frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 01 March 2024
DOI 10.3389/fgene.2024.1356205

OPEN ACCESS

EDITED BY
Wen Zhang,
Huazhong Agricultural University, China

REVIEWED BY
Advait Balaji,
Occidental Petroleum Corporation, United States
XianFang Tang,
Wuhan Textile University, China
Guohua Huang,
Shaoyang University, China
Ying Liang,
Jiangxi Agricultural University, China

*CORRESPONDENCE
Lijun Zeng,
zenglijun@hnit.edu.cn
Lihong Peng,
plhhnu@163.com

RECEIVED 18 December 2023
ACCEPTED 01 February 2024
PUBLISHED 01 March 2024

CITATION
Zhou L, Peng X, Zeng L and Peng L (2024), Finding potential lncRNA–disease associations using a boosting-based ensemble learning model.
Front. Genet. 15:1356205.
doi: 10.3389/fgene.2024.1356205

Finding potential lncRNA–disease associations using a boosting-based ensemble learning model

Liqian Zhou[1], Xinhuai Peng[1], Lijun Zeng[2]* and Lihong Peng[1]*

[1]School of Computer Science, Hunan University of Technology, Zhuzhou, Hunan, China, [2]School of Computer Science, Hunan Institute of Technology, Hengyang, China

Introduction: Long non-coding RNAs (lncRNAs) have been in the clinical use as potential prognostic biomarkers of various types of cancer. Identifying associations between lncRNAs and diseases helps capture the potential biomarkers and design efficient therapeutic options for diseases. Wet experiments for identifying these associations are costly and laborious.

Methods: We developed LDA-SABC, a novel boosting-based framework for lncRNA–disease association (LDA) prediction. LDA-SABC extracts LDA features based on singular value decomposition (SVD) and classifies lncRNA–disease pairs (LDPs) by incorporating LightGBM and AdaBoost into the convolutional neural network.

Results: The LDA-SABC performance was evaluated under five-fold cross validations (CVs) on lncRNAs, diseases, and LDPs. It obviously outperformed four other classical LDA inference methods (SDLDA, LDNFSGB, LDASR, and IPCAF) through precision, recall, accuracy, F1 score, AUC, and AUPR. Based on the accurate LDA prediction performance of LDA-SABC, we used it to find potential lncRNA biomarkers for lung cancer. The results elucidated that 7SK and HULC could have a relationship with non-small-cell lung cancer (NSCLC) and lung adenocarcinoma (LUAD), respectively.

Conclusion: We hope that our proposed LDA-SABC method can help improve the LDA identification.

KEYWORDS

lncRNA–disease association, singular value decomposition, LightGBM, AdaBoost, convolutional neural network

1 Introduction

Long non-coding RNAs (lncRNAs) are important RNA molecules comprising more than 200 nucleotides (Jiang et al., 2015; Liu et al., 2021; Chen et al., 2023). lncRNAs have been in the clinical use as prognostic biomarkers of many complex diseases, including cancers (Tang et al., 2022; 2021; Huo et al., 2021). For example, liver-specific lncRNA FAM99A plays a cancer-inhibiting role in hepatocellular carcinoma and might serve as its prognostic biomarker (Mo et al., 2022). Exosomal RP5-977B1 might be a diagnostic biomarker of non-small-cell lung cancer (NSCLC) (Min et al., 2022). MALAT1 has been broadly applied for its oncogenic properties in lung cancer (Xin et al., 2023), bladder cancer (Li et al., 2017), breast cancer (Adewunmi et al., 2023), and ovarian cancer (Mao et al., 2021). Identifying possible

relationships between lncRNAs and diseases helps capture potential biomarkers for various cancers and provide clues for their diagnosis and treatment (Wang et al., 2021). Traditional wet experiments for detecting new lncRNA–disease associations (LDAs) are costly and have low success rates; computational techniques have been increasingly developed to discover new LDAs (Chen et al., 2021; Zhao et al., 2023). Meanwhile, various lncRNA-related databases, such as MNDR v2.0 (Cui et al., 2018), Lnc2Cancer (Ning et al., 2016), LncRNADisease 3.0 (Lin et al., 2023b), and NRED (Dinger et al., 2009), provide diverse LDA data resources. Based on these resources, many computational methods, especially network-based and machine learning methods, have been applied to LDA prediction (Chen et al., 2017; Chen and Huang, 2022; Sheng et al., 2023).

Network-based methods predict new LDAs through label propagation and multi-information fusion on the heterogeneous lncRNA–disease networks (Jiang et al., 2010; Zou et al., 2016; Hu et al., 2017; Wang et al., 2019; Yu et al., 2020; Qiu et al., 2023b). Chen et al. conducted many research studies and significantly promoted LDA prediction (Chen and Yan, 2013; Chen et al., 2015; Chen, 2015a; Chen, 2015b). Based on these studies, they comprehensively concluded the current computational methods for non-coding RNA analysis and unfolded existing challenges and corresponding solutions (Chen and Huang, 2022; 2023). Xie et al. used the unbalanced bi-random walk algorithm (Xie et al., 2020b; a) and bidirectional linear neighborhood label propagation (Xie et al., 2023) for LDA identification. In addition, a random walk with a restart algorithm (Wang et al., 2022) has been still applied to find new LDAs. Network-based methods found many possible LDAs, but they did not analyze the topological features of LDA networks.

Machine learning methods have been applied to various association discovery tasks (Zou et al., 2018; Peng et al., 2019; 2022b; Shen et al., 2022; Wu et al., 2022; Yu et al., 2022; Lin et al., 2023a; Peng et al., 2023a; Peng et al., 2023b; Peng et al., 2024b; Han et al., 2023; Liu and Zhang, 2023; Qi and Zou, 2023; Xiong et al., 2023; Xu et al., 2024; Zhang et al., 2024). Consequently, machine learning algorithms have been broadly applied in LDA prediction, for example, collaborative filtering (Yu et al., 2019), graph regularization (Liu et al., 2020; Wang et al., 2021), matrix factorization (Fu et al., 2018; Wang et al., 2020; Xi et al., 2022), heterogeneous graph learning framework, (Cao et al., 2023), and ensemble learning models (Peng et al., 2022a). Notably, deep learning has been broadly applied due to its powerful classification performance (Sun et al., 2022; Wang et al., 2023; Wang et al., 2023b; Hu et al., 2023; Jiang et al., 2023; Zhang et al., 2023; Zhang and Wu, 2023; Zhou et al., 2024a), such as in the graph convolution network (Wang W. et al., 2022), node2vec (Li et al., 2021), collaborative deep learning (Lan et al., 2020), deep neural network (Wei et al., 2020), deep multi-network embedding (Ma, 2022), graph autoencoder (Liang et al., 2023; Zhou et al., 2024b), and a capsule network with the attention mechanism (Zhang et al., 2023). In particular, to identify new LDAs, a few models first extracted LDA features and classified unknown lncRNA–disease pairs (LDPs) by combining machine leaning models. SDLDA (Zeng et al., 2020) effectively integrated deep learning and singular value decomposition (SVD), LDASR (Guo et al., 2019) combined autoencoder and rotating forest, LDNFSGB (Zhang et al., 2020) used autoencoder and the gradient boosting model, IPCARF (Zhu et al., 2021) applied the incremental principal component analysis and random forest, CapsNet-LDA (Zhang et al., 2023) utilized stacked autoencoder and attention mechanism, and LDAEXC (Lu and Xie, 2023) integrated deep autoencoder and XGBoost. Machine learning-based methods boosted LDA prediction, but they neglect noisy and irrelevant data.

To boost the LDA prediction performance, here, we developed LDA-SABC, a novel boosting-based framework for LDA prediction. LDA-SABC extracts LDA features based on SVD and classifies LDPs by integrating LightGBM (Wang et al., 2023) and AdaBoost combined with the convolutional neural network (AdaBoost-CNN) (Taherkhani et al., 2020; Peng et al., 2023c). The LDA-SABC performance was evaluated under fivefold cross validations (CVs) on lncRNAs, diseases, and LDPs. This approach accurately found a few potential lncRNAs for lung cancer. LDA-SABC is publicly available at https://github.com/plhhnu/LDA-SABC.

2 Materials and methods

2.1 Overview of LDA-SABC

LDA-SABC contains two main steps: 1) LDA feature extraction: the LDP linear features are extracted through SVD. 2) LDA classification: the association probability of each LDP is computed by integrating AdaBoost-CNN and LightGBM. The details are shown in Figure 1.

2.2 Data preparation

LDA-SABC was evaluated on two human LDA datasets (Peng et al., 2024a), namely, LncRNADisease (Chen et al., 2012) and MNDR (Cui et al., 2018). After deleting diseases without regular names or MeSH data and lncRNAs without sequence data, the number of lncRNAs, one of the diseases, and one of the LDAs in two LDA datasets are listed in Table 1. Subsequently, an LDA network containing n lncRNAs and m diseases is denoted as $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$, where $y_{ij} = 1$ if lncRNA l_i is associated with disease d_j, otherwise $y_{ij} = 0$.

2.3 LDA feature extraction

SVD (Abdi, 2007) can effectively extract features by eigen decomposition. By selecting larger singular values, SVD can reduce the dimensionality of the data and remove features that contribute less to data variability, thereby reducing the storage and calculation costs of the data. In addition, the feature vectors corresponding to smaller singular values represent noise or redundant parts in the data. By selecting larger singular values, SVD can retain the main linear features, thereby removing noise and redundant information. Furthermore, the size of singular values represents important features in the data, and SVD helps us understand the structure and variation patterns of the data by observing the size of singular values and their corresponding feature vectors. Thus, SVD is used to extract lncRNA and disease features: the LDA matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$ is factorized using Eq. 1:

$$\boldsymbol{Y} = \boldsymbol{U} \sum \boldsymbol{V}^T, \tag{1}$$

10.3389/fgene.2024.1356205

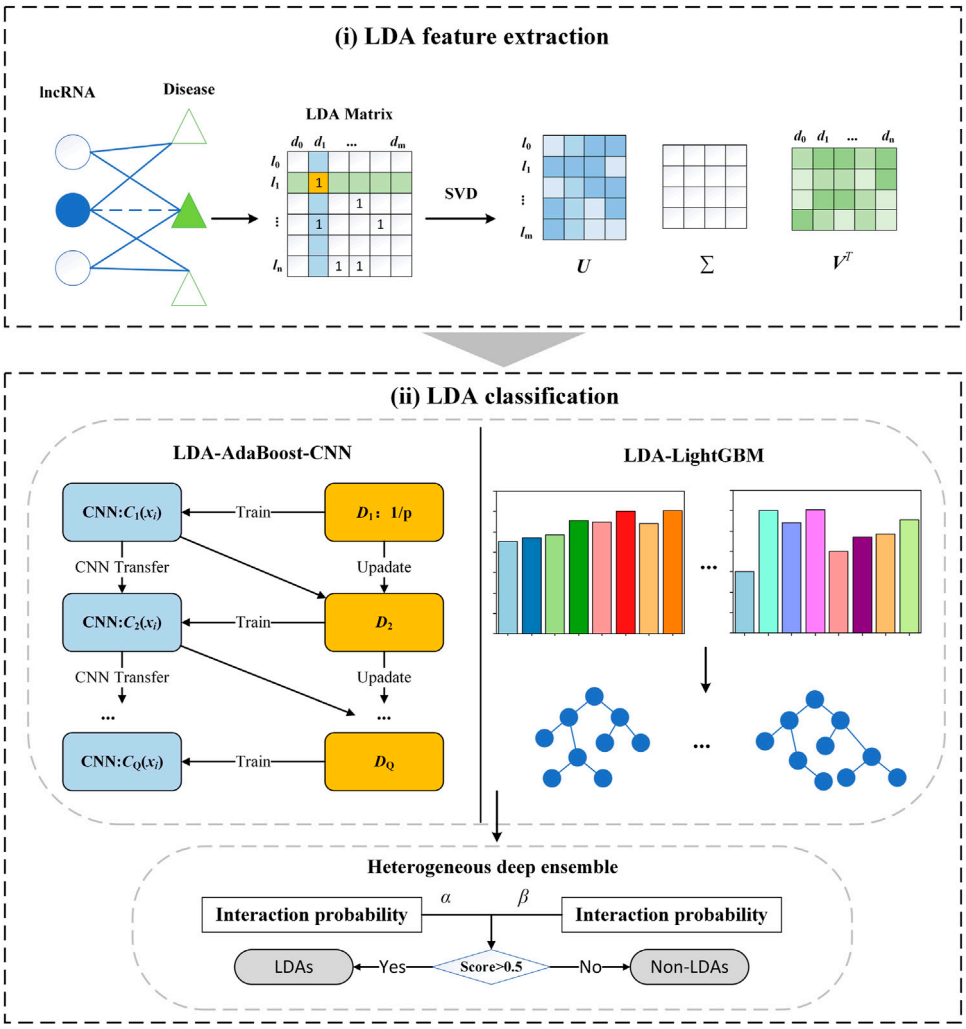

FIGURE 1
Flowchart of the LDA prediction model LDA-SABC: (i) LDA feature extraction. (ii) LDA classification.

TABLE 1 Introduction of two LDA datasets.

Dataset	lncRNA	Disease	LDA
LncRNADisease	82	157	605
MNDR	89	190	1,529

where V^T represents the transpose of V, $U \in R^{n\times n}$ and $V \in R^{m\times m}$ are two real matrices, and Σ denotes a diagonal matrix composed of n singular values.

Subsequently, the e largest singular values are selected to build an approximation representation using Eq. 2:

$$R \approx U_i \sum_e \left(V_j\right)^T. \quad (2)$$

Consequently, U_i and ${V_j}^T$ denote the features of the ith lncRNA l_i and the jth disease d_j, respectively.

As a result, the features of each lncRNA can be represented as an a-dimensional vector, and the features of each disease can be represented as a b-dimensional vector. The two features are concatenated as a d ($d = a + b$)-dimensional vector for characterizing each LDP.

2.4 LDA prediction

For an LDA dataset $D = (\boldsymbol{X}, \hat{\boldsymbol{Y}})$, with p ($p = n \times m$) samples (i.e., p LDPs), let $\boldsymbol{x}_i \in \boldsymbol{X}$ denote the ith LDP with d-dimensional features, and $\boldsymbol{y}_i \in \hat{\boldsymbol{Y}}$ denotes its label.

2.4.1 LDA-AdaBoost-CNN

Inspired by AdaBoost-CNN proposed by Hastie et al. (2009) and Taherkhani et al. (2020), we exploit an LDA identification algorithm LDA-AdaBoost-CNN by integrating AdaBoost and CNNs based on transfer learning. Given Q CNNs, LDA-AdaBoost-CNN uses CNNs as base estimators for predicting LDAs. During training, we use a vector $\boldsymbol{D}$ with initial values $\frac{1}{p}$ to measure the importance of each sample. Next, the weights of all training samples are updated and normalized. Finally, LDA-AdaBoost-CNN outputs a binary vector

TABLE 2 Performance of five LDA inference methods under CV_l.

	Dataset	SDLDA	LDNFSGB	IPCARF	LDASR	LDA-SABC
Precision	LncRNADisease	0.8514 ± 0.0509	0.7004 ± 0.0639	0.4878 ± 0.1309	0.6726 ± 0.1200	**0.8980 ± 0.0306**
	MNDR	0.9399 ± 0.0154	0.8552 ± 0.0393	0.6615 ± 0.0966	0.8405 ± 0.0300	**0.9494 ± 0.0172**
Recall	LncRNADisease	0.6521 ± 0.0732	0.6092 ± 0.0790	0.5721 ± 0.1580	0.5129 ± 0.0946	**0.7709 ± 0.0622**
	MNDR	0.8239 ± 0.0437	0.8021 ± 0.0498	0.6434 ± 0.1545	0.7358 ± 0.0562	**0.8436 ± 0.0513**
Accuracy	LncRNADisease	0.7799 ± 0.0341	0.6769 ± 0.0423	0.4906 ± 0.0951	0.6417 ± 0.0597	**0.8444 ± 0.0445**
	MNDR	0.8857 ± 0.0283	0.8323 ± 0.0230	0.6526 ± 0.0775	0.7972 ± 0.0268	**0.8989 ± 0.0317**
F1 score	LncRNADisease	0.7365 ± 0.0563	0.6462 ± 0.0451	0.5125 ± 0.1100	0.5668 ± 0.0536	**0.8278 ± 0.0363**
	MNDR	0.8775 ± 0.0278	0.8260 ± 0.0230	0.6401 ± 0.1017	0.7827 ± 0.0260	**0.8925 ± 0.0307**
AUC	LncRNADisease	0.8023 ± 0.0477	0.7346 ± 0.0465	0.5096 ± 0.1432	0.7057 ± 0.0420	**0.9328 ± 0.0243**
	MNDR	0.9366 ± 0.0195	0.8839 ± 0.0270	0.7104 ± 0.0997	0.8641 ± 0.0256	**0.9675 ± 0.0147**
AUPR	LncRNADisease	0.8461 ± 0.0553	0.7239 ± 0.0626	0.5336 ± 0.1423	0.6775 ± 0.0971	**0.9304 ± 0.0252**
	MNDR	0.9533 ± 0.0129	0.8832 ± 0.0307	0.7128 ± 0.1012	0.8671 ± 0.0252	**0.9709 ± 0.0106**

TABLE 3 Performance of five LDA inference methods under CV_d.

	Dataset	SDLDA	LDNFSGB	IPCARF	LDASR	LDA-SABC
Precision	LncRNADisease	0.8854 ± 0.0377	0.7548 ± 0.0639	0.5583 ± 0.0910	0.7462 ± 0.0613	**0.9218 ± 0.0242**
	MNDR	0.9232 ± 0.0331	0.8005 ± 0.0625	0.5557 ± 0.1473	0.7625 ± 0.0749	**0.9573 ± 0.0217**
Recall	LncRNADisease	0.7182 ± 0.0694	0.7309 ± 0.0646	0.7538 ± 0.1067	0.6431 ± 0.0757	**0.8745 ± 0.0353**
	MNDR	0.8579 ± 0.0655	0.6936 ± 0.0794	0.5279 ± 0.1969	0.5758 ± 0.0894	**0.9231 ± 0.0400**
Accuracy	LncRNADisease	0.8187 ± 0.0282	0.7552 ± 0.0291	0.5766 ± 0.0740	0.7165 ± 0.0339	**0.9008 ± 0.0232**
	MNDR	0.9043 ± 0.0174	0.7670 ± 0.0432	0.5593 ± 0.1159	0.7010 ± 0.0463	**0.9455 ± 0.0146**
F1 score	LncRNADisease	0.7917 ± 0.0519	0.7407 ± 0.0526	0.6339 ± 0.0715	0.6873 ± 0.0512	**0.8970 ± 0.0218**
	MNDR	0.8886 ± 0.0475	0.7402 ± 0.0577	0.5190 ± 0.1434	0.6485 ± 0.0555	**0.9394 ± 0.0260**
AUC	LncRNADisease	0.8788 ± 0.0274	0.8329 ± 0.0273	0.6402 ± 0.1004	0.7951 ± 0.0317	**0.9630 ± 0.0122**
	MNDR	0.9559 ± 0.0160	0.8603 ± 0.0363	0.5992 ± 0.1601	0.8045 ± 0.0362	**0.9860 ± 0.0057**
AUPR	LncRNADisease	0.8934 ± 0.0387	0.8163 ± 0.0537	0.6355 ± 0.1217	0.7914 ± 0.0542	**0.9605 ± 0.0130**
	MNDR	0.9561 ± 0.0354	0.8292 ± 0.0680	0.6040 ± 0.1476	0.7630 ± 0.0717	**0.9836 ± 0.0101**

$\boldsymbol{o}_l^k(\boldsymbol{x}_i)$ with the last CNN to identify one LDP as LDA ($k = 1$) or non-LDA ($k = 2$).

For the ith feature map in the lth layer y_i^l, its activity is computed using Eq. 3:

$$\boldsymbol{y}_i^l = \sum_j f\left(\boldsymbol{w}_{i,j}^l * \boldsymbol{y}_j^{l-1} + \boldsymbol{b}_i^l\right), \tag{3}$$

where $\boldsymbol{w}_{i,j}^l$ represents the weight of a convolutional kernel, which maps the jth feature at the $(l-1)$th CNN layer to the ith feature at the lth CNN layer, and $\boldsymbol{b}_i^l$ is the bias of the ith feature in the lth layer. Finally, the output $\boldsymbol{F}^l$ at the lth hidden layer is computed using Eq. 4:

$$\boldsymbol{F}^l = f\left(\boldsymbol{W}^l\left(\boldsymbol{F}^{l-1}\right)^T + \boldsymbol{b}^l\right), \tag{4}$$

where $f(\cdot)$ denotes a non-linear function. Consequently, the probability distribution matrix $\boldsymbol{Z}$ of all LDPs is computed via a softmax function using Eq. 5:

$$\boldsymbol{Z} = \mathrm{softmax}\left(\boldsymbol{W}^o\left(\boldsymbol{F}^L\right)^T + \boldsymbol{b}^o\right), \tag{5}$$

where $\boldsymbol{W}^o$ denotes a weight matrix linking the last hidden layer with the output layer, $\boldsymbol{b}^o$ indicates the bias, and $\boldsymbol{F}^L$ represents the output at the last hidden layer.

For the ith sample $\boldsymbol{x}_i$, after training Q CNNs, its output is computed based on its output $\boldsymbol{o}_q^k(\boldsymbol{x}_i)$ $(k = 1, 2)$ in the qth CNN using Eq. 6:

$$C(\boldsymbol{x}_i) = \underset{k}{\mathrm{argmax}} \sum_{q=1}^{Q} c_k^q(\boldsymbol{x}_i), \tag{6}$$

TABLE 4 Performance of five LDA inference methods under CV_{ld}.

	Dataset	SDLDA	LDNFSGB	IPCARF	LDASR	LDA-SABC
Precision	LncRNADisease	0.8782 ± 0.0306	0.7782 ± 0.0270	0.7069 ± 0.0478	0.7695 ± 0.0393	**0.9052 ± 0.0241**
	MNDR	0.9178 ± 0.0154	0.8548 ± 0.0156	0.7693 ± 0.0850	0.8553 ± 0.0189	**0.9525 ± 0.0153**
Recall	LncRNADisease	0.7256 ± 0.0376	0.8169 ± 0.0408	0.6155 ± 0.0652	0.6836 ± 0.0342	**0.9074 ± 0.0329**
	MNDR	0.8824 ± 0.0198	0.8818 ± 0.0204	0.5034 ± 0.1469	0.8204 ± 0.0238	**0.9459 ± 0.0131**
Accuracy	LncRNADisease	0.8120 ± 0.0216	0.7916 ± 0.0256	0.6793 ± 0.0403	0.7385 ± 0.0283	**0.9058 ± 0.0183**
	MNDR	0.9015 ± 0.0114	0.8658 ± 0.0127	0.6793 ± 0.0753	0.8405 ± 0.0129	**0.9493 ± 0.0109**
F1 score	LncRNADisease	0.7939 ± 0.0260	0.7965 ± 0.0262	0.6563 ± 0.0492	0.7233 ± 0.0289	**0.9058 ± 0.0190**
	MNDR	0.8996 ± 0.0119	0.8679 ± 0.0129	0.5995 ± 0.1312	0.8371 ± 0.0137	**0.9491 ± 0.0108**
AUC	LncRNADisease	0.8774 ± 0.0200	0.8578 ± 0.0234	0.7384 ± 0.0466	0.8133 ± 0.0218	**0.9628 ± 0.0132**
	MNDR	0.9560 ± 0.0081	0.9346 ± 0.0074	0.7680 ± 0.0882	0.9143 ± 0.0112	**0.9878 ± 0.0046**
AUPR	LncRNADisease	0.8952 ± 0.0177	0.8489 ± 0.0289	0.7409 ± 0.0515	0.8131 ± 0.0277	**0.9606 ± 0.0150**
	MNDR	0.9639 ± 0.0063	0.9273 ± 0.0098	0.7689 ± 0.0924	0.9100 ± 0.0136	**0.9881 ± 0.0055**

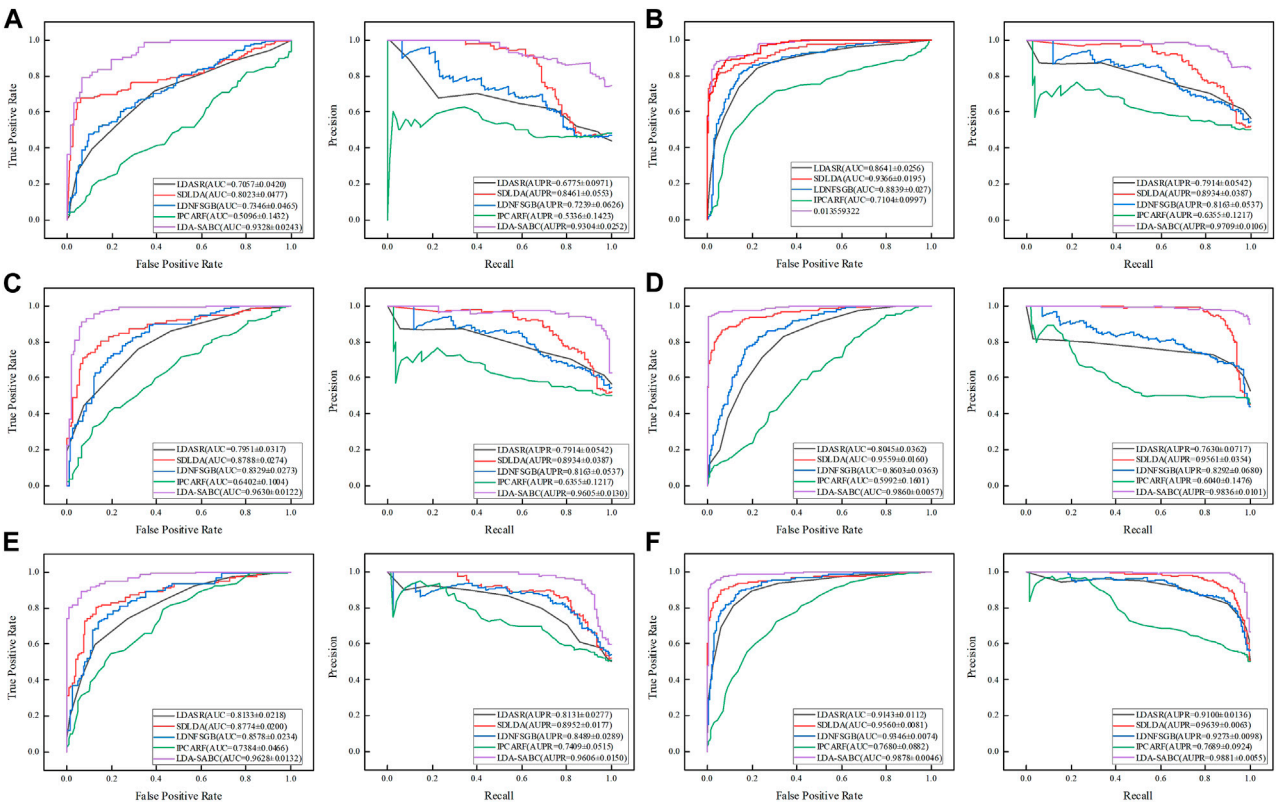

FIGURE 2
ROC and PR curves of LDA-SABC and four other methods: **(A,B)** ROC and PR curves on the LncRNADisease and MNDR databases under CV_l, respectively. **(C,D)** Curves under CV_d. **(E,F)** Curves under CV_{ld}.

where

$$c_k^q(\boldsymbol{x}_i) = \log\left(\boldsymbol{o}_k^q(\boldsymbol{x}_i)\right) - \frac{1}{2}\sum_{k'=1}^{2}\log\left(\boldsymbol{o}_{k'}^q(\boldsymbol{x}_i)\right). \quad (7)$$

2.4.2 LDA-LightGBM

LightGBM is a gradient-based model. It uses two powerful techniques to acquire the optimal split node and accurately classify unknown samples: one-side sampling and exclusive

10.3389/fgene.2024.1356205

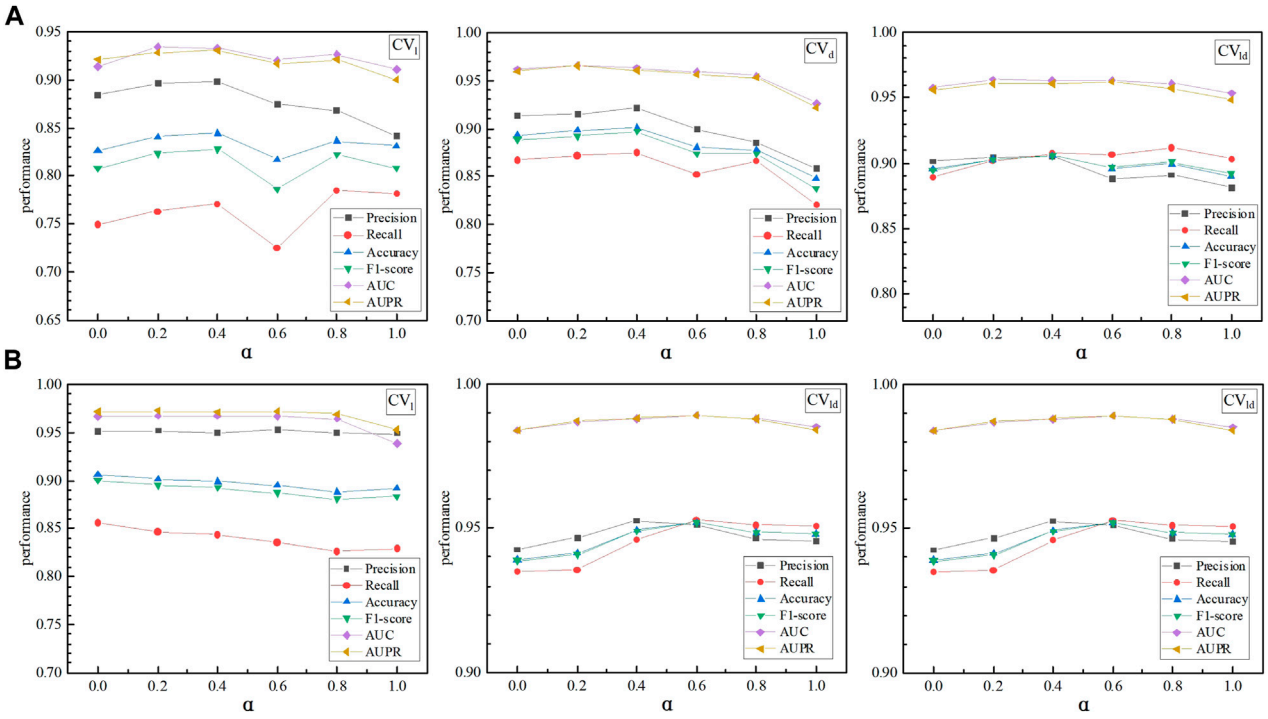

FIGURE 3
Effects of the parameters α and β on the LDA prediction performance: **(A)** performance of LDA-SABC based on α and β on LncRNADisease under CV_l, CV_d, and CV_{ld}, respectively. **(B)** Performance of LDA-SABC based on different α and β values on MNDR under CV_l, CV_d, and CV_{ld}, respectively.

TABLE 5 Performance of four boosting methods under CV_l.

	Dataset	AdaBoost-CNN	AdaBoost	LightGBM	LDA-SABC
Precision	LncRNADisease	0.8412 ± 0.0584	0.7641 ± 0.0536	0.8836 ± 0.0354	**0.8980 ± 0.0306**
	MNDR	0.9486 ± 0.0217	0.8826 ± 0.0331	**0.9510 ± 0.0175**	0.9494 ± 0.0172
Recall	LncRNADisease	**0.7815 ± 0.0844**	0.7151 ± 0.0805	0.7494 ± 0.0765	0.7709 ± 0.0622
	MNDR	0.8295 ± 0.0728	0.8483 ± 0.0374	**0.8561 ± 0.0506**	0.8436 ± 0.0513
Accuracy	LncRNADisease	0.8307 ± 0.0309	0.7571 ± 0.0320	0.8261 ± 0.0523	**0.8444 ± 0.0445**
	MNDR	0.8916 ± 0.0419	0.8685 ± 0.0307	**0.9059 ± 0.0284**	0.8989 ± 0.0317
F1 score	LncRNADisease	0.8079 ± 0.0606	0.7359 ± 0.0525	0.8079 ± 0.0419	**0.8278 ± 0.0363**
	MNDR	0.8833 ± 0.0447	0.8644 ± 0.0265	**0.9002 ± 0.0293**	0.8925 ± 0.0307
AUC	LncRNADisease	0.9107 ± 0.0262	0.8252 ± 0.0308	0.9139 ± 0.0406	**0.9328 ± 0.0243**
	MNDR	0.9384 ± 0.0441	0.9314 ± 0.0216	0.9664 ± 0.0190	**0.9675 ± 0.0147**
AUPR	LncRNADisease	0.8997 ± 0.0575	0.8283 ± 0.0559	0.9209 ± 0.0262	**0.9304 ± 0.0252**
	MNDR	0.9526 ± 0.0250	0.9371 ± 0.0241	**0.9715 ± 0.0134**	0.9709 ± 0.0106

feature bundling. Here, inspired by LightGBM (Ke et al., 2017), we propose a LightGBM-based LDA inference algorithm LDA-LightGBM. First, gradients of all LDPs in the training set are computed, and the $a\%$ LDPs with the smallest gradients are taken as A. Next, a sample set B is constructed by randomly selecting $b \times |A^r|$ samples from the remaining LDPs A^r. Finally, all LDPs are split on the node p_d according to information gain $I_j(p_d)$ on $A \cup B$ using Eq. 8:

$$I_j(p_d) = \frac{1}{p}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i\right)^2}{p_r^j(d)}\right) + \frac{1}{p}\left(\frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i\right)^2}{n_r^j(d)}\right), \quad (8)$$

where $A_l = \{x_i \in A: x_{ij} \le p_d\}$, $A_r = \{x_i \in A: x_{ij} > p_d\}$, $B_l = \{x_i \in B: x_{ij} \le p_d\}$, $B_r = \{x_i \in B: x_{ij} > p_d\}$, and g_i represents the negative gradient.

TABLE 6 Performance of four boosting methods under CV_d.

	Dataset	AdaBoost-CNN	AdaBoost	LightGBM	LDA-SABC
Precision	LncRNADisease	0.8581 ± 0.0502	0.7788 ± 0.0560	0.9134 ± 0.0321	**0.9218 ± 0.0242**
	MNDR	0.9467 ± 0.0224	0.8750 ± 0.0380	0.9358 ± 0.0257	**0.9573 ± 0.0217**
Recall	LncRNADisease	0.8208 ± 0.0514	0.7746 ± 0.0576	0.8669 ± 0.0423	**0.8745 ± 0.0353**
	MNDR	0.9006 ± 0.0458	0.8521 ± 0.0665	0.9156 ± 0.0360	**0.9231 ± 0.0400**
Accuracy	LncRNADisease	0.8476 ± 0.0336	0.7832 ± 0.0288	0.8928 ± 0.0217	**0.9008 ± 0.0232**
	MNDR	0.9280 ± 0.0234	0.8769 ± 0.0177	0.9321 ± 0.0185	**0.9455 ± 0.0146**
F1 score	LncRNADisease	0.8376 ± 0.0378	0.7748 ± 0.0449	0.8884 ± 0.0218	**0.8970 ± 0.0218**
	MNDR	0.9223 ± 0.0260	0.8627 ± 0.0508	0.9254 ± 0.0288	**0.9394 ± 0.0260**
AUC	LncRNADisease	0.9263 ± 0.0226	0.8548 ± 0.0246	0.9615 ± 0.0124	**0.9630 ± 0.0122**
	MNDR	0.9758 ± 0.0107	0.9395 ± 0.0154	0.9825 ± 0.0068	**0.9860 ± 0.0057**
AUPR	LncRNADisease	0.9215 ± 0.0290	0.8453 ± 0.0581	0.9596 ± 0.0147	**0.9605 ± 0.0130**
	MNDR	0.9746 ± 0.0131	0.9290 ± 0.0367	0.9793 ± 0.0144	**0.9836 ± 0.0101**

TABLE 7 Performance of four boosting methods under CV_{ld}.

	Dataset	AdaBoost-CNN	AdaBoost	LightGBM	LDA-SABC
Precision	LncRNADisease	0.8810 ± 0.0285	0.7989 ± 0.0262	0.9012 ± 0.0263	**0.9052 ± 0.0241**
	MNDR	0.9455 ± 0.0115	0.8755 ± 0.0157	0.9426 ± 0.0140	**0.9525 ± 0.0153**
Recall	LncRNADisease	0.9031 ± 0.0242	0.8040 ± 0.0323	0.8893 ± 0.0335	**0.9074 ± 0.0329**
	MNDR	**0.9507 ± 0.0119**	0.8691 ± 0.0230	0.9350 ± 0.0131	0.9459 ± 0.0131
Accuracy	LncRNADisease	0.8901 ± 0.0203	0.8003 ± 0.0214	0.8955 ± 0.0227	**0.9058 ± 0.0183**
	MNDR	0.9479 ± 0.0087	0.8726 ± 0.0129	0.9389 ± 0.0097	**0.9493 ± 0.0109**
F1 score	LncRNADisease	0.8916 ± 0.0194	0.8009 ± 0.0220	0.8948 ± 0.0232	**0.9058 ± 0.0190**
	MNDR	0.9480 ± 0.0087	0.8721 ± 0.0135	0.9387 ± 0.0096	**0.9491 ± 0.0108**
AUC	LncRNADisease	0.9532 ± 0.0144	0.8657 ± 0.0177	0.9575 ± 0.0122	**0.9628 ± 0.0132**
	MNDR	0.9850 ± 0.0043	0.9447 ± 0.0090	0.9839 ± 0.0042	**0.9878 ± 0.0046**
AUPR	LncRNADisease	0.9482 ± 0.0194	0.8610 ± 0.0189	0.9561 ± 0.0119	**0.9606 ± 0.0150**
	MNDR	0.9840 ± 0.0058	0.9454 ± 0.0106	0.9839 ± 0.0041	**0.9881 ± 0.0055**

However, LDA features have high dimensions and multiple zero values, that is, the features cannot simultaneously have nonzero values. To solve this problem, LRI-LightGBM first uses weights to characterize the whole conflict between all LDA features and construct a weighted graph. Subsequently, all LDA features are sorted and are set to a defined bundle or create a new bundle. Finally, all LDPs are classified using Eq. 9:

$$F_{I_q}(\boldsymbol{x}_i) = \sum_{q=1}^{I_q} \gamma_q h_q(\boldsymbol{x}_i), \tag{9}$$

where T_q is the maximum iteration number and $h_q(\boldsymbol{x}_i)$ is the qth basic decision tree.

2.4.3 Ensemble learning

Ensemble learning exhibits strong classification performance compared to individual classifiers. Thus, we combined LDA-AdaBoost-CNN and LDA-LightGBM for LDA identification. For one LDP $\boldsymbol{x}_i$, let $C(\boldsymbol{x}_i)$ and $F(\boldsymbol{x}_i)$ represent its association scores computed by LDA-AdaBoost-CNN and LDA-LightGBM, respectively; its final association probability $p(\boldsymbol{x}_i)$ is obtained Eq. 10:

$$P(\boldsymbol{x}_i) = \alpha C(\boldsymbol{x}_i) + \beta F(\boldsymbol{x}_i), \tag{10}$$

where α and $\beta(\beta = 1 - \alpha)$ are used to evaluate the importance of LDA-AdaBoost-CNN and LDA-LightGBM with respect to the LDA inference performance, respectively.

10.3389/fgene.2024.1356205

TABLE 8 Predicted top 15 lncRNAs associated with NSCLC on LncRNADisease and MNDR.

LncRNADisease			MNDR		
Rank	LncRNA	Evidence	Rank	LncRNA	Evidence
1	HULC	Lnc2Cancer 3.0, RNADisease, and LncRNADisease v3.0	1	PTENP1	Unknown
2	MIAT	Lnc2Cancer 3.0, RNADisease, and LncRNADisease v3.0	2	WRAP53	RNADisease and Lnc2Cancer 3.0
3	MINA	Unknown	3	PRINS	Unknown
4	CCDC26	Unknown	4	MINA	Unknown
5	CRNDE	Lnc2Cancer 3.0, RNADisease, and LncRNADisease v3.0	5	RRP1B	Unknown
6	PCAT1	Lnc2Cancer 3.0, RNADisease, and LncRNADisease v3.0	6	MYCNOS	Unknown
7	HNF1A-AS1	Lnc2Cancer 3.0, RNADisease, and LncRNADisease v3.0	7	DLEU1	LncRNADisease v3.0
8	7SK	Unknown	8	LINC00032	Unknown
9	WT1-AS	LncRNADisease v3.0	9	SNHG16	Lnc2Cancer 3.0, RNADisease, and LncRNADisease v3.0
10	GHET1	RNADisease and LncRNADisease v3.0	10	SRA1	Unknown
11	SOX2-OT	RNADisease, and LncRNADisease v3.0	11	7SK	Unknown
12	PTENP1	Unknown	12	MKRN3-AS1	Unknown
13	CASC2	Lnc2Cancer 3.0, RNADisease, and LncRNADisease v3.0	13	DISC2	Unknown
14	HIF1A-AS2	LncRNADisease v3.0	14	NRON	Unknown
15	LSINCT5	Lnc2Cancer 3.0, RNADisease, and LncRNADisease v3.0	15	MESTIT1	Unknown

TABLE 9 Predicted top 15 lncRNAs associated with LUAD on LncRNADisease and MNDR.

LncRNADisease			MNDR		
Rank	LncRNA	Evidence	Rank	LncRNA	Evidence
1	CDKN2B-AS1	RNADisease and Lnc2Cancer 3.0	1	TUG1	RNADisease and LncRNADisease v3.0
2	PVT1	Lnc2Cancer 3.0, RNADisease, and LncRNADisease v3.0	2	CDKN2B-AS1	RNADisease and Lnc2Cancer 3.0
3	H19	Lnc2Cancer 3.0 and LncRNADisease v3.0	3	PVT1	Lnc2Cancer 3.0, RNADisease, and LncRNADisease v3.0
4	TUG1	RNADisease and LncRNADisease v3.0	4	UCA1	Lnc2Cancer 3.0, RNADisease, and LncRNADisease v3.0
5	CCAT2	Lnc2Cancer 3.0 and LncRNADisease v3.0	5	KCNQ1OT1	RNADisease and Lnc2Cancer 3.0
6	XIST	RNADisease and Lnc2Cancer 3.0	6	CBR3-AS1	LncRNADisease v3.0
7	HULC	Unknown	7	SNHG4	Unknown
8	DANCR	Lnc2Cancer 3.0, RNADisease, and LncRNADisease v3.0	8	WT1-AS	LncRNADisease v3.0
9	MINA	Unknown	9	SPRY4-IT1	RNADisease and Lnc2Cancer 3.0
10	BCYRN1	Unknown	10	BCYRN1	Unknown
11	BANCR	Unknown	11	HULC	Unknown
12	PANDAR	Unknown	12	PTENP1	Unknown
13	CASC2	Lnc2Cancer 3.0, RNADisease, and LncRNADisease v3.0	13	HIF1A-AS1	RNADisease
14	LSINCT5	Unknown	14	CCAT2	Lnc2Cancer 3.0, and LncRNADisease v3.0
15	CCDC26	Unknown	15	HIF1A-AS2	LncRNADisease v3.0

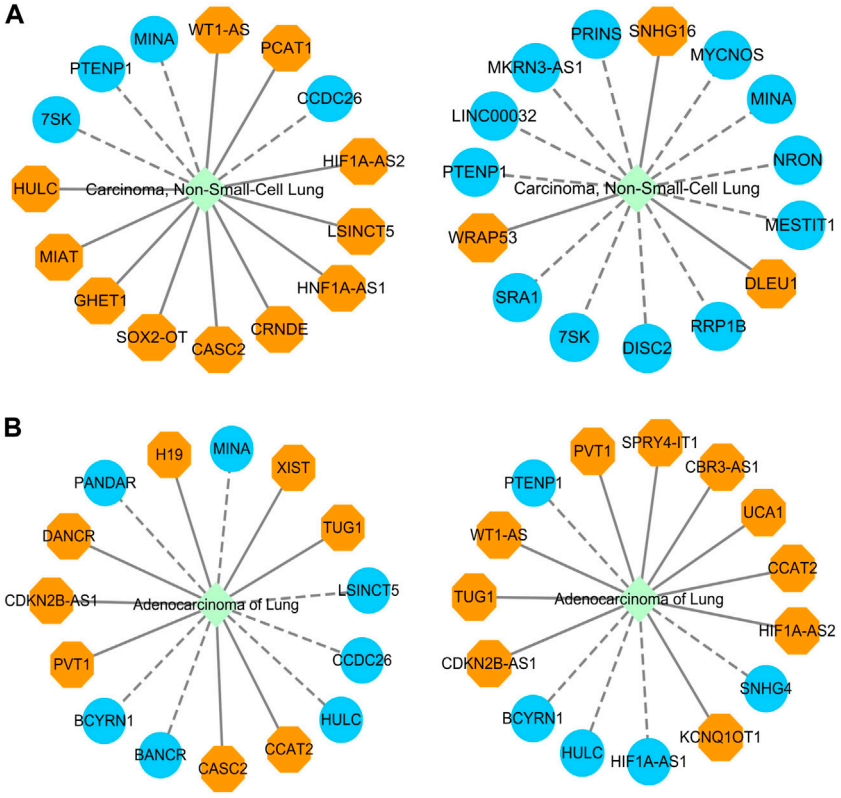

FIGURE 4
(A) Inferred top 15 lncRNAs associated with NSCLC on LncRNADisease and MNDR databases. **(B)** Inferred top 15 lncRNAs associated with LUAD on LncRNADisease and MNDR databases.

3 Results

3.1 Experimental settings

To assess the LDA inference performance of LDA-SABC, we implemented three fivefold CVs to compare it with four representative LDA prediction approaches, namely, SDLDA (Zeng et al., 2020), LDNFSGB (Zhang et al., 2020), IPCARF (Zhu et al., 2021), and LDASR (Guo et al., 2019). The parameters in the above four methods were derived from their corresponding literatures. For the LDA-SABC model, we set n_estimators, learning rate, and epochs to 100, 0.1, and 10, respectively, in LDA-AdaBoost-CNN and n_estimators and learning rate to 100 and 0.1, respectively, in LRI-LightGBM. The dimension d of an LDA feature vector was set to 64.

3.2 Comparison with four classical LDA prediction methods

We used six evaluation metrics (precision, recall, accuracy, F1 score, AUC, and AUPR (Shen et al., 2022; Liu et al., 2023; Qiu et al., 2023a)) to assess the performance of LDA-SABC and four other LDA prediction algorithms (SDLDA, LDNFSGB, IPCARF, and LDASR) under three different fivefold cross validations. The three CVs are fivefold CV on lncRNAs (CV_l), five-fold CV on diseases (CV_d), and fivefold CV on LDPs (CV_{ld}). The details refer to Peng et al. (2024a). Tables 2–4 depict the performance of LDA-SABC and four other methods on two databases (i.e., LncRNADisease and MNDR) under the three CVs. Figure 2 characterizes the corresponding ROC and precision–recall (PR) curves.

CV_l was used to compare the performance of LDA-SABC with SDLDA, LDNFSGB, LDASR, and IPCAF when identifying diseases linking to a new lncRNA. Under CV_l, all five methods randomly selected 80% of lncRNAs as the training set and used the remaining as the test set. The results are listed in Table 2 and Figure 2. We found that LDA-SABC outperformed in terms of precision, recall, accuracy, F1 score, AUC, and AUPR compared with the four classical LDA prediction algorithms. For example, LDA-SABC obtained the highest AUC values of 0.9328 and 0.9675, outperforming by 13.05% and 3.09% compared to those of the second best algorithm, on the LncRNADisease and MNDR databases, respectively. It also calculated the highest AUPR values of 0.9304 and 0.9703, outperforming by 8.43% and 1.76% compared to those of the second best algorithm, respectively. These results imply that LDA-SABC could accurately capture the underlying diseases linking to a new lncRNA.

CV_d was applied to compare the performance of LDA-SABC with SDLDA, LDNFSGB, LDASR, and IPCAF when identifying lncRNAs linking to a new disease. Under CV_d, all five methods randomly selected 80% of diseases as the training set and used the remaining as the test set. As demonstrated in Table 3 and Figure 2, LDA-SABC significantly surpassed four other algorithms on the two

datasets. For example, LDA-SABC obtained the highest AUC values of 0.9630 and 0.9860, outperforming by 8.42% and 3.01% compared to those of the second best algorithm (i.e., SDLDA), on the LncRNADisease and MNDR databases, respectively. It also calculated the highest AUPR values of 0.9605 and 0.9836, outperforming by 6.71% and 2.75% compared to those of the second best algorithm (i.e., SDLDA), on the LncRNADisease and MNDR databases, respectively. These results suggest that LDA-SABC could accurately infer potential lncRNAs linking to a new disease.

CV_{ld} is used to compare the performance of all five LDA inference methods when identifying new LDAs from unknown LDPs. Under CV_{ld}, all five methods randomly selected 80% of LDPs as the training set and used the remaining as the test set. As demonstrated in Table 4 and Figure 2, LDA-SABC significantly improved LDA prediction in comparison with the four other methods. For example, LDA-SABC achieved the highest AUC values of 0.9628 and 0.9878, outperforming by 8.54% and 3.18% compared to those of the second best algorithm (i.e., SDLDA), on the LncRNADisease and MNDR databases, respectively. It also calculated the highest AUPR values of 0.9606 and 0.9881, outperforming by 6.54% and 2.42% compared to those of the second best algorithm (i.e., SDLDA), on the LncRNADisease and MNDR databases, respectively. Thus, LDA-SABC could more accurately infer the underlying LDAs through known LDAs.

3.3 Ablation study

LDA-SABC combined AdaBoost-CNN and LightGBM for LDA prediction. In model Ensemble, α and β were used to evaluate the effects of LDA-AdaBoost-CNN and LDA-LightGBM on the LDA inference performance, respectively. As shown in Figure 3, when α was set to 0, 0.2, 0.4, 0.6, 0.8, and 1, respectively, LDA-SABC achieved the best performance on the LncRNADisease and MNDR databases under fivefold CVs on lncRNAs, diseases, and LDPs. Supplementary Tables S1–S3 show the detailed performance of LDA-SABC when α was set to the above six values, respectively. Thus, we set α and β to 0.4 and 0.6, respectively.

To better understand the performance of ensemble learning, we compared LDA-SABC with other boosting algorithms, i.e., AdaBoost-CNN, AdaBoost, and LightGBM, under three different CVs. The boosting algorithms used the same feature extraction procedures as LDA-SABC except for using different boosting models for classifying unknown LDPs. Tables 5–7 show their LDA prediction performance under fivefold CVs on lncRNAs, diseases, and LDPs, respectively. The results demonstrate that LDA-SABC computed the best LDA inference accuracy on the two LDA databases under the three CVs in most cases, thereby elucidating the powerful LDP classification performance of our proposed ensemble learning model with LightGBM and AdaBoost-CNN.

3.4 Case study

Lung cancer is one of the most frequent malignant tumors and has a very high incidence and mortality rate. More importantly, its 5-year survival rate is much lower compared to other leading cancers (Huang et al., 2023). Non-small-cell lung cancer and lung adenocarcinoma (LUAD) are two prevalent lung cancers, wherein NSCLC accounts for approximately 85% of lung cancers (Tan et al., 2023) and LUAD is the most predominant subtype (Li et al., 2023). lncRNAs have close associations with various complex diseases and are potential biomarkers of many types of cancers. Therefore, it is very important to discover potential lncRNAs and further provide therapeutic options for lung cancer.

Through performance comparison, we validated the accurate LDA classification performance of LDA-SABC. Subsequently, we utilized LDA-SABC to discover the potential lncRNAs for NSCLC and LUAD. We computed the association probabilities between all lncRNAs and NSCLC and LUAD. Tables 8 and 9 demonstrate the top 15 lncRNAs with the highest association probability with NSCLC and LUAD among all lncRNAs which have no observed association with NSCLC and LUAD on the LncRNADisease and MNDR databases, respectively. Figure 4 elucidates two predicted LDA networks for NSCLC and LUAD.

Among the inferred top 15 lncRNAs associated with NSCLC, 11 and 3 lncRNAs, predicted on the LncRNADisease and MNDR databases, have been confirmed by Lnc2Cancer 3.0 (Gao et al., 2021), LncRNADisease v3.0 (Lin et al., 2023b), and/or RNADisease (Chen et al., 2023), respectively. Particularly, 7SK was linked to NSCLC, which was ranked 8 and 11, respectively. lncRNA 7SK acts as a transcription regulator. Its exosomal delivery could inhibit the proliferation and aggressiveness of tumor cells in triple-negative breast cancer (Farhadi et al., 2023). Furthermore, 7SK could suppress human tongue squamous carcinoma (Zhang et al., 2021). 7SK was predicted to be associated with NSCLC, which needs further confirmation.

Among the inferred top 15 lncRNAs associated with LUAD, 8 and 11 lncRNAs, predicted on the LncRNADisease and MNDR databases, have been reported by Lnc2Cancer 3.0, LncRNADisease v3.0, and/or RNADisease, respectively. We found that HULC could be associated with LUAD, which was ranked 7 and 11, respectively. HULC is an oncogenic lncRNA and may serve as a prognostic biomarker of hepatocellular carcinoma development (Liu S. et al., 2023). Moreover, it displays the potential to be a novel biomarker for assisting acute myocardial infarction diagnosis when combined with other biomarkers (Xie et al., 2022).

4 Discussion and conclusion

Inferring possible LDAs can advance our understanding of human complex diseases in the context of lncRNAs. However, traditional experimental techniques for LDA prediction are costly, laborious, and time-consuming, which restricts the number of the verified LDAs. Thus, substantive computational frameworks have been exploited. In this manuscript, we proposed a novel computational LDA inference framework LDA-SABC by combining SVD and an ensemble model of LightGBM and AdaBoost-CNN.

LDA-SABC first acquired LDP linear features using SVD. Next, it computed the association probability for each LDP with LDA-LightGBM and LDA-AdaBoost-CNN. Finally, all LDPs were classified through ensemble learning. To illustrate the effectiveness of LDA-SABC, it was compared with four classical

computational methods (SDLDA, LDNFSGB, IPCARF, and LDASR) under three CVs. The results elucidated that its performance was significantly improved. To validate the performance of LDA-SABC, we further performed case studies to find potential biomarkers of NSCLC and LUAD and discovered the top 15 lncRNAs linked to them from all unknown LDPs. The results demonstrated that among the inferred top lncRNAs reported by RNADisease, LncRNADisease v3.0, or/and Lnc2Cancer 3.0 databases, 7SK and HULC could have a relationship with NSCLC and LUAD, respectively.

The novelty of this study is the use of SVD for extracting LDP features and designing an ensemble model with LightGBM and AdaBoost-CNN for improving the LDA prediction accuracy. Differing from traditional LDA prediction performance validation, LDA-SABC was assessed under fivefold CVs on lncRNAs, diseases, and LDPs. However, in the process of negative LDA selection, a random selection strategy was adopted, which affected the overall performance of the model. In the future, we will design a reasonable negative LDA selection strategy based on positive-unlabeled learning. More importantly, we will still explore a stronger classification model for LDP classification by integrating various data and deep learning methods. We hope that our proposed LDA-SABC could contribute to the lncRNA biomarker discovery of various complex diseases, especially cancers, and further help find new therapeutic options for various types of cancers.

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found at: https://github.com/plhhnu/LDA-SABC.

Author contributions

LZh: writing–original draft and writing–review and editing. XP: writing–original draft and writing–review and editing. LP: writing–original draft and writing–review and editing. LZe: writing–original draft and writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. LZh received funding from the National Natural Science Foundation of China under grant no. 62072172 and Natural Science Foundation of Hunan Province under grant no. 2021JJ30219. LP was supported by the National Natural Science Foundation of China under grant no. 61803151.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1356205/full#supplementary-material

References

Abdi, H. (2007). Singular value decomposition (svd) and generalized singular value decomposition. *Encycl. Meas. statistics* 907, 912.

Adewunmi, O., Shen, Y., Zhang, X. H.-F., and Rosen, J. M. (2023). Targeted inhibition of lncrna malat1 alters the tumor immune microenvironment in preclinical syngeneic mouse models of triple-negative breast cancer. *Cancer Immunol. Res.* 11, 1462–1479. doi:10.1158/2326-6066.CIR-23-0045

Cao, Y., Xiao, J., Sheng, N., Qu, Y., Wang, Z., Sun, C., et al. (2023). X-lda: an interpretable and knowledge-informed heterogeneous graph learning framework for lncrna-disease association prediction. *Comput. Biol. Med.* 167, 107634. doi:10.1016/j.compbiomed.2023.107634

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., et al. (2012). Lncrnadisease: a database for long-non-coding rna-associated diseases. *Nucleic acids Res.* 41, D983–D986. doi:10.1093/nar/gks1099

Chen, J., Lin, J., Hu, Y., Ye, M., Yao, L., Wu, L., et al. (2023a). 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, X. (2015a). Katzlda: katz measure for the lncrna-disease association prediction. *Sci. Rep.* 5, 16840–16911. doi:10.1038/srep16840

Chen, X. (2015b). Predicting lncrna-disease associations and constructing lncrna functional similarity network based on the information of mirna. *Sci. Rep.* 5, 13186–13211. doi:10.1038/srep13186

Chen, X., Clarence Yan, 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, 11338–11412. doi:10.1038/srep11338

Chen, X., and Huang, L. (2022). Computational model for ncrna research. *Briefings Bioinforma.* 23, bbac472. doi:10.1093/bib/bbac472

Chen, X., and Huang, L. (2023). Computational model for disease research. *Briefings Bioinforma.* 24, bbac615. doi:10.1093/bib/bbac615

Chen, X., Sun, L.-G., and Zhao, Y. (2021). Ncmcmda: mirna–disease association prediction through neighborhood constraint matrix completion. *Briefings Bioinforma.* 22, 485–496. doi:10.1093/bib/bbz159

Chen, X., Yan, C. C., Zhang, X., and You, Z.-H. (2017). Long non-coding rnas and complex diseases: from experimental results to computational models. *Briefings Bioinforma.* 18, 558–576. doi:10.1093/bib/bbw060

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.-g., Yang, X., Li, C., Lin, X., and Zhang, W. (2023b). Non-coding rna identification with pseudo rna sequences and feature representation learning. *Comput. Biol. Med.* 165, 107355. doi:10.1016/j.compbiomed.2023.107355

Cui, T., Zhang, L., Huang, Y., Yi, Y., Tan, P., Zhao, Y., et al. (2018). Mndr v2. 0: an updated resource of ncrna–disease associations in mammals. *Nucleic acids Res.* 46, D371–D374. doi:10.1093/nar/gkx1025

Dinger, M. E., Pang, K. C., Mercer, T. R., Crowe, M. L., Grimmond, S. M., and Mattick, J. S. (2009). Nred: a database of long noncoding rna expression. *Nucleic acids Res.* 37, D122–D126. doi:10.1093/nar/gkn617

Farhadi, S., Mohammadi-Yeganeh, S., Kiani, J., Hashemi, S. M., Koochaki, A., Sharifi, K., et al. (2023). Exosomal delivery of 7sk long non-coding rna suppresses viability, proliferation, aggressiveness and tumorigenicity in triple negative breast cancer cells. *Life Sci.* 322, 121646. doi:10.1016/j.lfs.2023.121646

Fu, G., Wang, J., Domeniconi, C., and Yu, G. (2018). Matrix factorization-based data fusion for the prediction of lncrna–disease associations. *Bioinformatics* 34, 1529–1537. doi:10.1093/bioinformatics/btx794

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

Guo, Z.-H., You, Z.-H., Wang, Y.-B., Yi, H.-C., and Chen, Z.-H. (2019). A learning-based method for lncrna-disease association identification combing similarity information and rotation forest. *IScience* 19, 786–795. doi:10.1016/j.isci.2019.08.030

Han, S., Fu, H., Wu, Y., Zhao, G., Song, Z., Huang, F., et al. (2023). Himgnn: a novel hierarchical molecular graph representation learning framework for property prediction. *Briefings Bioinforma.* 24, bbad305. doi:10.1093/bib/bbad305

Hastie, T., Rosset, S., Zhu, J., and Zou, H. (2009). Multi-class adaboost. *Statistics its Interface* 2, 349–360. doi:10.4310/sii.2009.v2.n3.a8

Hu, H., Feng, Z., Lin, H., Cheng, J., Lyu, J., Zhang, Y., et al. (2023). Gene function and cell surface protein association analysis based on single-cell multiomics data. *Comput. Biol. Med.* 157, 106733. doi:10.1016/j.compbiomed.2023.106733

Hu, Y., Zhou, M., Shi, H., Ju, H., Jiang, Q., and Cheng, L. (2017). Measuring disease similarity and predicting disease-related ncrnas by a novel method. *BMC Med. genomics* 10, 71–74. doi:10.1186/s12920-017-0315-9

Huang, S., Yang, J., Shen, N., Xu, Q., and Zhao, Q. (2023). Artificial intelligence in lung cancer diagnosis and prognosis: current application and future perspective. *Seminars Cancer Biol.* 89, 30–37. doi:10.1016/j.semcancer.2023.01.006

Huo, J., Cai, J., Guan, G., Liu, H., and Wu, L. (2021). A ferroptosis and pyroptosis molecular subtype-related signature applicable for prognosis and immune microenvironment estimation in hepatocellular carcinoma. *Front. Cell Dev. Biol.* 9, 761839. doi:10.3389/fcell.2021.761839

Jiang, J., Xu, J., Liu, Y., Song, B., Guo, X., Zeng, X., et al. (2023). Dimensionality reduction and visualization of single-cell rna-seq data with an improved deep variational autoencoder. *Briefings Bioinforma.* 24, bbad152. doi:10.1093/bib/bbad152

Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., et al. (2010). Prioritization of disease micrornas through a human phenome-micrornaome network. *BMC Syst. Biol.* 4, S2–S9. doi:10.1186/1752-0509-4-S1-S2

Jiang, Q., Ma, R., Wang, J., Wu, X., Jin, S., Peng, J., et al. (2015). Lncrna2function: a comprehensive resource for functional investigation of human lncrnas based on rna-seq data. *BMC genomics BioMed Cent.* 16, S2–S11. doi:10.1186/1471-2164-16-S3-S2

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: a highly efficient gradient boosting decision tree. *Adv. neural Inf. Process. Syst.* 30.

Lan, W., Lai, D., Chen, Q., Wu, X., Chen, B., Liu, J., et al. (2020). Ldicdl: lncrna-disease association identification based on collaborative deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 19, 1715–1723. doi:10.1109/TCBB.2020.3034910

Li, C., Cui, Y., Liu, L.-F., Ren, W.-B., Li, Q.-Q., Zhou, X., et al. (2017). High expression of long noncoding rna malat1 indicates a poor prognosis and promotes clinical progression and metastasis in bladder cancer. *Clin. Genitourin. cancer* 15, 570–576. doi:10.1016/j.clgc.2017.05.001

Li, J., Li, J., Kong, M., Wang, D., Fu, K., and Shi, J. (2021). Svdnvlda: predicting lncrna-disease associations by singular value decomposition and node2vec. *BMC Bioinforma.* 22, 538–618. doi:10.1186/s12859-021-04457-1

Li, L., Cai, Q., Wu, Z., Li, X., Zhou, W., Lu, L., et al. (2023). Bioinformatics construction and experimental validation of a cuproptosis-related lncrna prognostic model in lung adenocarcinoma for immunotherapy response prediction. *Sci. Rep.* 13, 2455. doi:10.1038/s41598-023-29684-9

Liang, Q., Zhang, W., Wu, H., and Liu, B. (2023). Lncrna-disease association identification using graph auto-encoder and learning to rank. *Briefings Bioinforma.* 24, bbac539. doi:10.1093/bib/bbac539

Lin, X., Dai, L., Zhou, Y., Yu, Z.-G., Zhang, W., Shi, J.-Y., et al. (2023a). Comprehensive evaluation of deep and graph learning on drug–drug interactions prediction. *Briefings Bioinforma.* 24, bbad235. doi:10.1093/bib/bbad235

Lin, X., Lu, Y., Zhang, C., Cui, Q., Tang, Y.-D., Ji, X., et al. (2023b). Lncrnadisease v3. 0: an updated database of long non-coding rna-associated diseases. *Nucleic Acids Res.*, gkad828. doi:10.1093/nar/gkad828

Liu, J.-X., Cui, Z., Gao, Y.-L., and Kong, X.-Z. (2020). Wgrcmf: a weighted graph regularized collaborative matrix factorization method for predicting novel lncrna-disease associations. *IEEE J. Biomed. Health Inf.* 25, 257–265. doi:10.1109/JBHI.2020.2985703

Liu, N., Zhang, Z., Wu, Y., Wang, Y., and Liang, Y. (2023a). Crbsp: prediction of circrna-rbp binding sites based on multimodal intermediate fusion. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 20, 2898–2906. doi:10.1109/TCBB.2023.3272400

Liu, S., Huttad, L., He, G., He, W., Liu, C., Cai, D., et al. (2023b). Long noncoding rna hulc regulates the nf-κb pathway and represents a promising prognostic biomarker in liver cancer. *Cancer Med.* 12, 5124–5136. doi:10.1002/cam4.5263

Liu, S. J., Dang, H. X., Lim, D. A., Feng, F. Y., and Maher, C. A. (2021). Long noncoding rnas in cancer metastasis. *Nat. Rev. Cancer* 21, 446–460. doi:10.1038/s41568-021-00353-1

Liu, X., and Zhang, W. (2023). A subcomponent-guided deep learning method for interpretable cancer drug response prediction. *PLOS Comput. Biol.* 19, e1011382. doi:10.1371/journal.pcbi.1011382

Lu, C., and Xie, M. (2023). Ldaexc: lncrna–disease associations prediction with deep autoencoder and xgboost classifier. *Interdiscip. Sci. Comput. Life Sci.* 1–13. doi:10.1007/s12539-023-00573-z

Ma, Y. (2022). Deepmne: deep multi-network embedding for lncrna-disease association prediction. *IEEE J. Biomed. Health Inf.* 26, 3539–3549. doi:10.1109/JBHI.2022.3152619

Mao, T.-L., Fan, M.-H., Dlamini, N., and Liu, C.-L. (2021). Lncrna malat1 facilitates ovarian cancer progression through promoting chemoresistance and invasiveness in the tumor microenvironment. *Int. J. Mol. Sci.* 22, 10201. doi:10.3390/ijms221910201

Min, L., Zhu, T., Lv, B., An, T., Zhang, Q., Shang, Y., et al. (2022). Exosomal lncrna rp5-977b1 as a novel minimally invasive biomarker for diagnosis and prognosis in non-small cell lung cancer. *Int. J. Clin. Oncol.* 27, 1013–1024. doi:10.1007/s10147-022-02129-5

Mo, M., Ma, X., Luo, Y., Tan, C., Liu, B., Tang, P., et al. (2022). Liver-specific lncrna fam99a may be a tumor suppressor and promising prognostic biomarker in hepatocellular carcinoma. *BMC cancer* 22, 1098–1119. doi:10.1186/s12885-022-10186-2

Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., et al. (2016). Lnc2cancer: a manually curated database of experimentally supported lncrnas associated with various human cancers. *Nucleic acids Res.* 44, D980–D985. doi:10.1093/nar/gkv1094

Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., et al. (2019). A learning-based framework for mirna-disease association identification using neural networks. *Bioinformatics* 35, 4364–4371. doi:10.1093/bioinformatics/btz254

Peng, L., He, X., Peng, X., Li, Z., and Zhang, L. (2023a). Stgnnks: identifying cell types in spatial transcriptomics data based on graph neural network, denoising auto-encoder, and k-sums clustering. *Comput. Biol. Med.* 166, 107440. doi:10.1016/j.compbiomed.2023.107440

Peng, L., Huang, L., Su, Q., Tian, G., Chen, M., and Han, G. (2024a). Lda-vghb: identifying potential lncrna-disease associations with singular value decomposition, variational graph auto-encoder and heterogeneous Newton boosting machine. *Briefings Bioinforma.* 5, bbad466. doi:10.1093/bib/bbad466

Peng, L., Tan, J., Tian, X., and Zhou, L. (2022a). Enanndeep: an ensemble-based lncrna–protein interaction prediction framework with adaptive k-nearest neighbor classifier and deep models. *Interdiscip. Sci. Comput. Life Sci.* 14, 209–232. doi:10.1007/s12539-021-00483-y

Peng, L., Tan, J., Xiong, W., Zhang, L., Wang, Z., Yuan, R., et al. (2023b). Deciphering ligand–receptor-mediated intercellular communication based on ensemble deep learning and the joint scoring strategy from single-cell transcriptomic data. *Comput. Biol. Med.* 163, 107137. doi:10.1016/j.compbiomed.2023.107137

Peng, L., Wang, F., Wang, Z., Tan, J., Huang, L., Tian, X., et al. (2022b). Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies. *Briefings Bioinforma.* 23, bbac234. doi:10.1093/bib/bbac234

Peng, L., Xiong, W., Han, C., Li, Z., and Chen, X. (2024b). Celldialog: a computational framework for ligand-receptor-mediated cell-cell communication analysis. *IEEE J. Biomed. Health Inf.* 28, 580–591. doi:10.1109/jbhi.2023.3333828

Peng, L., Yuan, R., Han, C., Han, G., Tan, J., Wang, Z., et al. (2023c). Cellenboost: a boosting-based ligand-receptor interaction identification model for cell-to-cell communication inference. *IEEE Trans. NanoBioscience* 22, 705–715. doi:10.1109/TNB.2023.3278685

Qi, R., and Zou, Q. (2023). Editorial: machine learning methods in single-cell immune and drug response prediction. *Front. Genet.* 14, 1233078. doi:10.3389/fgene.2023.1233078

Qiu, S., Liu, R., and Liang, Y. (2023a). Gr-m6a: prediction of n6-methyladenosine sites in mammals with molecular graph and residual network. *Comput. Biol. Med.* 163, 107202. doi:10.1016/j.compbiomed.2023.107202

Qiu, S., Wang, M., Yang, Y., Yu, G., Wang, J., Yan, Z., et al. (2023b). Meta multi-instance multi-label learning by heterogeneous network fusion. *Inf. Fusion* 94, 272–283. doi:10.1016/j.inffus.2023.02.010

Shen, L., Liu, F., Huang, L., Liu, G., Zhou, L., and Peng, L. (2022). Vda-rwlrls: an anti-sars-cov-2 drug prioritizing framework combining an unbalanced bi-random walk and laplacian regularized least squares. *Comput. Biol. Med.* 140, 105119. doi:10.1016/j.compbiomed.2021.105119

Sheng, N., Huang, L., Lu, Y., Wang, H., Yang, L., Gao, L., et al. (2023). Data resources and computational methods for lncrna-disease association prediction. *Comput. Biol. Med.* 153, 106527. doi:10.1016/j.compbiomed.2022.106527

Sun, F., Sun, J., and Zhao, Q. (2022). A deep learning method for predicting metabolite–disease associations via graph neural network. *Briefings Bioinforma.* 23, bbac266. doi:10.1093/bib/bbac266

Taherkhani, A., Cosma, G., and McGinnity, T. M. (2020). Adaboost-cnn: an adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning. *Neurocomputing* 404, 351–366. doi:10.1016/j.neucom.2020.03.064

Tan, K., Zhang, C., He, Z., and Zeng, P. (2023). Construction of an anoikis-associated lncrna-mirna-mrna network reveals the prognostic role of β-elemene in non-small cell lung cancer. *Sci. Rep.* 13, 20185. doi:10.1038/s41598-023-46480-7

Tang, R., Wu, Z., Rong, Z., Xu, J., Wang, W., Zhang, B., et al. (2022). Ferroptosis-related lncrna pairs to predict the clinical outcome and molecular characteristics of pancreatic ductal adenocarcinoma. *Briefings Bioinforma.* 23, bbab388. doi:10.1093/bib/bbab388

Tang, Y., Li, C., Zhang, Y.-J., and Wu, Z.-H. (2021). Ferroptosis-related long non-coding rna signature predicts the prognosis of head and neck squamous cell carcinoma. *Int. J. Biol. Sci.* 17, 702–711. doi:10.7150/ijbs.55552

Wang, F., Yang, H., Wu, Y., Peng, L., and Li, X. (2023a). Saelgmda: identifying human microbe–disease associations based on sparse autoencoder and lightgbm. *Front. Microbiol.* 14, 1207209. doi:10.3389/fmicb.2023.1207209

Wang, H., Tang, J., Ding, Y., and Guo, F. (2021a). Exploring associations of non-coding rnas in human diseases via three-matrix factorization with hypergraph-regular terms on center kernel alignment. *Briefings Bioinforma.* 22, bbaa409. doi:10.1093/bib/bbaa409

Wang, L., Shang, M., Dai, Q., and He, P.-a. (2022a). Prediction of lncrna-disease association based on a laplace normalized random walk with restart algorithm on heterogeneous networks. *BMC Bioinforma.* 23, 5–20. doi:10.1186/s12859-021-04538-1

Wang, M.-N., You, Z.-H., Wang, L., Li, L.-P., and Zheng, K. (2021b). Ldgrnmf: lncrna-disease associations prediction based on graph regularized non-negative matrix factorization. *Neurocomputing* 424, 236–245. doi:10.1016/j.neucom.2020.02.062

Wang, T., Sun, J., and Zhao, Q. (2023b). Investigating cardiotoxicity related with herg channel blockers using molecular fingerprints and graph attention mechanism. *Comput. Biol. Med.* 153, 106464. doi:10.1016/j.compbiomed.2022.106464

Wang, W., Zhang, L., Sun, J., Zhao, Q., and Shuai, J. (2022b). Predicting the potential human lncrna-mirna interactions based on graph convolution network with conditional random field. *Briefings Bioinforma.* 23, bbac463. doi:10.1093/bib/bbac463

Wang, Y., Yu, G., Domeniconi, C., Wang, J., Zhang, X., and Guo, M. (2019). Selective matrix factorization for multi-relational data fusion. *Int. Conf. database Syst. Adv. Appl.* 11446, 313–329. doi:10.1007/978-3-030-18576-3_19

Wang, Y., Yu, G., Wang, J., Fu, G., Guo, M., and Domeniconi, C. (2020). Weighted matrix factorization on multi-relational data for lncrna-disease association prediction. *Methods* 173, 32–43. doi:10.1016/j.ymeth.2019.06.015

Wei, H., Liao, Q., and Liu, B. (2020). ilncrnadis-fb: identify lncrna-disease associations by fusing biological feature blocks through deep neural network. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 18, 1946–1957. doi:10.1109/TCBB.2020.2964221

Wu, H., Wu, Y., Jiang, Y., Zhou, B., Zhou, H., Chen, Z., et al. (2022). schicstackl: a stacking ensemble learning-based method for single-cell hi-c classification using cell embedding. *Briefings Bioinforma.* 23, bbab396. doi:10.1093/bib/bbab396

Xi, W.-Y., Zhou, F., Gao, Y.-L., Liu, J.-X., and Zheng, C.-H. (2022). Ldcmfc: predicting long non-coding rna and disease association using collaborative matrix factorization based on correntropy. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 20, 1774–1782. doi:10.1109/TCBB.2022.3215194

Xie, G., Jiang, J., and Sun, Y. (2020a). Lda-lnsubrw: lncrna-disease association prediction based on linear neighborhood similarity and unbalanced bi-random walk. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 19, 989–997. doi:10.1109/TCBB.2020.3020595

Xie, G., Wu, C., Gu, G., and Huang, B. (2020b). Haubrw: hybrid algorithm and unbalanced bi-random walk for predicting lncrna-disease associations. *Genomics* 112, 4777–4787. doi:10.1016/j.ygeno.2020.08.024

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. *Briefings Bioinforma.* 24, bbac595. doi:10.1093/bib/bbac595

Xie, J., Liao, W., Chen, W., Lai, D., Tang, Q., and Li, Y. (2022). Circulating long non-coding rna ttty15 and hulc serve as potential novel biomarkers for predicting acute myocardial infarction. *BMC Cardiovasc. Disord.* 22, 86. doi:10.1186/s12872-022-02529-5

Xin, R., Hu, B., Qu, D., and Chen, D. (2023). WITHDRAWN: oncogenic lncRNA MALAT-1 recruits E2F1 to upregulate RAD51 expression and thus promotes cell autophagy and tumor growth in non-small cell lung cancer. *Pulm. Pharmacol. Ther.*, 102199. doi:10.1016/j.pupt.2023.102199

Xiong, Z., Liu, S., Huang, F., Wang, Z., Liu, X., Zhang, Z., et al. (2023). Multi-relational contrastive learning graph neural network for drug-drug interaction event prediction. *Proc. AAAI Conf. Artif. Intell.* 37, 5339–5347. doi:10.1609/aaai.v37i4.25665

Xu, Z., Song, L., Liu, S., and Zhang, W. (2024). Deepcrbp: improved predicting function of circrna-rbp binding sites with deep feature learning. *Front. Comput. Sci.* 18, 182907. doi:10.1007/s11704-023-2798-1

Yu, G., Wang, Y., Wang, J., Domeniconi, C., Guo, M., and Zhang, X. (2020). Attributed heterogeneous network fusion via collaborative matrix tri-factorization. *Inf. Fusion* 63, 153–165. doi:10.1016/j.inffus.2020.06.012

Yu, H., Shen, Z.-A., Zhou, Y.-K., and Du, P.-F. (2022). Recent advances in predicting protein-lncrna interactions using machine learning methods. *Curr. Gene Ther.* 22, 228–244. doi:10.2174/1566523221666210712190718

Yu, J., Xuan, Z., Feng, X., Zou, Q., and Wang, L. (2019). A novel collaborative filtering model for lncrna-disease association prediction based on the naïve bayesian classifier. *BMC Bioinforma.* 20, 396–413. doi:10.1186/s12859-019-2985-0

Zeng, M., Lu, C., Zhang, F., Li, Y., Wu, F.-X., Li, Y., et al. (2020). Sdlda: lncrna-disease association prediction based on singular value decomposition and deep learning. *Methods* 179, 73–80. doi:10.1016/j.ymeth.2020.05.002

Zhang, B., Min, S., Guo, Q., Huang, Y., Guo, Y., Liang, X., et al. (2021). 7sk acts as an anti-tumor factor in tongue squamous cell carcinoma. *Front. Genet.* 12, 642969. doi:10.3389/fgene.2021.642969

Zhang, P., and Wu, H. (2023). Ichrom-deep: an attention-based deep learning model for identifying chromatin interactions. *IEEE J. Biomed. Health Inf.* 27, 4559–4568. doi:10.1109/JBHI.2023.3292299

Zhang, Q., Liu, J., Zhang, W., Yang, F., Yang, Z., and Zhang, X. (2024). A multi-stream network for retrosynthesis prediction. *Front. Comput. Sci.* 18, 182906. doi:10.1007/s11704-023-3103-z

Zhang, Y., Ye, F., Xiong, D., and Gao, X. (2020). Ldnfsgb: prediction of long non-coding rna and disease association using network feature similarity and gradient boosting. *BMC Bioinforma.* 21, 377–427. doi:10.1186/s12859-020-03721-0

Zhang, Z., Xu, J., Wu, Y., Liu, N., Wang, Y., and Liang, Y. (2023). Capsnet-lda: predicting lncrna-disease associations using attention mechanism and capsule network based on multi-view data. *Briefings Bioinforma.* 24, bbac531. doi:10.1093/bib/bbac531

Zhao, X., Wu, J., Zhao, X., and Yin, M. (2023). Multi-view contrastive heterogeneous graph attention network for lncrna–disease association prediction. *Briefings Bioinforma.* 24, bbac548. doi:10.1093/bib/bbac548

Zhou, Z., Zhuo, L., Fu, X., Lv, J., Zou, Q., and Qi, R. (2024a). Joint masking and self-supervised strategies for inferring small molecule-mirna associations. *Mol. Therapy-Nucleic Acids* 35, 102103. doi:10.1016/j.omtn.2023.102103

Zhou, Z., Zhuo, L., Fu, X., and Zou, Q. (2024b). Joint deep autoencoder and subgraph augmentation for inferring microbial responses to drugs. *Briefings Bioinforma.* 25, bbad483. doi:10.1093/bib/bbad483

Zhu, R., Wang, Y., Liu, J.-X., and Dai, L.-Y. (2021). Ipcarf: improving lncrna-disease association prediction using incremental principal component analysis feature selection and a random forest classifier. *BMC Bioinforma.* 22, 175–217. doi:10.1186/s12859-021-04104-9

Zou, Q., Li, J., Song, L., Zeng, X., and Wang, G. (2016). Similarity computation strategies in the microrna-disease network: a survey. *Briefings Funct. genomics* 15, 55–64. doi:10.1093/bfgp/elv024

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., and Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Front. Genet.* 9, 515. doi:10.3389/fgene.2018.00515