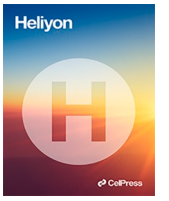
Heliyon 9 (2023) e17726

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



A lncRNA-disease association prediction model based on the two-step PU learning and fully connected neural networks

Hou Biyu, Tan GuangWen, Zeng Ming, Guan Lixin, Li Mengshan *

College of Physics and Electronic Information, Gannan Normal University, Ganzhou, Jiangxi, 341000, China

A R T I C L E I N F O

Keywords:
PU learning
Two-step approach
lncRNA-disease association
Deep learning

A B S T R A C T

Long non-coding RNAs (lncRNAs) have been shown to play a regulatory role in various processes of human diseases. However, lncRNA experiments are inefficient, time-consuming and highly subjective, so that the number of experimentally verified associations between lncRNA and diseases is limited. In the era of big data, numerous machine learning methods have been proposed to predict the potential association between lncRNA and diseases, but the characteristics of the associated data were seldom explored. In these methods, negative samples are randomly selected for model training and the model is prone to learn the potential positive association error, thus affecting the prediction accuracy. In this paper, we proposed a cyclic optimization model of predicting lncRNA-disease associations (COPTLDA in short). In COPTLDA, the two-step training strategy is adopted to search for the samples with the greater probability of being negative examples from unlabeled samples and the determined samples are treated as negative samples, which are combined together with known positive samples to train the model. The searching and training steps are repeated until the best model is obtained as the final prediction model. In order to evaluate the performance of the model, 30% of the known positive samples are used to calculate the model accuracy and 10% of positive samples are used to calculate the recall rate of the model. The sampling strategy used in this paper can improve the accuracy and the AUC value reaches 0.9348. The results of case studies showed that the model could predict the potential associations between lncRNA and malignant tumors such as colorectal cancer, gastric cancer, and breast cancer. The predicted top 20 associated lncRNAs included 10 colorectal cancer lncRNAs, 2 gastric cancer lncRNAs, and 8 breast cancer lncRNAs.

1. Introduction

The high-throughput sequencing technology allows researchers to glimpse the full picture of species' genes. The human genome contains about 3.16 billion DNA base pairs, but the number of exons only accounts for 1–2% of the total length of the genome and the remaining 98% cannot be encoded as protein sequences. According to the size, non-coding RNAs are divided into large non-coding RNAs and small non-coding RNAs [1,2]. Long non-coding RNA (lncRNA) is a large non-coding RNA with a length of more than 200 nucleotides and the largest subspecies of non-coding RNA with the important genetic role, but it was initially thought to be only a short copy of DNA [3,4]. More evidences have shown that lncRNA can change mRNA splicing through the interaction with splicing factors, regulate the epigenetic state through chromosome remodeling proteins, promote or block transcription factors to affect gene

* Corresponding author. Gannan Normal University, China.
E-mail address: msli@gnnu.edu.cn (L. Mengshan).

https://doi.org/10.1016/j.heliyon.2023.e17726
Received 9 March 2023; Received in revised form 13 June 2023; Accepted 26 June 2023
Available online 28 June 2023

expression, and participate in multiple biological processes such as cell differentiation, proliferation, and apoptosis [5–8]. LncRNA has a close regulatory association with the occurrence and development of human diseases. For example, lncRNA H19 has been shown to increase the HMGA2-mediated epithelial-mesenchymal transformation (EMT) in pancreatic ductal adenocarcinoma (PDAC) by antagonizing let-7, at least partially promoting PDAC cells [9]. In the future, lncRNA is expected to become a new biomarker to study the pathogenesis of diseases and guide disease treatment [10–14]. Therefore, it is of great significance to study the associations between lncRNA and diseases.

Traditional biological experiments are time-consuming and costly and it is difficult to make achievements from numerous data, so the number of experimentally verified lncRNAs associated with diseases is limited. Therefore, it is necessary to predict the potential associations between lncRNAs and diseases with computational methods. Currently, the calculation methods of predicting the potential associations between lncRNAs and diseases can be roughly divided into the following three categories. Firstly, based on traditional calculation methods, Chen et al. constructed a semi-supervised learning framework named LRLSLDA to predict potential disease-related lncRNAs [15]. Lu et al. used the induction matrix completion model SIMCLDA to predict the associations between lncRNA and diseases [16]. Secondly, based on deep learning methods, Zeng et al. used a neural network model based on deep matrix factorization to predict potential lncRNA-disease associations [17]. Thirdly, with the calculation results from other biological information, the correlations between lncRNA and diseases were predicted. Chen et al. integrated the miRNA-disease correlation with lncRNA-miRNA interaction to predict the potential associations between lncRNA and diseases [18].

In addition, other interaction prediction models in computational biology can also provide valuable references for lncRNA-disease association prediction. Such as, Wang et al. propose a method based on graph convolutional neural (GCN) network and conditional random field (CRF) for predicting human lncRNA-miRNA associations, named GCNCRF [19]. The model uses the GCN network to obtain the initial embedding of nodes. At the same time, an attention mechanism is added to the CRF layer to re-weight nodes, so as to better grasp the feature information of important nodes. Li et al. developed a network distance analysis model (NDALMA) for lncRNA-miRNA association prediction [13]. Firstly, the similarity networks of lncRNA and miRNA were calculated, and the Gaussian interaction spectrum (GIP) nuclear similarity was used to integrate the model. Then the network distance analysis is carried out on the integrated similar network, and the final score is obtained through the confidence calculation and score conversion. Sun et al. proposed a new deep learning algorithm called Graph convolutional Networks and Graph Attention Networks (GCNAT) to predict potential associations of disease-related metabolites [14]. First, they constructed a heterogeneous network based on the known information, and then encoded and learned the metabolites and disease characteristics through the graph convolutional neural network. Then, the graph attention layer is used to combine the embedding of multiple convolutional layers, the corresponding attention coefficient is calculated, and different weights are assigned to the embedding of each layer. Finally, the predicted results are obtained by decoding and scoring the final composite embeddings.

In the studies on the association between lncRNA and disease, some researchers used the association matrix to describe whether there was any association between the research objects, and extracted the rows and columns of the matrix from the association matrix between lncRNA and diseases as feature vectors. However, the association matrix is a sparse matrix consisting of a few positive samples and a large number of unlabeled samples containing potential positive samples. In the training process of most models, the samples from unlabeled samples are randomly selected as negative samples and used together with existing positive samples. In this way, rough labeling of unlabeled samples with negative labels leads to biased classifiers, which inevitably affect the prediction accuracy of the model. Whether the negative samples used in model training are accurate and reasonable is crucial to the prediction accuracy of the model.

In order to better predict the associations between lncRNA and diseases, the two-step method of positive and unlabeled (PU) learning is used as the learning strategy of the model. A novel method called cyclic optimization model (COPTLDA) is proposed to predict potential associations between lncRNAs and diseases. The two-step method is a strategy of PU learning. According to the strategy, the samples with greater probability of being negative examples are obtained from the U set (unlabeled samples) and then used together with the known positive samples from the P set, to train the model. After repeated searching and training, the best model is chosen to predict the remaining unlabeled samples.

2. Theory and calculation

2.1. Establishment of the correlation matrix

Given m lncRNAs $L = \{L_1, L_2, \ldots, L_i, \ldots, \}$ and n diseases $D = \{D_1, D_2, \ldots, D_j, \ldots, D_n\}$, then the lncRNA-disease interaction matrix is defined as R ($R \in R_{m\times n}$), as shown in formula (1):

$$R_{ij} = \begin{cases} 1, \textit{LncRNAs have been linked to diseases} \\ 0, \textit{The relationship between lncRNA and diseases is unknown} \end{cases} \quad (1)$$

In the training process of the model, the rows and columns of the association matrix are respectively taken as the input of the model. It is a sparse matrix, in which the value of 1 accounts for 1.166%.

2.2. PU learning

PU learning (positive and unlabeled learning) is a semi-supervised binary classification method, in which P and U respectively

represents positive samples and unlabeled samples. In PU learning, the training dataset D consists of a set of positive of D_P and a set of unlabeled D_U, where $D = D_P \cup D_U$. D_P contains n_p positive samples x^p sampled from $P(x|Y = 1)$. D_U contains n_u unlabeled samples x^u sampled from P(x). The prior probabilities of positive and negative classes are respectively expressed as $\Pi_p = P(Y = 1)$ and $\Pi_n = P(Y = -1)$, where it is assumed that Π_p is known in the paper. The parameters are set as follows. G: $R^d \rightarrow R$ indicates the binary classifier; θ is the parameter of the binary classifier; L: $R \times \{1, -1\} \rightarrow R$ indicates the loss function. Then, the risk of classifier, $\widehat{R}_{Pu}(g)$, can be approximated to formula (2) as follows

$$\widehat{R}_{Pu}(g) = \frac{\pi_p}{n_p}\sum_{i=1}^{n_p} L(g(x_i^p), 1) + \frac{1}{n_u}\sum_{i=1}^{n_u} L(g(x_i^p), -1) - \frac{\pi_p}{n_p}\sum_{i=1}^{n_p} L(g(x_i^p), -1) \tag{2}$$

This approximation process is called unbiased risk estimation. However, this estimation yields negative values, which may lead to overfitting of the model. Therefore, non-negative PU unbiased estimation, nnPU, can be replaced with formula (3) as shown below:

$$\widehat{R}_{Pu}(g) = \frac{\pi_p}{n_p}\sum_{i=1}^{n_p} L(g(x_i^p), 1) + \max\left(0, \frac{1}{n_u}\sum_{i=1}^{n_u} L(g(x_i^p), -1) - \frac{\pi_p}{n_p}\sum_{i=1}^{n_p} L(g(x_i^p), -1)\right) \tag{3}$$

PU learning provides benefits for machine learning problems where there are only positive samples and the remaining samples are unreliable for binary classification. Since deep learning was proposed, it has been widely applied in computer vision, natural language processing, bioinformatics, and other fields [20,21]. Some researchers applied deep learning in the association prediction of lncRNA and diseases [22–25]. Inspired by the previous studies, we combined the PU learning strategy with deep learning algorithms to construct a neural network model based on two-step method in PU learning, COPTLDA.

2.3. COPTLDA model

In the basic model of COPTLDA, a fully connected neural network is used to train the data. The row vectors and column vectors of the association matrix are respectively used as the inputs of a three-layer neural network and a four-layer neural network. For the intersection of the row and column, the value is masked and replaced with 0. The basic model diagram is shown in Fig. 1.

The numbers of the nodes in the four-layer network of training row vectors are respectively 581, 290, 150, and 70. The numbers of the nodes of the three-layer network for training column vectors are respectively 215, 120, and 70. The ReLU function is set as the activation function for each middle layer, as shown in formula (4):

$$\mathrm{ReLU} = \max(0, x) \tag{4}$$

During model training, two neural networks are used to train the rows and columns of the correlation matrix respectively, and their output is two vectors with the same dimension. Then combine two different features by element-wise multiplication and expect to find something more representational. So, through element-wise multiplication, the vectors of a lncRNA and a disease are fused into a new vector, as shown in formula (5):

$$\mathrm{w} = \mathrm{Element} - \mathrm{wise\ multiplication}(V_1, V_2) \tag{5}$$

The obtained vector is then put into a two-layer neural network as an input vector. The number of the input nodes of this network is 70 and the number of the output node is 1. In the final output layer, sigmoid is set as the activation function, as shown in formula (6):

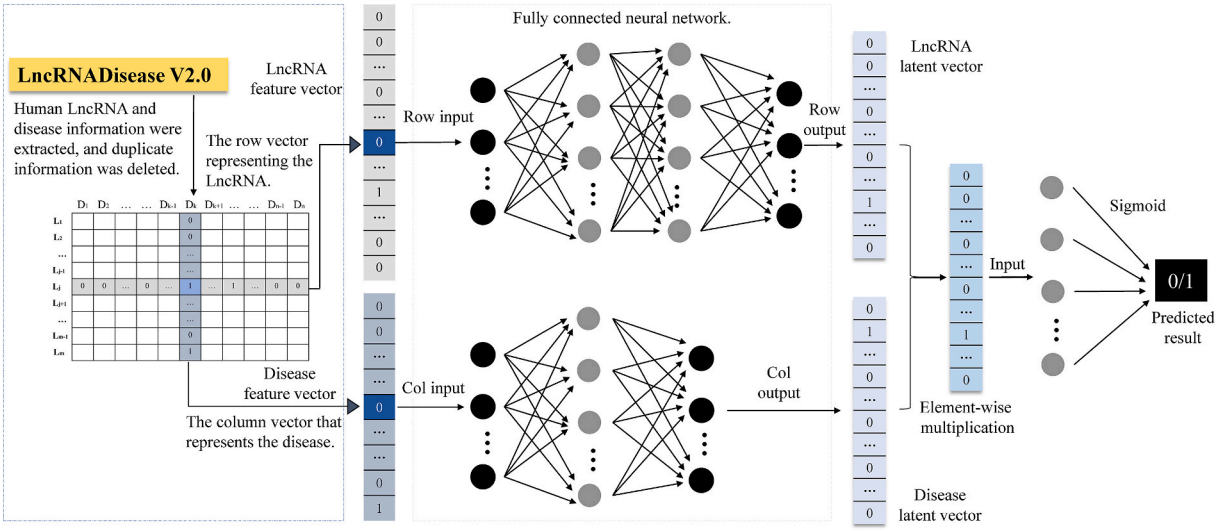


Fig. 1. Diagram of the basic model of COPTLDA.

$$\text{sigmoid}(x) = \frac{1}{1 + \exp(-x)} \tag{6}$$

The cross-entropy function is used as the loss function and Adam is used as the optimizer to train COPTLDA. The learning rate is set to 0.005 and the number of iterations is set to 100. The model prediction result is the prediction value of the association between lncRNAs and diseases. In this way, the first sub-model is obtained.

Fig. 2 shows the training strategy of COPTLDA model. The gray frame represents the process of obtaining a sub-model through one sampling and the pink frame represents the process of obtaining the final model through continuously selecting unlabeled samples that are more likely to be negative samples and training the sub-models.

In COPTLDA, each lncRNA is represented as the row L_i of the association matrix and each disease is represented as the column D_j of the association matrix, so that the association matrix R is obtained. Each lncRNA has a one-to-one association with all kinds of diseases. When the *i*-th lncRNA is associated with the *j*-th disease, Rij = 1 is a positive sample; otherwise, it is an unknown sample. In the

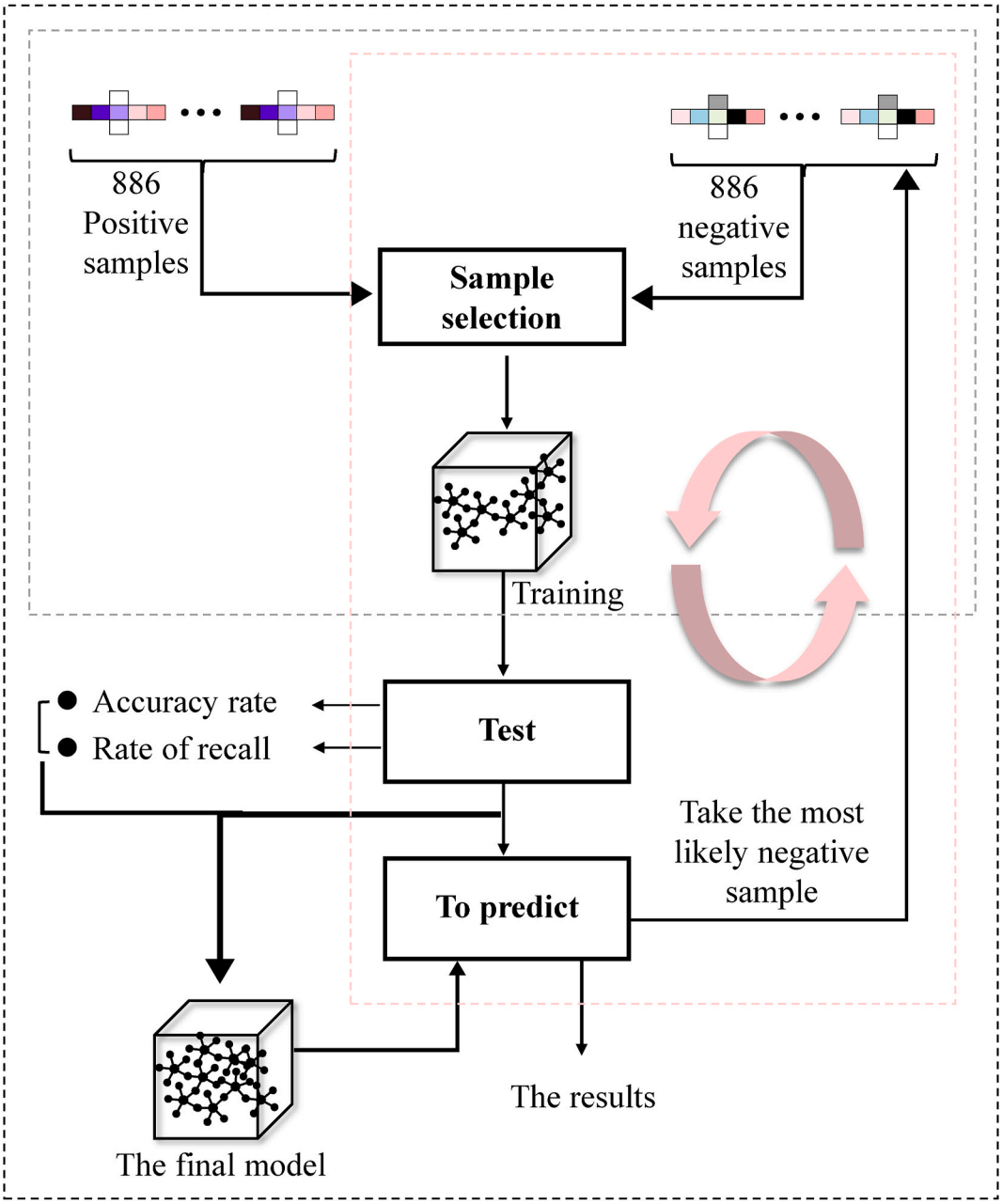


Fig. 2. Training strategy diagram.

association matrix, 1477 positive samples and 123,438 unlabeled samples are obtained. Then, 60% of the positive samples, 886 cases, are taken as the training set and 10% of positive samples, 147 cases, are doped into the unknown sample set after changing their labels. Finally, 30% of the samples, 444 cases, are taken as the test set.

The quantity of the data is small and negative samples are randomly selected, so the obtained model is rather rough. However, it is still considered to have a certain predictive ability. At this time, the untrained unknown samples are predicted with the first model and then the most possible negative samples, the highest ranked top 886 samples, are selected as the negative samples of the second model. Then, the 886 negative samples are combined with the 886 positive samples to train the second model. Then, the negative samples in the previous round of training are added into the unknown sample pool to repeat the training process of the second model. In the process, the most possible negative samples are continuously selected and combined with the fixed positive samples to train the models. Each obtained model is verified with the test set to evaluate its predictive ability and then the remaining unlabeled samples are predicted to obtain the predicted value. The negative samples selected in all the training models are statistically analyzed, and the highest ranked top 886 samples are selected as negative samples for training according to the descending order of selection times. In this way, the final model was obtained.

3. Experiments and result analysis

3.1. Data sources and model evaluation

Two datasets were downloaded from lncRNADisease Database V2.0 [26]. One dataset contained the experimentally verified associations between lncRNAs and diseases. After removing the information of other species and repeated entries, 581 lncRNAs and 215 diseases were obtained as the training model. The statistics are shown in Table 1.

From the lncRNA-disease association matrix, 1477 positive samples were obtained. In the obtained samples, 60% of positive samples were used as the training set and 30% of positive samples were used as the test set to determine the accuracy of the model. The remaining 10% of positive samples were used as the test set to determine the recall rate of the model.

The other dataset contained the association information between lncRNA and diseases predicted by other models and was used as the verification cases for case study. The two datasets were also downloaded from the RNADisease Database V4.0 [27]. One dataset was verified by experiments and the other was predicted by other models (Table 2). The two datasets were also used to evaluate the case study results.

In the plotting process of receiver operating characteristic (ROC) curve, no negative sample was obtained, so the samples with the largest probability of being negative samples (top unlabeled samples) were selected as negative samples. The proportions of positive and negative samples remained the same as the training dataset. The numbers of positive and negative samples were respectively 444 and 38,079. The data were input into the trained model to calculate the true positive rate (TPR) and false positive rate (FPR) with the predicted new values, as shown in formula (7) and formula (8):

$$TPR = \frac{\mathrm{TP}}{TP + FN} \tag{7}$$

$$FPR = \frac{\mathrm{FP}}{FP + TN} \tag{8}$$

where *TP* represents the number of correctly identified positive samples and *FN* represents the number of incorrectly identified positive samples. According to the settings of different classification thresholds, FPR and TPR were used as horizontal and vertical coordinates to plot receiver operating characteristic (ROC) curves. FPR and TPR were also used as the performance evaluation indexes of the model. The area under ROC curve was defined as AUC. It is generally believed that the larger value of AUC indicates the better performance of classifier.

A total of 444 cases (30% of the positive samples) were taken as the test set and the ratio of the number of positive samples correctly predicted by the test to the number of positive samples in the test was defined as accuracy. The accuracy calculation formula is shown in formula (9):

$$ACC = \frac{\mathrm{TP}}{TP + FN} \tag{9}$$

A total of 10% positive samples (147 cases) were doped into the unknown sample dataset for model training. After training, the prediction accuracy of these positive samples was used as one of the evaluation indexes of the model. The recall rate calculation formula is shown in formula (10):

Table 1
Training set of the model.

Dataset	LncRNAs	Diseases	Association	Association rate
LncRNADisease V2.0	581	215	1477	1.166%

Table 2
Test set of the model.

Datasets	File names	URL
LncRNADisease v2.0	Predicted lncRNA-disease information.xlsx	http://www.rnanut/lncrnadisease/
RNADisease V4.0	Experimental Data: lncRNA-disease information	http://rnadisease.org/
RNADisease V4.0	Predicted Data: lncRNA-disease information	http://rnadisease.org/

$$recall = \frac{TP}{TP + FN} \tag{10}$$

In the calculation of accuracy and recall rate, *TP* represents the number of correctly identified positive samples and *FN* represents the number of incorrectly identified positive samples.

3.2. Distribution analysis of the results

Fig. 3(a) shows the result distribution of the final model. After each sub-model was trained, the untrained unlabeled samples were predicted. When the prediction result was less than 0.001, corresponding sample was more probably a negative sample. When it was greater than 0.999, it was a positive sample. The distribution of prediction results of 60 sub-models was recorded. The negative samples for training the first sub-model were mixed with 147 known positive samples, thus resulting in the weak prediction ability of the model. The proportion of predicted samples in the unknown samples was small and only 5% of the samples had a predicted value greater than 0.999. Starting from the second sub-model, according to two-step strategy, the samples were trained and the number of samples determined as positive samples increased. The calculated average values indicated that the proportion of the most probably positive samples reached 25%, whereas the proportion of the most probably negative samples was 55%. The prediction results of most sub-models fluctuated within a certain range (Fig. 3(b) and (c)). This fluctuation might be interpreted as follows. The samples selected for each sub-model did not include the negative samples selected for the former sub-model, thus reducing the number of the most probably negative samples. Therefore, in the final prediction result, the proportion of positive samples decreased and the corresponding proportion of negative samples increased. In the prediction with the final model, the extreme thresholds of 0.001 and 0.999 were set for classification and the results were relatively stable, indicating that COPTLDA had the stable prediction ability.

3.3. Loss analysis of the prediction results

Table 3 lists the loss values of the two models respectively after training for 60 times and 80 times. The comparison results of the

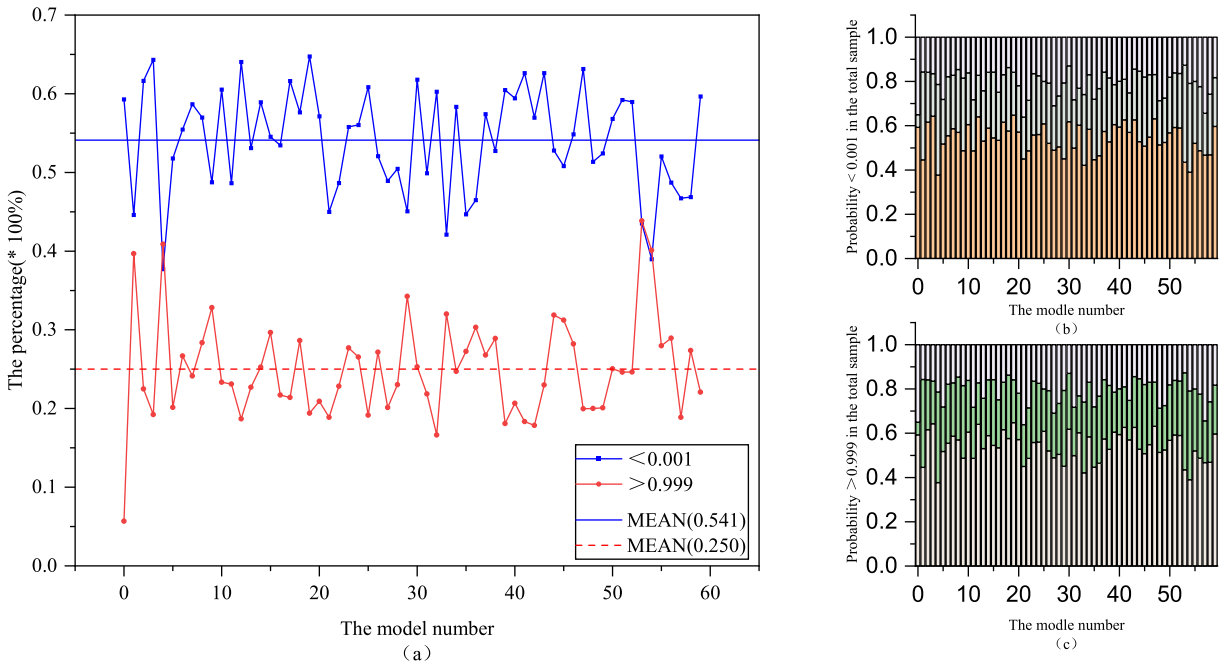


Fig. 3. Distribution of prediction results. (a) The distribution of the proportion of unknown samples predicted by the final model as positive and negative samples and their average proportion. (b) Distribution diagram of the proportion of unknown samples predicted by the final model as negative samples. (c) Distribution diagram of the proportion of unknown samples predicted by the final model to be positive.

loss values of the COPTLDA model and the single model indicated that after the trained model converged, the losses of sub-models in the COPTLDA model were similar, whereas the losses of the sub-models in the single model were quite different.

The loss difference of COPTLDA model between M50 and M40 (or M60) after training for 60 times was the largest and reached 0.017. The loss difference of single model between M30 and M60 after training for 60 times was the largest and reached 0.073. This gap might be interpreted as follows. Negative samples of the single model were randomly selected and more positive samples were misclassified, further indicating that the training strategy of the COPTLDA model was more reliable. Fig. 4 shows the loss curves of COPTLDA model and single model.

Fig. 4(a) and (c) respectively show the loss functions of the seven sub-models in the two models trained for 100 times. Each line represents the loss function of a sub-model. Fig. 4(b) and (d) respectively show the violin plots of 7 sub-models and each violin represents the loss distribution of a sub-model. Sub-model 1 (M1) of COPTLDA contained incorrectly identified samples, thus resulting in the slow training convergence speed. After training for 40 times, the loss of Sub-model 1 was 4–6 times of that of other sub-models and its loss was still larger than that of other models when the model converged after training for 100 times. After the two models were training for 80 times, the losses of several sub-models fluctuated sharply. This fluctuation might be interpreted as follows. The set learning rate was slightly higher and the softmax activation function adopted in the output layer might cause overflow. In addition, the weight in the neural network changed to NAN, which resulted in a steep increase in loss. COPTLDA model has a smaller loss range than the single model, and the loss value of the 60th sub-model was the smallest (Fig. 4(b) and (d)). The loss function did not fluctuate in the training process, indicating that COPTLDA model had a more stable performance in predicting the association between lncRNA and diseases.

3.4. Accuracy of the model

Fig. 5 shows the probability that 444 positive samples of the test set are correctly predicted in each sub-model.

Fig. 5(a) shows the variation of accuracy of the sub-models of the two models. Red lines and blue lines respectively represent the accuracy of COPTLDA model and single model and the purple dotted lines represent the accuracy of the first 886 samples that are repeatedly selected as the negative samples for the largest probability after training for 60 times. Fig. 5(b) shows the ridge maps of the accuracy of the two models. Red peaks represent the COPTLDA model and yellow peaks represent the single model. Except Sub-model 1 in COPTLDA whose accuracy was lower than that of Sub-model 1 of the single model because the training set contained multiple samples with label errors in the initial state, the accuracy of other sub-models of COPTLDA exceeded that of the corresponding sub-models of the single model. The red peak was larger and closer to the right (Fig. 5(b)), indicating that the sub-models in COPTLDA model had higher accuracy. It was also confirmed that the two-step strategy could improve the prediction ability of the model. The accuracy of the purple dotted line reached 0.876, which was basically equal to the highest accuracy of the sub-models. Therefore, the statistical selection times-based selection method of negative samples could select more reliable negative samples and further improve the accuracy of the model.

3.5. Model recall rate

Fig. 6 shows the probability that the positive samples doped in the unlabeled sample set were re-predicted as positive samples in the prediction with each sub-model of COPTLDA model (also called recall rate).

Fig. 6(a) shows the variations of recall rates of sub-models of the two models. Green lines and blue lines respectively represent the recall rates of COPTLDA model and single model. Like the accuracy calculation results, the red dotted lines in Fig. 6(a) indicate the recall rate of the first 886 samples that are repeatedly selected as the negative samples for the largest probability after training for 60 times. Fig. 6(b) shows the half-fiddle plot of recall rates of the two models. After stable prediction, the recall rates of sub-models of COPTLDA were always greater than those of sub-models of the single model. The red dotted lines showed a recall rate of 0.932, which was better than all sub-models of the single model. The results further confirmed that qualified negative samples could be selected with the two-step strategy so as to obtain the optimized model and improve the predictive ability of the association between lncRNA and diseases.

3.6. ROC curve of the final model

Through the comprehensive judgment of the accuracy and recall rate of the 60 sub-models, negative samples were selected according to the descending order of the number of selections to train the final model, which was used to predict the potential association between lncRNA and diseases. Fig. 7 shows the ROC curve of the final selected model in the COPTLDA model.

Table 3
Loss values of the two models after training for 60 and 80 times.

Model	loop	M10	M20	M30	M40	M50	M60
COPTLDA	60	0.018	0.016	0.019	0.012	0.029	0.012
	80	0.010	0.011	0.007	0.004	0.016	0.005
single	60	0.040	0.037	0.096	0.069	0.041	0.033
	80	0.018	0.014	0.076	0.034	0.021	0.013

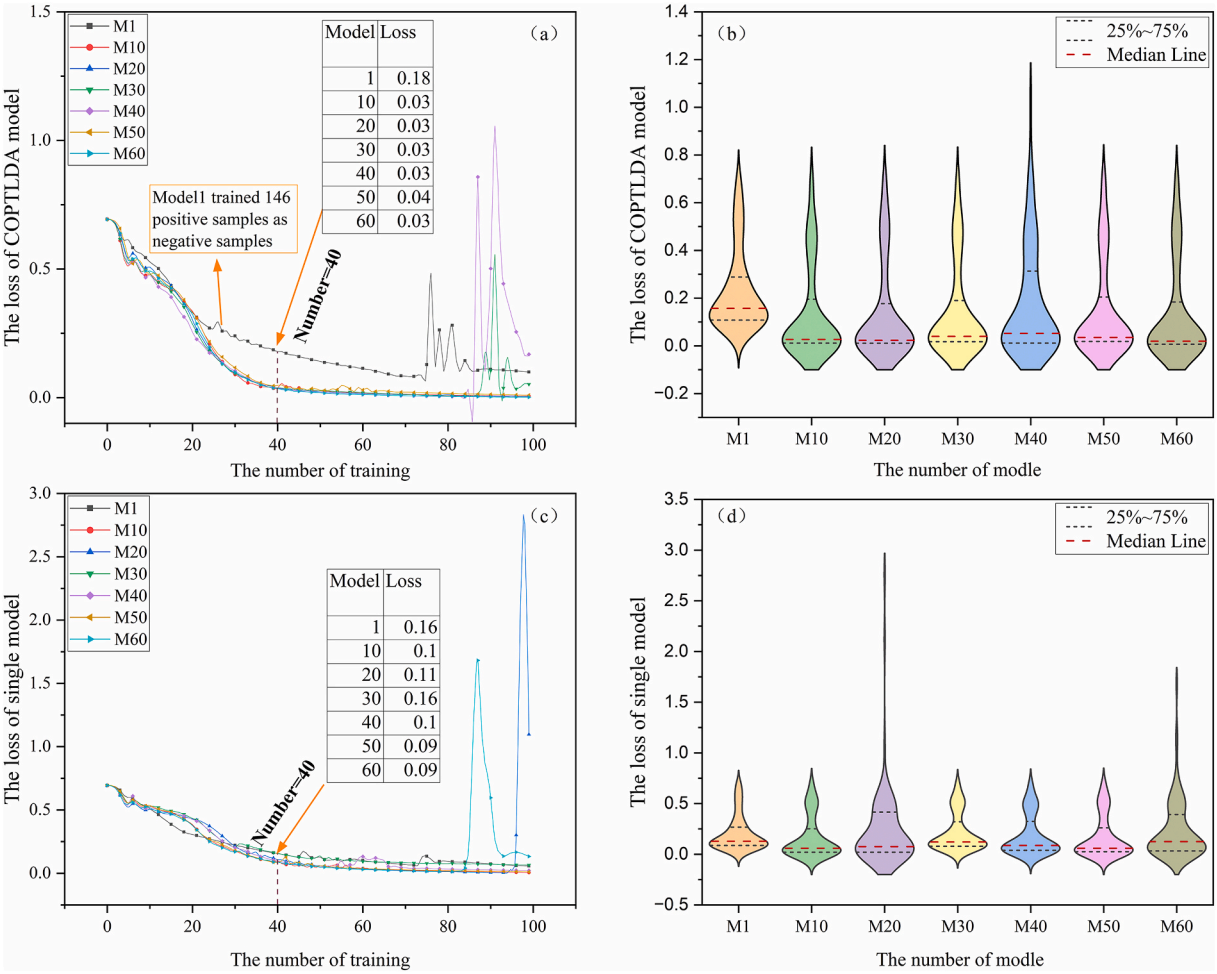


Fig. 4. Loss curves of COPTLDA model and single model. (a) Loss function of seven sub-models in COPTLDA model. (b) Loss distribution of seven sub-models in COPTLAD model. (c) Loss function of seven sub-models in single model. (d) Loss distribution of seven sub-models in single model.

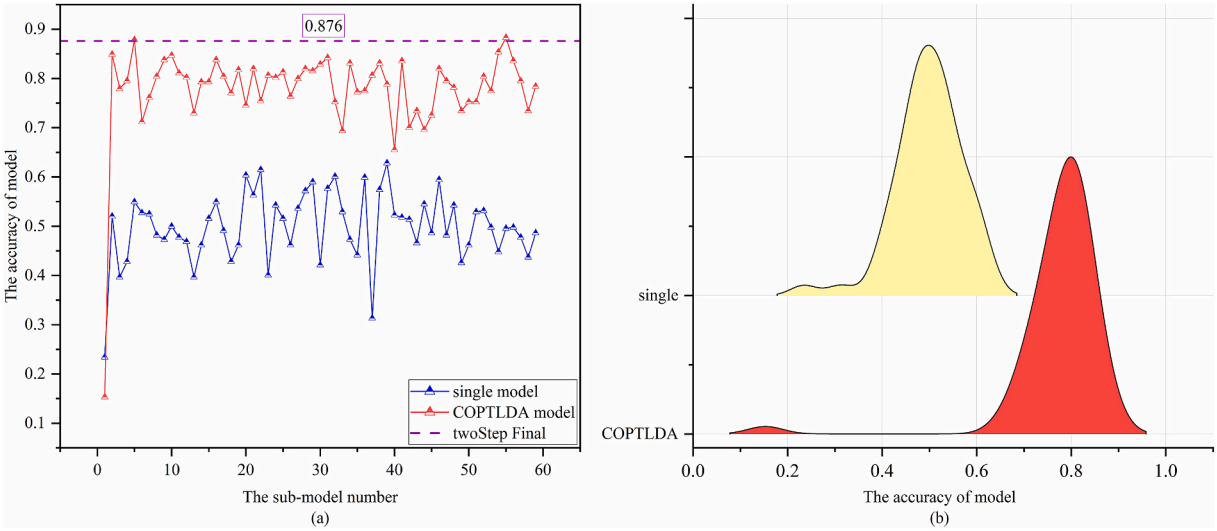


Fig. 5. Accuracy of the two models.(a) Prediction accuracy of 60 sub-models in COPTLAD model and single model and prediction accuracy of final model. (b) Accuracy distribution of COPTLAD model and single model.

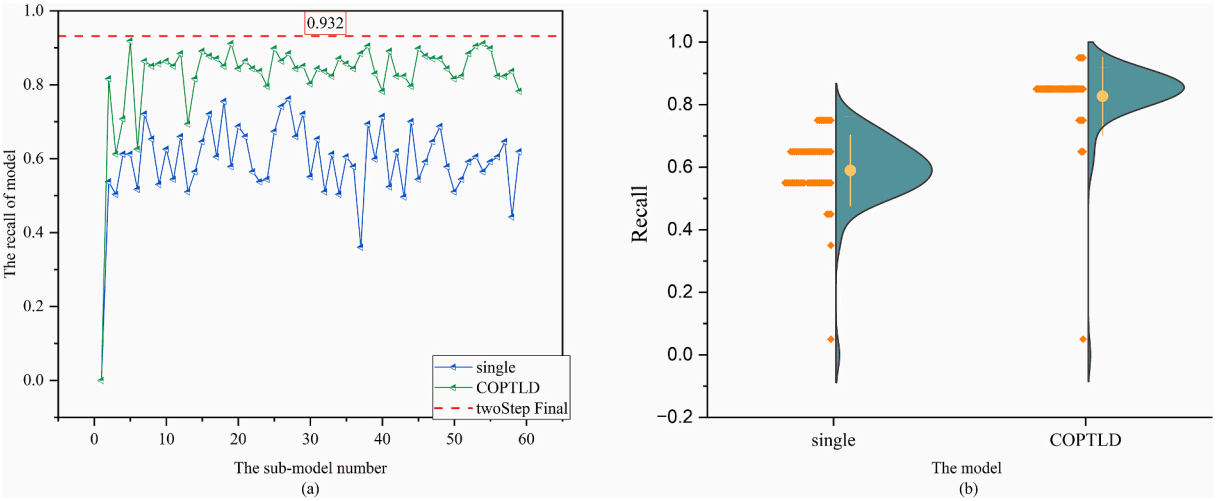


Fig. 6. Variations of recall rates of sub-models of the two models. (a) Recall rate of COPTLAD model and single model 60 sub-models and recall rate of final model. (b) Recall rate distribution of COPTLAD model and single model.

When FPR = 0, the TPR value was above 0.9. Because it's in the prediction, most of the predicted values of positive samples are greater than 0.99, while the predicted values of negative samples are basically close to 0. Therefore, when the threshold is set as 1-δ (δ is infinitesimal), the x-coordinate approaches 0 and the y-coordinate approaches 1. Then, the threshold is reduced, so that when the x-coordinate gradually increases from 0 to 1, the y-coordinate basically does not change, but when the threshold is set to 0 + δ (δ is infinitesimal), now all samples are predicted to be positive samples, so there's a sudden change at the tail of the ROC curve. But this has no effect on the performance of the model. The AUC value finally reached 0.9348. The results fully proved that COPTLDA model had the superior performance and contributed to the prediction of the association between LncRNA and diseases. The two-step sampling strategy could significantly improve the prediction ability of the model.

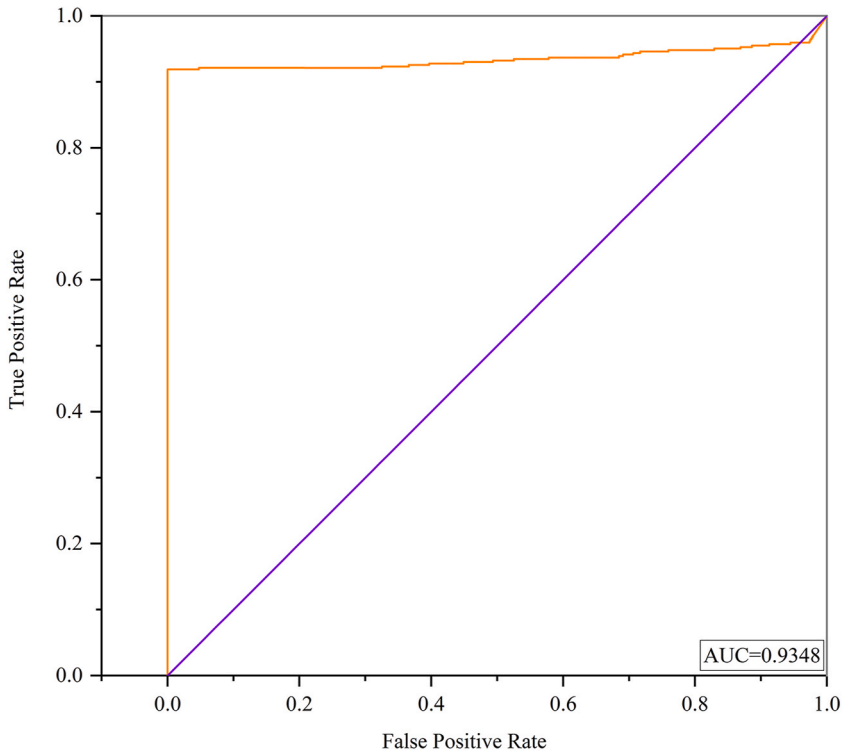


Fig. 7. ROC curve of the final model.

3.7. COPTLDA compared to other models

In order to further verify the advantages of COPTLDA model, 5 models with good prediction performances were selected as the benchmark comparison model in this paper. Table 4 provides relevant theories and parameter information of each model.

Fig. 8 shows the AUC values of the six models. The AUC values of COPTLDA, DMFLDA, SIMCLDA, TPGLDA, MFLDA, and LDAP models showed the descending order. In this paper, the AUC value of the COPTLDA model reached the maximum value of 0.9348, indicating that COPTLDA performed better than the other five models and had the higher confidence in predicting the potential association between lncRNA and diseases.

4. Case studies

The predicted results of three common cancers were output and the top 20 lncRNAs were selected and compared with the predicted results of other papers in the database or manually mined biomedical references.

COPTLDA predicted the top 20 lncRNAs associated with colorectal cancer, including 12 lncRNAs validated in the validation set (Table 5).

Colorectal cancer is a common malignant tumor whose onset age tends to be middle-aged and elderly and ranks second among cancers of males and third among cancers of females. It seriously endangers public health security [31]. Therefore, it is important to find the lncRNAs associated with colon cancer. In this paper, COPTLDA was used to predict the lncRNAs associated with colorectal cancer. Among the top 20 lncRNAs, 10 experimentally verified lncRNAs and 2 lncRNAs predicted by other models in the validation dataset were NBR2 [32], DANCR [33], Lnc00312 [34], HIF1A-AS2 [35], lnc34a [36], ADAMTS9-AS2 [37], PCBP2-OT1 [38], LINC00339 [39], UHRF1 [40], LncBRM [41], GIHCG, and LINC00968.

COPTLDA predicted the top 20 lncRNAs associated with gastric cancer, including 13 lncRNAs validated in the validation set (Table 6).

Gastric cancer is a common malignant tumor in the digestive system, mostly in middle-aged and elderly patients, most of which are male. Most of the diagnosed patients of gastric cancer have entered the middle and late stages of the cancer and gastric cancer ranks second among the most common causes for cancer death [42]. Similarly, COPTLDA was used in this paper to predict lncRNAs associated with gastric cancer. Among the top 20 lncRNAs, 2 lncRNAs found in the validation dataset and 11 predicted lncRNAs were HIF1A-AS2 [43], BLACAT1 [44], GIHCG, LINC01844, TRIM52-AS1, ADAMTS9-AS2, PCBP2-OT1, LINC00339, SBF2-AS1, TSIX, LINC01116, TCF7, and FTX.

COPTLDA predicted the top 20 lncRNAs associated with breast cancer, including 9 lncRNAs validated in the validation set (Table 7).

According to the latest data of the International Agency for Research on Cancer (IARC) in 2018, the global incidence of breast cancer in female cancers was 24.2% and breast cancer ranked first among the cases of female cancers, among which 52.9% cases occurred in developing countries [45]. In this paper, COPTLDA was used to predict lncRNAs associated with breast cancer. Among the top 20 lncRNAs, 8 lncRNAs validated in the validation dataset and 1 predicted lncRNA were GIHCG [46], HIF1A-AS2 [47], ADAMTS9-AS2 [48], LINC00339 [49], BLACAT1 [50], SNHG20 [51], SBF2-AS1 [52], TSIX [53], and SCHLAP1.

In the paper, COPTLDA model was used to predict the associations between three types of cancers and lncRNAs. In the prediction results, a total of 20 experimentally verified lncRNAs included 10 lncRNAs associated with colorectal cancer, 8 lncRNAs associated with breast cancer, and 2 lncRNAs associated with gastric cancer. In the prediction results, a total of 14 lncRNAs predicted by other researchers included 2 lncRNAs associated with colorectal cancer, 1 lncRNA associated with breast cancer, and 11 lncRNAs associated with gastric cancer. The obtained results proved that COPTLDA model had the ability to predict the potential associations between lncRNA and diseases. Due to different data sources and scarce positive samples, the real prediction ability of the model should be more accurate than the above data.

5. Conclusions

LncRNAs play an important role in various biological processes. The identification of disease-related lncRNAs is of great significance for understanding the pathogenesis of diseases at the lncRNA level and contributes to disease prevention and treatment. The COPTLDA model uses the following strategies: (1) The two-step strategy is used to select more likely negative samples. The selected negative samples and some known positive samples are used to train the first sub-model and record the evaluation indicators of the sub-model. (2) Without putting back the negative samples selected in the last round, the process of selecting negative samples and training sub-models is repeated constantly, and the evaluation index of each sub-model is recorded. (3) Count the number of negative samples that have been selected, sort by the number of selection times, and select a certain number of negative samples with higher ranking. The selected negative samples are trained together with the positive samples to obtain the final model, and then all unknown associations are predicted.

In this paper, the two-step strategy in PU learning was adopted to select the training data and train the model. Unknown positive samples mixed with unlabeled samples for training affected the predictive performance of the model, indicating that PU learning method performed better in predicting the associations between lncRNA and diseases. Then, COPTLDA model was compared with the single model. The results of multiple parameters indicated that the two-step strategy in training data could greatly improve the predictive performance of lncRNA-disease association model. Based on the statistics of the first 60 sub-models, the top 886 samples with the largest probability of being selected were obtained and the higher accuracy and recall rate were finally realized, thus further

Table 4
Introduction to comparative models.

Models	Descriptions	References
DMFLDA	Developed by Zeng et al. a neural network model based on deep matrix factorization could predict potential lncRNA-disease associations.	[17]
SIMCLDA	Developed by Lu et al. and based on inductive matrix completeness, SIMCLDA found a low-rank matrix that could integrate the prior knowledge of lncRNA and diseases to form the lncRNA-disease association matrix.	[16]
TPGLDA	Developed by Ding et al. it combined gene-disease associations with lncRNA-disease associations and used effective resource allocation algorithms to predict potential lncRNA-disease associations.	[28]
MFLDA	Developed by Fu et al. it decomposed the data matrix of heterogeneous data sources into a low-rank matrix through matrix triple decomposition in order to explore and utilize its inherent and shared structure.	[29]
LDAP	Developed by Lan et al. an integrated SVM was used to predict potential lncRNA-disease associations through fusing lncRNA similarity and disease similarity.	[30]

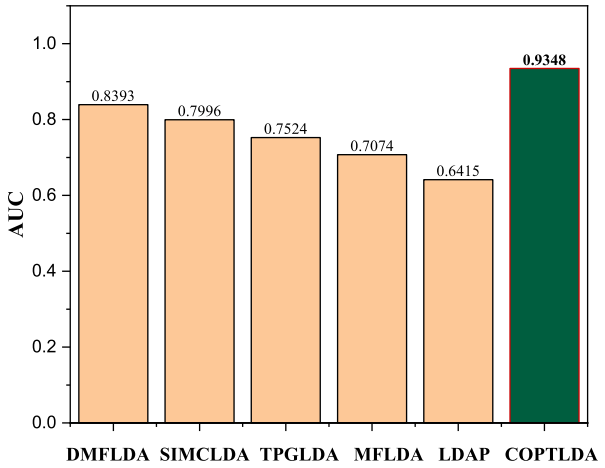


Fig. 8. AUC values of six models.

Table 5
Prediction results of lncRNAs associated with colorectal cancer.

Ranks	LncRNAs	References	PMID
1	GIHCG	LncRNADiseaseV2.0 (pre)	/
2	NBR2	RNADisease V4.0 (ex)	31571148
3	DANCR	RNADisease V4.0 (ex)	31863900
4	Lnc00312	RNADisease V4.0 (ex)	29461596
5	HIF1A-AS2	RNADisease V4.0 (ex)	29278853
6	lnc34a	RNADisease V4.0 (ex)	27077950
7	ADAMTS9-AS2	RNADisease V4.0 (ex)	27596298
8	PCBP2-OT1	RNADisease V4.0 (ex)	27914101
9	LINC00339	RNADisease V4.0 (ex)	31269584
10	UHRF1	RNADisease V4.0 (ex)	26768845
11	LINC00968	LncRNADiseaseV2.0 (pre)	/
12	LncBRM	RNADisease V4.0 (ex)	30563768

confirming the effectiveness and superiority of the two-step sampling strategy.

In COPTLDA, only the association between lncRNAs and diseases is used to construct the association matrix without using other biological data. However, other kinds of biological data are also significant for the prediction of association between lncRNAs and diseases. In future studies, it is necessary to integrate various biological data for joint prediction and develop more effective algorithms to adapt to complex data relationships.

Author contribution statement

Li Mengshan: conceived and designed the experiments; wrote the paper.
Hou Biyu: conceived and designed the experiments; performed the experiments; wrote the paper.
Zeng Ming: performed the experiments; analyzed and interpreted the data; wrote the paper.

Table 6
Prediction results of lncRNAs associated with gastric cancer.

Ranks	LncRNA	References	PMID
1	GIHCG	RNADisease V4.0 (pre)	/
2	LINC01844	LncRNADiseaseV2.0 (pre)	/
3	TRIM52-AS1	LncRNADiseaseV2.0 (pre)	/
4	HIF1A-AS2	RNADisease V4.0 (ex)	25686739
5	ADAMTS9-AS2	LncRNADisease V2.0 (pre)	/
6	PCBP2-OT1	LncRNADisease V2.0 (pre)	/
7	LINC00339	LncRNADisease V2.0 (pre)	/
8	BLACAT1	RNADisease V4.0 (ex)	25755750
9	SBF2-AS1	LncRNADisease V2.0 (pre)	/
10	TSIX	LncRNADisease V2.0 (pre)	/
11	LINC01116	RNADisease V4.0 (pre)	/
12	TCF7	LncRNADisease V2.0 (pre)	/
13	FTX	RNADisease V4.0 (pre)	/

Table 7
Prediction results of lncRNAs associated with breast cancer.

Ranks	LncRNA	References	PMID
1	GIHCG	RNADisease V4.0 (ex)	31858553
2	HIF1A-AS2	RNADisease V4.0 (ex)	14580258
3	ADAMTS9-AS2	RNADisease V4.0 (ex)	30840279
4	LINC00339	RNADisease V4.0 (ex)	29453409
5	BLACAT1	RNADisease V4.0 (ex)	30733855
6	SNHG20	RNADisease V4.0 (ex)	29236315
7	SCHLAP1	LncRNADiseaseV2.0 (pre)	/
8	SBF2-AS1	RNADisease V4.0 (ex)	31952549
9	TSIX	RNADisease V4.0 (ex)	31998636

Tan Guangwen: analyzed and interpreted the data; wrote the paper.
Guan Lixin: contributed reagents, materials, analysis tools or data; wrote the paper.

Data availability statement

Data associated with this study has been deposited at https://github.com/HouBiyu/LncRNA-disease-association-prediction.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Y.A. Huang, et al., Predicting microRNA-disease associations from lncRNA-microRNA interactions via Multiview Multitask Learning, Briefings Bioinf. 22 (3) (2021) bbaa133.
[2] H. Hu, et al., Gene function and cell surface protein association analysis based on single-cell multiomics data, Comput. Biol. Med. 157 (2023), 106733.
[3] H.X. Dang, et al., Long non-coding RNA LCAL62/LINC00261 is associated with lung adenocarcinoma prognosis, Heliyon 6 (3) (2020), e03521.
[4] Y. Xu, R. Liu, Analysis of the role of m6A and lncRNAs in prognosis and immunotherapy of hepatocellular carcinoma, Heliyon 8 (9) (2022), e10612.
[5] A. Silva, E.J. Spinosa, Graph convolutional auto-encoders for predicting novel lncRNA-disease associations, IEEE ACM Trans. Comput. Biol. Bioinf 19 (4) (2022) 2264–2271.
[6] G.B. Xie, J.W. Jiang, Y.P. Sun, LDA-LNSUBRW: lncRNA-disease association prediction based on linear neighborhood similarity and unbalanced bi-random walk, IEEE ACM Trans. Comput. Biol. Bioinf 19 (2) (2022) 989–997.
[7] J. Li, et al., Integrative network analysis reveals subtype-specific long non-coding RNA regulatory mechanisms in head and neck squamous cell carcinoma, Comput. Struct. Biotechnol. J. 21 (2023) 535–549.
[8] K. Zheng, et al., iMDA-BN: identification of miRNA-disease associations based on the biological network and graph embedding algorithm, Comput. Struct. Biotechnol. J. 18 (2020) 2391–2400.
[9] C. Ma, et al., H19 promotes pancreatic cancer metastasis by derepressing let-7's suppression on its target HMGA2-mediated EMT, Tumour Biol. 35 (9) (2014) 9163–9169.
[10] W. Han, et al., Pan-cancer analysis of LncRNA XIST and its potential mechanisms in human cancers, Heliyon 8 (10) (2022), e10786.
[11] Q. Lin, et al., Risk score = LncRNAs associated with doxorubicin metabolism can be used as molecular markers for immune microenvironment and immunotherapy in non-small cell lung cancer, Heliyon 9 (3) (2023), e13811.
[12] T.-T. Sun, et al., Regulatory effect of long-stranded non-coding RNA-CRNDE on neurodegeneration during retinal ischemia-reperfusion, Heliyon 8 (10) (2022), e10994.
[13] L. Zhang, et al., Using network distance analysis to predict lncRNA-miRNA interactions, Interdiscip. Sci. 13 (3) (2021) 535–545.
[14] F. Sun, J. Sun, Q. Zhao, A deep learning method for predicting metabolite-disease associations via graph neural network, Briefings Bioinf. 23 (4) (2022) bbac266.

[15] C. Xing, Y. Gui-Ying, Novel human lncRNA-disease association inference based on lncRNA expression profiles, Bioinformatics 29 (20) (2013) 2617–2624.
[16] L. Chengqian, et al., Prediction of lncRNA-disease associations based on inductive matrix completion, Bioinformatics 34 (19) (2018) 3357–3364.
[17] M. Zeng, et al., DMFLDA: a deep learning framework for predicting lncRNA-disease associations, IEEE ACM Trans. Comput. Biol. Bioinf 18 (6) (2021) 2353–2363.
[18] Chen, Xing, Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA, Sci. Rep. 5 (2015), 13186.
[19] W. Wang, et al., Predicting the potential human lncRNA-miRNA interactions based on graph convolution network with conditional random field, Briefings Bioinf. 23 (6) (2022) bbac463.
[20] Y. Zhang, F. Ye, X.P. Gao, MCA-net: multi-feature coding and attention convolutional neural network for predicting lncRNA-disease association, IEEE ACM Trans. Comput. Biol. Bioinf 19 (5) (2022) 2907–2919.
[21] T. Wang, J. Sun, Q. Zhao, Investigating cardiotoxicity related with hERG channel blockers using molecular fingerprints and graph attention mechanism, Comput. Biol. Med. 153 (2023), 106464.
[22] Y. Wang, et al., LncRNA functional annotation with improved false discovery rate achieved by disease associations, Comput. Struct. Biotechnol. J. 20 (2022) 322–332.
[23] L. Guo, et al., A comprehensive analysis of ncRNA-mediated interactions reveals potential prognostic biomarkers in prostate adenocarcinoma, Comput. Struct. Biotechnol. J. 20 (2022) 3839–3850.
[24] V.B. de Souza, J.C. Nobre, K. Becker, DAC stacking: a deep learning ensemble to classify anxiety, depression, and their comorbidity from reddit texts, IEEE J. Biomed. Health Info. 26 (7) (2022) 3303–3311.
[25] M.M. Gao, et al., Multi-label fusion collaborative matrix factorization for predicting LncRNA-disease associations, IEEE J. Biomed. Health Info. 25 (3) (2021) 881–890.
[26] Z. Bao, et al., LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases, Nucleic Acids Res. 47 (D1) (2019) D1034–d1037.
[27] J. Chen, et al., RNADisease v4.0: an updated resource of RNA-associated diseases, providing RNA-disease analysis, enrichment and prediction, Nucleic Acids Res. 51 (D1) (2023) D1397–d1404.
[28] L. Ding, et al., TPGLDA: novel prediction of associations between lncRNAs and diseases via lncRNA-disease-gene tripartite graph, Sci. Rep. 8 (1) (2018) 1065.
[29] G. Fu, et al., Matrix factorization-based data fusion for the prediction of lncRNA-disease associations, Bioinformatics 34 (9) (2018) 1529–1537.
[30] W. Lan, et al., LDAP: a web server for lncRNA-disease association prediction, Bioinformatics 33 (3) (2017) 458–460.
[31] W. Tavanapong, et al., Artificial intelligence for colonoscopy: past, present, and future, IEEE J. Biomed. Health Info. 26 (8) (2022) 3950–3965.
[32] J. Bai, et al., LncRNA NBR2 suppresses migration and invasion of colorectal cancer cells by downregulating miRNA-21, Hum. Cell 33 (1) (2020) 98–103.
[33] J. Lian, et al., Long non-coding RNA DANCR promotes colorectal tumor growth by binding to lysine acetyltransferase 6A, Cell. Signal. 67 (2019), 109502.
[34] L.M. Feng, et al., Association of the upregulation of LncRNA00673 with poor prognosis for colorectal cancer, Eur. Rev. Med. Pharmacol. Sci. 22 (3) (2018) 687–694.
[35] J. Lin, et al., LncRNA HIF1A-AS2 positively affects the progression and EMT formation of colorectal cancer through regulating miR-129-5p and DNMT3A, Biomed. Pharmacother. 98 (2018) 433–439.
[36] L. Wang, et al., A long non-coding RNA targets microRNA miR-34a to regulate colon cancer stem cell asymmetric division, Elife 5 (2016), e14620.
[37] Q. Li, et al., Differentially expressed long non-coding RNAs and the prognostic potential in colorectal cancer, Neoplasma 63 (6) (2016) 977–983.
[38] C. Wang, et al., TUC.338 promotes invasion and metastasis in colorectal cancer, Int. J. Cancer 140 (6) (2017) 1457–1464.
[39] H. Ye, et al., The SP1-induced long noncoding RNA, LINC00339, promotes tumorigenesis in colorectal cancer via the miR-378a-3p/MED19 Axis, OncoTargets Ther. 13 (2020) 11711–11724.
[40] K. Taniue, et al., Long noncoding RNA UPAT promotes colon tumorigenesis by inhibiting degradation of UHRF1, Proc. Natl. Acad. Sci. U.S.A. 113 (5) (2016) 1273–1278.
[41] R. Li, et al., Long noncoding RNA lncBRM promotes proliferation and invasion of colorectal cancer by sponging miR-204-3p and upregulating TPT1, Biochem. Biophys. Res. Commun. 508 (4) (2019) 1259–1263.
[42] H. Brenner, D. Rothenbacher, V. Arndt, Epidemiology of stomach cancer, Methods Mol. Biol. 472 (2009) 467–477.
[43] W.M. Chen, et al., Antisense long noncoding RNA HIF1A-AS2 is upregulated in gastric cancer and associated with poor prognosis, Dig. Dis. Sci. 60 (6) (2015) 1655–1662.
[44] Y. Hu, et al., Long noncoding RNA linc-UBC1 is negative prognostic factor and exhibits tumor pro-oncogenic activity in gastric cancer, Int. J. Clin. Exp. Pathol. 8 (1) (2015) 594–600.
[45] H. Sung, et al., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, Cancer J. Clinic. 71 (3) (2021) 209–249.
[46] L.Y. Fan, et al., LncRNA GIHCG regulates microRNA-1281 and promotes malignant progression of breast cancer, Eur. Rev. Med. Pharmacol. Sci. 23 (24) (2019) 10842–10850.
[47] A. Cayre, et al., aHIF but not HIF-1 alpha transcript is a poor prognostic marker in human breast cancer, Breast Cancer Res. Treat. 5 (6) (2003) R223–R230.
[48] Y.F. Shi, H. Lu, H.B. Wang, Downregulated lncRNA ADAMTS9-AS2 in breast cancer enhances tamoxifen resistance by activating microRNA-130a-5p, Eur. Rev. Med. Pharmacol. Sci. 23 (4) (2019) 1563–1573.
[49] X. Wang, et al., Long noncoding RNA Linc00339 promotes triple-negative breast cancer progression through miR-377-3p/HOXC6 signaling pathway, J. Cell. Physiol. 234 (8) (2019) 13303–13317.
[50] X. Hu, et al., Long non-coding RNA BLACAT1 promotes breast cancer cell proliferation and metastasis by miR-150-5p/CCR2, Cell Biosci. 9 (2019) 14.
[51] Y.X. Guan, et al., Lnc RNA SNHG20 participated in proliferation, invasion, and migration of breast cancer cells via miR-495, J. Cell. Biochem. 119 (10) (2018) 7971–7981.
[52] W. Xia, et al., Down-regulated lncRNA SBF2-AS1 inhibits tumorigenesis and progression of breast cancer by sponging microRNA-143 and repressing RRS1, J. Exp. Clin. Cancer Res. 39 (1) (2020) 18.
[53] E.A. Salama, R.E. Adbeltawab, H.M. El Tayebi, XIST and TSIX: novel cancer immune biomarkers in PD-L1-overexpressing breast cancer patients, Front. Oncol. 9 (2019) 1459.