282 (2020).
57. Spaander, M. C. et al. Young-onset colorectal cancer. *Nat. Rev. Dis. Primers.* **9**, 21 (2023).
58. Chen, C. et al. lncrna-ccdc26, as a novel biomarker, predicts prognosis in acute myeloid leukemia. *Oncol. Lett.* **18**, 2203–2211 (2019).
59. Hirano, T. et al. Long noncoding rna, ccdc26, controls myeloid leukemia cell growth through regulation of kit expression. *Mol. Cancer* **14**, 1–16 (2015).
60. Wang, S., Hui, Y., Li, X. & Jia, Q. Silencing of lncrna ccdc26 restrains the growth and migration of glioma cells in vitro and in vivo via targeting mir-203. *Oncol. Res.* **26**, 1143 (2018).
61. Yan, J. et al. Downregulation of lncrna ccdc26 contributes to imatinib resistance in human gastrointestinal stromal tumors through igf-1r upregulation. *Braz. J. Med. Biol. Res.* **52**, e8399 (2019).
62. Nolan, E., Lindeman, G. J. & Visvader, J. E. Deciphering breast cancer: from biology to the clinic. *Cell* **186**, 1708–1728 (2023).
63. Bray, F. et al. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263 (2024).
64. Ma, J., Cao, K., Ling, X., Zhang, P. & Zhu, J. Lncrna har1a suppresses the development of non-small cell lung cancer by inactivating the stat3 pathway. *Cancers* **14**, 2845 (2022).
65. Lee, C.-P., Ko, A.M.-S., Nithiyanantham, S., Lai, C.-H. & Ko, Y.-C. Long noncoding rna har1a regulates oral cancer progression through the alpha-kinase 1, bromodomain 7, and myosin iia axis. *J. Mol. Med.* **99**, 1323–1334 (2021).
66. Zhang, L. et al. Effects of har1 on cognitive function in mice and the regulatory network of har1 determined by rna sequencing and applied bioinformatics analysis. *Front. Genet.* **14**, 947144 (2023).
67. Saberi-Movahed, F., Berahman, K., Sheikhpour, R., Li, Y. & Pan, S. Nonnegative matrix factorization in dimensionality reduction: a survey. arXiv:2405.03615 (2024).
68. Fan, W. et al. Idssim: an lncrna functional similarity calculation model based on an improved disease semantic similarity method. *BMC Bioinform.* **21**, 1–14 (2020).
69. Liang, Q., Zhang, W., Wu, H. & Liu, B. Lncrna-disease association identification using graph auto-encoder and learning to rank. *Brief. Bioinform.* **24**, bbac539 (2023).
70. Luo, J., Quan, Y. & Xu, S. Robust-gbdt: a novel gradient boosting model for noise-robust classification. arXiv:2310.05067 (2023).

Author contributions

Conceptualization: L-LT, L-LH, and Y-Y; Investigation: L-LT and L-LH; Methodology: L-LT, L-LH, and Y-Y; Project administration: L-LT, L-LH, and Y-Y; Software: L-LT and L-LH; Writing-original draft: L-LT and L-LH; Writing-review and editing: L-LT, L-LH, and Y-Y. All authors reviewed the manuscript.

Funding
This research was funded by the Natural Science Foundation of Hunan province (Grant 2023J50203).

Competing interests
All authors declare that they have no competing interests.

Additional information
Correspondence and requests for materials should be addressed to L.H. or Y.Y.

Reprints and permissions information is available at www.nature.com/reprints.

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