DBNLDA: Deep Belief Network based representation learning for lncRNA-disease association prediction

Manu Madhavan¹,² · G. Gopakumar¹

Abstract
The advancements in the field of high throughput analysis show abnormal expression of long non-coding RNAs (lncRNAs) in many complex diseases. Accurately identifying the disease association of lncRNA is essential in understanding their role in disease mechanism and subsequent therapy. The contemporary methods for predicting lncRNA-disease association use heterogeneous information learned from different biological sources such as lncRNAs, miRNAs, and diseases. However, learning topological features from diverse network structured data is one of the limiting factors of these methods. To address this challenge, we propose a method for lncRNA-disease association prediction based on Deep Belief Network (DBN), referred to as DBNLDA. In this method, three interaction networks such as lncRNA-miRNA similarity (LMS), disease-miRNA similarity (DMS), and lncRNA-disease association (LDA) network are constructed. A new framework based on the node embedding, DBN, and a neural network regression model is used to learn network and local representation of lncRNA-disease pairs. From the node embedding matrices of LMS, DMS, and LDA networks, lncRNA-disease features are learned by DBN layers. These DBN features are used to predict the association score by an ANN regression model. Compared to several state-of-the-art methods, DBNLDA obtained better AUC (0.96) and AUPR (0.967) under five-fold cross-validation. Case studies on breast, lung, and stomach cancer also affirmed the ability of DBNLDA in predicting potential lncRNAs associated with various diseases.

Keywords Long non-coding RNA · Deep belief network · LncRNA-disease association · Deep learning · Functional similarity networks · Unsupervised feature learning

1 Introduction
Evidence from the high throughput sequencing analysis rescripted the then prominent central dogma of molecular biology by defining the function of non-coding RNAs, which does not convert into proteins. Based on the length of transcripts, these non-coding RNAs could be short or long. Short non-coding RNAs, including microRNA (miRNA) and transfer RNA (tRNA), have been studied extensively for their biological characteristics and functional roles. Long non-coding RNAs (lncRNA) are relatively recent entries that become popular among non-coding RNAs. LncRNAs are RNA transcripts with more than 200 nucleotides in length and lack protein-coding potential [14]. The studies revealed that lncRNAs are involved as a regulator in many biological processes such as ageing, cell differentiation, epigenetic mechanisms, and protein synthesis [21]. The aberrant expressions of lncRNAs are also associated with many complex diseases like cancers [18], Alzheimer’s disease [29], and heart failure [37]. Therefore, identifying the role of lncRNAs in diseases would help to improve the understanding of the disease mechanisms and provide new insights on drug therapeutics [2].

Earlier methods [3, 22] for lncRNA-disease association (LDA) prediction used biological information related to lncRNAs such as genome location and tissue specificity. However, these methods could not be used for diseases without tissue-specific gene records and details of genes.
adjacent to the locations of lncRNAs. A wide range of computational models, especially from the machine learning genre are often used for predicting LDA and identifying disease-associated candidate lncRNAs. However, these methods were challenged due to the less conserved biological properties of lncRNAs. The most successful LDA models considered interactions between lncRNAs, diseases, miRNAs, and protein-coding RNAs and these interactions are effectively represented as networks. Feature extraction from this heterogeneous network doubles the challenge of LDA prediction methods. In recent works [45, 46], introduction of deep learning models eliminated the need for feature extraction by enabling unsupervised representation learning. Methods effectively integrating the deep topological features are important for improving the performance of LDA prediction models.

Addressing these challenges, we propose a deep belief network (DBN) based representation learning model, DBNLDA, for novel lncRNA-disease association prediction. Instead of using a single heterogeneous network of lncRNAs, miRNAs and diseases as in previous works [45, 46], DBNLDA constructs three functional similarity networks-lncRNA-miRNA (LMS), disease-miRNA (DMS), and lncRNA-Disease association (LDA). The node embedding of lncRNA-disease nodes from these networks are utilized by multiple levels of DBN layers for feature learning. For each pair of lncRNA-disease, DBN subnetwork-1 learns lncRNA and disease representations from LMS and DMS networks. Similarly, DBN subnetwork-2 learns lncRNA-disease representations from LDA network. The representations learned by subnetworks are combined for learning higher level representation using the third DBN (DBN-combined). This learning approach enables a multi-modal representation learning of lncRNA and disease pairs. In order to reduce the obstructive impact of features learned from sparse network, the features from DBN-combined are recomputed using an attention layer. Finally, a neural network based regression model is used to score the lncRNA-disease association. The Area under ROC curve (AUC-ROC) and Area under Precision-Recall curve (AUPR) values under five-fold cross-validation shows the proposed model have significant improvement over several contemporary LDA models. Moreover, case studies on cancer dataset show the effectiveness of DBNLDA in predicting potential lncRNAs associated with major diseases. The main contributions of this paper are as follows:

1. Proposed a deep belief network (DBN) based representation learning for lncRNA-disease association (LDA) prediction
2. Proposed a multi-modal learning approach for lncRNA-disease representation from lncRNA, miRNA, and disease interactions
3. Potential lncRNA associated with 412 major diseases in the dataset are identified using the score computed by the proposed method

The rest of the paper is organized as follows. Remaining part of this section gives necessary background of deep belief networks. Section 2 discusses a review of recent methods in the area of lncRNA-disease association prediction. Architectural details of DBNLDA is discussed in Section 3. Section 4 discusses the results of the proposed method and Section 5 concludes the paper.

1.1 Deep belief networks

Deep Belief Networks (DBN), proposed by Hinton et al., [17] are stacked layers of Restricted Boltzmann Machines (RBM). DBNs are generative graphical models to extract deep hierarchical representation from training data. DBN architecture contains an input layer (visible layer) followed by a stack of hidden layers, where each hidden layer is the visible layer of the next layer. A DBN model compute the joint distribution between visible unit v and hidden layers \((μ^1, ..., μ^ℓ)\) as in (1).

\[
Prob(v, μ^1, μ^2, ..., μ^ℓ) = \left(Π_{k=0}^{ℓ-2} Prob(μ^k|μ^{k+1})\right) Prob(μ^{ℓ-1}, μ^ℓ)
\]  

where \(μ^0 = v\), \(Prob(μ^k|μ^{k+1})\) is the conditional probability at hidden layer k, and \(Prob(μ^{ℓ-1}, μ^ℓ)\) is the joint distribution at final layer. The probability distribution given in (1) can be treated as latent representation learned by DBN from the input.

2 Related work

An array of computational models to predict lncRNA-disease association are available in the literature with varying performance. All these works utilise varieties of lncRNA features from known lncRNA-disease associations and their interactions with other molecules like microRNAs (miRNA), proteins, and messenger RNAs (mRNA). The earlier methods for LDA prediction used various biological properties of lncRNAs, as discussed in the previous section. The performance and scalability of these methods are limited due to the lack of biological knowledge about lncRNAs. With the advancement of machine learning based methods, more accurate and scalable methods were reported. This section reviews some of the recent machine learning based methods for LDA. These methods can be grouped into three major categories. The first category of works utilises knowledge of lncRNA functional similarity under the assumption that functionally similar lncRNAs associate with similar
Based on this assumption, an lncRNA-disease association network was constructed. Among this category of works, Sun et al., [36] proposed LDA model named RWRlncD using random walk with restart on lncRNA similarity network. Similarly, the works such as IRWRLDA [8], BRWLDA [47] used variants of random walk algorithm to predict LDA. However, none of the above approaches can be used for new diseases that do not have any associated experimentally supported lncRNAs. This limitation was addressed by Chen et al., [4] by proposing KATZLDA, which applies Katz page ranking algorithm for analysing the similarity network. This work integrated known lncRNA-disease associations, lncRNA expression profiles, lncRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity for LDA prediction. All these methods in the first category relied upon the network structure features and results were biased towards nodes having high degree and centrality.

Advanced research on lncRNA mechanism revealed that the regulation of lncRNA is largely determined by co-expressed miRNAs [31]. The second category of works used expression levels of lncRNAs, mRNAs, and miRNAs in various diseases. The earlier methods [46] in this category used experimentally validated disease associated mRNAs/miRNAs and lncRNA co-expression data. The LDA model proposed by Liu et al., [26] used expression profiles of experimentally validated lncRNAs, mRNAs, and miRNAs whereas Li et al., [22] proposed a genome location based model. These methods were not useful with lncRNAs which have no experimentally validated mRNA/miRNA interactions. The method MFLDA proposed by Lu et al., [27] used matrix factorization on various lncRNA association matrices for the LDA prediction. Moreover, there are graph-based algorithms such as TPGLDA [10] and DisLncPri [39]) to predict LDA. TPGLDA proposed by Ding et al., was based on lncRNA-disease-mRNA tripartite graph construction. DisLncPri proposed by Wang et al., explored competing endogenous RNA (ceRNA) network theory which maps lncRNAs to their functional genomic context.

The third category of works constructed heterogeneous interaction network based on lncRNA, miRNA and mRNA functional similarity. Chen et al., proposed a Laplacian regularized least squares based LDA prediction model (LRLSLDA) [5]. LRLSLDA was a semi-supervised learning model which did not require negative samples to train. Later, Chen et al., improved LRLSLDA with LNCSIM by combining functional and expression features of lncRNAs. Yao et al., proposed a random walk based feature selection method for predicting LDA [46]. This method take the advantage of ensemble learning to improve the prediction accuracy. Lan et al., proposed a web tool called LDAP, which utilised support vector machine (SVM) for the LDA prediction [20].

Major challenge in the above methods was the effective representation of lncRNA-disease features. Introduction of deep learning models eliminated the need for feature extraction by enabling unsupervised representation learning. Among them, CNNLDA [44] used convolutional neural networks and GCNLDA [45] used graph convolutional neural networks [19] to learn global representations of lncRNA, miRNA, and disease nodes. A recent work named GAMCLDA [43] used graph autoencoder and matrix completion to predict lncRNA-disease association.

Deep Belief Networks (DBN) are used recently in computer vision and text mining for representation learning and classification [17]. DBN based models are successfully used in Bioinformatics to predict drug-target [41], multiple types of miRNA-disease association [7, 28], and cancer sub-type prediction [24], but not applied in the field of lncRNA-disease association prediction. Based on the status of the above-mentioned studies, we are proposing an LDA prediction model (DBNLDA) using features learned by multiple DBNs.

### 3 Proposed method

DBNLDA make use of heterogeneous information on functional similarity, co-expression and interactions between lncRNAs and diseases for making the prediction. The architecture of DBNLDA contains 3 modules namely-(i) network construction, (ii) DBN based feature learning, and (iii) association prediction. The proposed architecture is shown in Fig 1. Following subsections describe the dataset used for LDA prediction and the proposed method.

#### 3.1 The dataset

Datasets used in the previous works [13, 45, 46] are used for lncRNA-disease association, lncRNA-miRNA interaction, lncRNA functional similarity, and disease semantic similarity. LncRNA-disease associations were downloaded from two reference databases: LncRNADisease [3] and lnc2cancer [30]. The miRNA-lncRNA interactions and miRNA-disease interactions were obtained from miRNet [12] and Starbase [23] databases respectively. All these downloaded associations and similarities were then compiled for 240 lncRNAs, 412 diseases, 495 miRNAs, and 2697 known lncRNA-disease interactions. These known interaction pairs constituted the positive samples for training the model. All other pairs between lncRNAs and diseases, which are not listed in reference databases, were considered to be negative samples. We randomly selected 2697 samples from negative samples to construct a balanced dataset. Summary of the dataset is given in Table 1.
3.2 Construction of LMS, DMS and LDA networks

The first step in the DBNLDA architecture is the construction of three similarity networks such as LMS, DMS, and LDA as defined in Section 1. The known interactions between lncRNAs, miRNAs and diseases form edges in the networks. Let \( n_l \), \( n_m \), and \( n_d \) be the number of lncRNAs, miRNAs and diseases respectively in the dataset. LMS network (with \( n_l \) number of lncRNAs and \( n_m \) number of miRNAs) was constructed using lncRNA-lncRNA similarity and lncRNA-miRNA interactions. Functional similarities between lncRNAs are computed by Chen’s method [6] in which functional similarity is based on the semantic similarity between their associated disease groups. An edge is added if the similarity score is greater than 0. For the list of lncRNAs, the known miRNA targets downloaded from the miRNet are used as lncRNA-miRNA edges.

Similar to LMS, DMS network is constructed (with \( n_d \) number of diseases and \( n_m \) number of miRNAs) using disease-disease similarities and known disease-miRNA associations. Disease-disease similarity is calculated using the method proposed by Chen et al., [6]. This model represents disease as a directed acyclic graph (DAG) of disease terms. Then disease semantic similarity between two diseases are calculated based on the nodes shared by two DAGs. Finally, LDA network contains \( n_l \) number of lncRNAs and \( n_d \) number of diseases, where an un-directed edge is used to represent the known association between lncRNA and disease.

3.3 Computing node embedding features from networks

Node embedding is an important step in network modeling which maps a node into a real vector. In DBNLDA, Node2vec [15] algorithm is used to get node embedding from each network. For each node in a network, Node2vec outputs a \( d \)-dimensional real vector based on the neighbourhood information of the node.

First, LMS network is passed through Node2vec layer to get node representation of lncRNAs and miRNAs. From the output of Node2vec, the embedding vector for each lncRNA is obtained as \( L_{lms}^{i} = [l_1^i, l_2^i, ..., l_e^i] \in \mathbb{R}^e \), where \( e \) is the embedding dimension. All these embedding vectors are combined to form the embedding matrix \( L_{lms} \in \mathbb{R}^{n_l \times e} \). Similarly, the embedding matrix for diseases from DMS is obtained as \( D_{dms} \in \mathbb{R}^{n_d \times e} \). Then, for each lncRNA-disease pair \((l_i, d_j)\) in the dataset (both positive and negative samples), lncRNA embedding vector from LMS, \( L_{lms}^{i} \) and disease embedding vector from DMS \( D_{dms}^{j} \) are concatenated to form the vector \( LD_{lmd}^{ij} \in \mathbb{R}^{2e} \). This resulted in a feature matrix \( LD_{lmd} \in \mathbb{R}^{n \times 2e} \), where \( n \) is the total number of samples in the dataset. Finally, the embedding matrices of lncRNAs and diseases, represented as \( L_{lda} \in \mathbb{R}^{n_l \times e}, D_{lda} \in \mathbb{R}^{n_d \times e} \) are obtained from LDA network. Then as in the case of \( LD_{lmd} \), for each pairs of lncRNA-disease, vectors from \( L_{lda} \) and \( D_{lda} \) are concatenated to form feature matrix \( LD_{lda} \in \mathbb{R}^{n \times 2e} \).
3.4 DBN based feature learning

Following the Node2vec embedding layer, DBNLDA implements Deep Belief Network based feature learning. In this work, DBN is used to learn latent representation of lncRNA and disease nodes and encode them to a new dimension \( h \), where \( h \geq e \). The architecture consists of two DBN subnetworks-one to learn lncRNA-disease representation from functional similarity networks (LMS and DMS) and other from LDA network. The DBN subnetwork-1 receives embedded feature matrix \( LD_{lmd} \) as input and produces \( LD_{db1} \in \mathbb{R}^{n \times h} \) (\( h \), the number of hidden units in DBN) as output. Similarly, the DBN subnetwork-2 receives \( LD_{lda} \) as input and produces \( LD_{db2} \in \mathbb{R}^{n \times h} \) as output. Then, both \( LD_{db1} \) and \( LD_{db2} \) are concatenated to form combined feature representation, \( LD \in \mathbb{R}^{n \times 2h} \). The network DBN-combined accepts \( LD \) as input and learn the feature representation as \( LD_{db} \in \mathbb{R}^{n \times h'} \), where \( h' \) is the number of hidden units in the DBN-combined and \( h' \geq h \).

3.5 Feature attention layer

Since all three networks used in the DBNLDA architecture are sparse in nature, the latent representation learned by DBN has a chance to lose feature importance. Furthermore, the use of three DBN modules has increased the number of learning parameters and may cause overfitting problems. The attention mechanism in deep learning is used to solve these issues by recomputing the feature values from all available information so that their contributions could be different and unique. In this work, attention mechanism similar to the one used in GCNLDA [45] is applied to \( LD_{db} \) features.

Let \( LD_{db} = [ld_1^1, ld_1^2, ..., ld_1^h] \) represents the feature vector of \( i^{th} \) entry in the dataset learned by DBN-combined, where \( ld_i^j \in \mathbb{R}, \forall k = 1, 2, ..., h' \). The attention score for each element in \( LD_{db} \) is calculated by introducing attention weight parameters \( H_{att} \in \mathbb{R}^{h' \times h'} \), \( W_{att} \in \mathbb{R}^{h' \times h'} \) and bias \( b_{att} \in \mathbb{R}^{h'} \) as in (2). \( a_{i}^{att} = Softmax(H_{att} \cdot tanh(W_{att} LD_i + b_{att})) \) \hspace{1cm} (2)

Next, the attention enhanced feature values are recomputed as in (3). \( LD_{i}^{att} = a_{i}^{att} \otimes LD_{db} \) \hspace{1cm} (3) where \( \otimes \) represents pairwise multiplication. Finally, the matrix \( LD_{i}^{att} = [LD_{i}^{att}] \) was used as input for association prediction.

3.6 Prediction Layer

The final module of DBNLDA architecture is a neural network regression model to score the association between lncRNA and disease. The regression network follows a 1-2-1 topology, where the input layer has \( h' \) nodes to receive input from \( LD_{att} \) and output layer has one node to compute the association score. The number of nodes in hidden layers is kept as hyperparameter. Activation function used in all layers except the output is ReLU. Output layer used a sigmoid function to compute the association score. In order to reduce the overfitting, a dropout of 0.02 probability is added between the hidden layers. The learning is measured by computing binary cross entropy function, as this network outputs probability distribution for binary classification. Let \( y_i \in \{0, 1\} \) is the actual label and \( p(y_i) \in \mathbb{R} \) is the probability predicted by the classification model. Then the cross entropy loss \( L(Y) \), for \( Y = [y_1, y_2, ..., y_n] \), is calculated as in (4).

\[
L(Y) = -\frac{1}{n} \sum_{y_i \in Y} y_i \log(p(y_i)) + (1 - y_i) \log(p(1 - y_i)) \hspace{1cm} (4)
\]

In addition, the network is optimized by ADAM optimizer with learning rate 0.01. The entire workflow of neural network classifier can be summarized as in (5). \( score = Sigmoid(ReLU(Linear(ReLU(Linear(LD_{att})))))) \) \hspace{1cm} (5)

3.7 Hyperparameters

Various hyperparameters determine the performance of DBNLDA in different modules. The values of these hyperparameters are empirically tuned using a grid search method. The dimension of Node2vec embedding, \( e \), is selected from \{16, 32, 64, 128, 256, 512\} for all networks by keeping other parameters to default values as in [15]. Following the implementation of [28], the DBN architectures in this work uses three stacked layers of RBM, with \( h \) and \( h' \) takes values from \{64, 128, 256, 512\}. DBNLDA gave best performance when \( e = 64, h = 128, \) and \( h' = 256 \). It was found that very low values of \( h \) and \( h' \) degrade the performance and high values have no effect on the performance of the model. Figure 2 shows the details. For the classification module, the number of neurons in both hidden layers set as 128. The classifier iterate over 30 epochs, since the model performance became stable after 30\(^{th}\) epoch.

A detailed description DBNLDA steps and implementation details are available in the SupplementaryFile-S1.pdf.
4 Results and discussion

4.1 Performance evaluation metrics

The experiments used five-fold cross-validation to evaluate the performance of DBNLDA and other LDA prediction models. Four evaluation indexes—TPR (True Positive Rate), FPR (False Positive Rate), AUPR (Area under Precision-Recall curve), and ROC (Receiver Operating Characteristic)—were used to evaluate the models. For five-fold cross validation, all known lncRNA-disease associations were treated as positive samples. Conversely, pairs not known to be associated were treated as negative samples. The number of positive samples in the entire dataset was far less compared with negative samples. So, the same number of negative samples as the total number of positive samples were picked according to random sampling.

The experiments were repeated with different random selection of negative samples and mean results with standard deviation are reported. When the score calculated by the model was above a threshold (say \( \theta \)), the prediction was treated as positive. Then TP (true positive) and TN (true negative) represented the number of correctly identified positive and negative samples by the model. FP and FN are their misidentified counterparts. Then True positive rate (TPR) defines how many positive results occurs among all positive samples and False positive rate (FPR) defines incorrect positive results occurs among all negative samples. Similarly, precision is the portion of TP among all the samples predicted as positive by the model and recall measures the ratio of TP with actual positive samples in the dataset [48].

The ROC curve was plotted by TPR and FPR under each value of \( \theta \), and area under the ROC curve (AUC) was calculated to evaluate the performance of the lncRNA-disease prediction models. Similarly, the plot with precision and recall values calculated for different \( \theta \) was used to calculate AUPR. Since the dataset is negatively skewed, the use of AUPR gave a better comparison of algorithms [9]. Particularly AUPR is critical as identifying positive relation is more important and data set containing low true positive baseline.

4.2 Overall performance

The DBNLDA trained over 30 epochs, and the learning curve (refer Fig. 3) showed consistent characteristics in all folds of cross-validation. The model gave an average AUC of 0.96 (±0.0002) over cross-validations. The ROC curve is shown in Fig. 4. The model reported accuracy of 0.957 (±0.0002) and AUPR value of 0.968 (±0.00265). A Chi-square significance test (with a significant level, \( \alpha = 0.05 \)) was also conducted to test the statistical significance of the predictions by the proposed model. The p-value obtained is \( 1 \times 10^{-5} \), which is less than the significant level 0.05, which confirms that the predictions by DBNLDA model were statistically significant.

In order to analyse how DBN based features improve the performance of the prediction model, we had repeated the experiments with different levels of feature combinations from Node2vec embedding to various DBN representations. Table 2 summarises the result of the experiments. It is clear from the results that introduction of DBN based learning significantly improved the accuracy of the prediction model.
4.3 Comparison with other methods

The performance of DBNLDA was compared with other state-of-the-art methods such as RFLDA [46], GCNLDA [45], SIMCLDA [27], Ping’s Method [33], MFLDA [13], LDAP [20], GAMCLDA [43], and CNNLDA [44]. The above methods used knowledge from heterogeneous information from lncRNA, miRNA and disease associations to predict LDA. The same experimental setup described in [45, 46] were followed with 412 diseases and 2697 known LDA. All the above methods expect RFLDA used network based deep learning model for LDA prediction. RFLDA used an ensemble learning based feature selection method. The results of the comparison are given in Table 3. The AUC and AUPR values of all methods except DBNLDA were taken from [46] and [43]. It is evident from the table that DBNLDA reported second best AUC (0.96) which is closer (1.3% less) to the highest AUC value reported by RFLDA. Compared to the deep learning models, AUC value of DBNLDA shows 10.38% improvement over LDAP, 9.55% improvement over Ping’s method, 5.81% improvement over GAMCLDA, 1.14% improvement over CNNLDA, and 0.4% improvement over GCNLDA. This clearly shows the significance of DBN based feature representation compared to other convolutional and autoencoder based models. DBNLDA also gave 34.99% improvement over MFLDA and 22.53% improvement over SIMCLDA. These methods did not consider network topology features which cause inferior performance. However, RFLDA gave slight improvement
over DBNLDA since it used a bagging approach for selecting strong features. Fine tuning the hyperparameters may improve the performance of DBNLDA further.

On the other hand, DBNLDA achieved AUPR score of 0.968 for all tested 412 diseases, which was higher than all other methods involved in comparison. Specifically, DBNLDA shows 19.52% improvement over the second highest value (RFLDA). Among the deep learning methods, DBNLDA give 76.96% improvement over GCNLDA, 74.04% improvement over CNNLDA, 77.37% improvement over Ping’s method, and 96.17% improvement over GAMCLDA. Compared to the other methods, DBNLDA combines local and global node embeddings of various interaction networks to learn higher level representation of lncRNA-disease association. The results obtained clearly show the significance of proposed approach in predicting lncRNA-disease associations.

To verify whether the proposed model was significantly better than the other deep learning based models, paired Wilcoxon tests on DBNLDA and other methods (GCNLDA, GAMCLDA, and CNNLDA) was conducted. It is clear from the p-values of the tests (Table 4), the relative performance of DBNLDA is statistically significant compared to other models.

### 4.4 Case studies

To further investigate the ability of DBNLDA in predicting significant lncRNA-disease associations, case studies on breast cancer, lung cancer, and stomach cancer were conducted. For this study, first trained the DBNLDA model on a dataset containing all lncRNA-disease associations except the validated associations between lncRNAs and the disease of interest (breast/lung/stomach cancer). Then the association score for all lncRNAs to the particular disease was calculated using the trained model and analysed the top 15 candidate lncRNAs for each disease. Tables 5, 6, and 7 show the top 15 candidate lncRNAs respectively in breast, lung, and stomach cancers, predicted by DBNLDA. The evidence column shows the reference to the associations either from reference databases or from literature.

It was found that 13 (86.67%) lncRNAs associated with breast cancer predicted by DBNLDA were also confirmed by lnc2cancer or LncRNADisease database. For the two unconfirmed predictions, we could find the evidence from recent publications. In the case of lung cancer, among the top 15 predicted lncRNAs, 10 were confirmed by reference databases and remaining were reported in recent literature.

In case of stomach cancer, DBNLDA could predict 12 associations reported either by reference databases or literature. The three new associations (HCP5, HCG4, and MIR99AHG, indicated by ‘*’ in Table 7) which could be considered as new suggestions for further laboratory

---

**Table 2** Accuracy of the model based on feature combinations

| Experiment | Feature | Accuracy |
|------------|---------|----------|
| Exp 1 only | Node2vec features | 0.817 |
| Exp 2 | Node2vec, DBN subnetwork-1 and DBN subnetwork-2 | 0.896 |
| Exp 3 | Node2vec, DBN subnetwork-1, DBN subnetwork-2 and DBN-combined | 0.956 |

**Table 3** Comparison of performance of DBNLDA with state-of-the-art methods

| Method      | AUC  | AUPR  |
|-------------|------|-------|
| MFLDA [13]  | 0.626| 0.066 |
| SIMCLDA [27]| 0.746| 0.095 |
| LDAP [20]   | 0.863| 0.166 |
| Ping’s Method [33] | 0.871| 0.219 |
| GAMCLDA [43]| 0.907| 0.037 |
| CNNLDA [44]| 0.952| 0.251 |
| GCNLDA [45]| 0.959| 0.223 |
| RFLDA [46]  | 0.976| 0.779 |
| **DBNLDA**  | 0.963| 0.968 |

**Table 4** p-values obtained for paired Wilcoxon test with DBNLDA and other methods

| Method     | ROC-AUC | AUPR  |
|------------|---------|-------|
| GCNLDA     | 0.042   | 0.021 |
| GAMCLDA    | 0.021   | 0.020 |
| CNNLDA     | 0.021   | 0.0206|

**Table 5** Top 15 DBNLDA predicted lncRNAs associated with breast cancer

| IncRNA      | Rank | Evidence            |
|-------------|------|---------------------|
| GAS5        | 1    | Lnc2Cancer          |
| DLEU2       | 2    | LncRNA-Disease      |
| HCP5        | 3    | Literature (PMID: 31864836) [42] |
| HOTAIR      | 4    | LncRNA-Disease, Lnc2Cancer |
| MEG3        | 5    | LncRNA-Disease, Lnc2Cancer |
| HULC        | 6    | Lnc2Cancer          |
| BCYRN1      | 7    | LncRNA-Disease      |
| HOTTIP      | 8    | Lnc2Cancer          |
| UCA1        | 9    | LncRNA-Disease, Lnc2Cancer |
| CDKN2B-AS1  | 10   | LncRNA-Disease      |
| NEAT1       | 11   | LncRNA-Disease, Lnc2Cancer |
| TUG1        | 12   | LncRNA-Disease, Lnc2Cancer |
| AFAP1-AS1   | 13   | Lnc2Cancer          |
| MIR100HG    | 14   | Literature (PMID:30042378) [40] |
| TINCR       | 15   | Lnc2Cancer          |
Table 6  Top 15 DBNLDA predicted lncRNAs associated with lung cancer

| lncRNA   | Rank | Evidence                  |
|----------|------|---------------------------|
| TUG1     | 1    | Literature (PMID:31532756) [16] |
| PVT1     | 2    | Lnc2Cancer                |
| AFAP1-AS1| 3    | LncRNADisease, Lnc2Cancer  |
| XIST     | 4    | Literature (PMID: 28448993) [38] |
| CCAT2    | 5    | LncRNADisease              |
| MALAT1   | 6    | LncRNADisease, Lnc2Cancer  |
| HOTTIP   | 7    | LncRNADisease, Lnc2Cancer  |
| SOX2-OT  | 8    | LncRNADisease              |
| HULC     | 9    | Literature (PMID:30575912) [25] |
| MIR155HG | 10   | Literature (PMID:32129458) [35] |
| CDKN2B-AS1| 11   | Literature (PMID:29541247) [11] |
| BANCR    | 12   | LncRNADisease, Lnc2Cancer  |
| BCYRN1   | 13   | LncRNADisease              |
| UCA1     | 14   | LncRNADisease, Lnc2Cancer  |
| H19      | 15   | LncRNADisease, Lnc2Cancer  |

validations. The detailed comparison of LDA prediction by DBNLDA is available in the SupplementaryFile-S2.pdf.

4.5 Limitations of DBNLDA

There are some limitations in DBNLDA model. As a supervised model, DBNLDA requires both positive and negative samples to learn the association function. However, it is not practical to obtain reliable negative samples for the lncRNA-disease association. The method of random sampling may affect the performance of DBNLDA. This issue is addressed by taking the statistical summary by repeating the experiments on different random samples. Besides, limited knowledge about diseases, lncRNA, and miRNAs also challenges the prediction performance of all lncRNA-disease prediction models. Additionally, the complex architecture with three DBN creates a large number of learnable parameters, increasing the running time of the model. A thorough fine-tuning of model hyperparameters and the use of parallel computing can solve this challenge to an extent.

The complexity analysis of DBNLDA model reveals that the computational complexity is \( \max(O(n \times 2h \times h'^{3}), O(\epsilon \times n \times (h' \times i + i \times j + j)) \) \( (h \text{ and } h' \text{ are the number of neurons in the DBN-subnetworks and DBN-combined respectively, } \epsilon \text{ is the number of epoch, } n \text{ is the number of training examples and } i \text{ and } j \text{ are the number of neurons in the hidden layers of ANN}), \text{ where the first component is the running time of DBN based feature learning and the second component is that of ANN based regressor. Since } i, j < h', \text{ it can be inferred from the above equation that the complexity of DBNLDA depends on the complexity } (O(n \times 2h \times h'^{3})) \text{ of DBN based feature learning, which is polynomial in terms of the number neurons in DBN hidden layers. Details of complexity analysis is provided in SupplementaryFile-S1.pdf.}

5 Conclusion

In this work, a deep learning framework consisting of DBN for predicting lncRNA-disease association was proposed. The model implements a DBN based multi-modal feature learning approach to extract features from lncRNA, miRNA, and disease interaction networks. The experimental evaluations and statistical tests confirm that DBNLDA achieves comparable performance in AUC and significant improvement in AUPR. Case studies on breast cancer, lung cancer and stomach cancer also show the ability of DBNLDA to predict potential disease-associated lncRNAs. However, the use of randomly selected negative samples and a large number of learnable DBN parameters are some of the limitations of this model. The model could be extended further with multi-modal data such as lncRNA-drug-target interactions and lncRNA-epigenetic-disease interactions. Therefore, DBNLDA can provide help for the mechanism studies of lncRNAs in diseases in the future.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10489-021-02675-x.

Acknowledgements This research work is an outcome of the R&D work under the Visvesvaraya PhD Scheme of Ministry of Electronics and Information Technology, Government of India.
Author Contributions Both authors contributed equally in conception, design and implementation of the proposed idea and manuscript preparation.

Availability of data and material https://github.com/manumad/DBNLDA

Code availability https://github.com/manumad/DBNLDA

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Bahari F, Emadi-Baygi M, Nikpour P et al (2015) miR-17-92 host gene, uderexpressed in gastric cancer and its expression was negatively correlated with the metastasis. Ind J Cancer 52(1):22
2. Bhartiya D, Kapoor S, Jalali S, Sati S, Kaushik K, Sachidanandan C, Sivasubbu S, Scaria V (2012) Conceptual approaches for IncRNA drug discovery and future strategies. Expert Opin Drug Discov 7(6):503–513
3. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q (2012) LncRNA-Disease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res 41(D1):D983–D986
4. Chen X (2015) KATZLDA: KATZ Measure for the IncRNA-disease association prediction. Sci Rep 5:16840
5. Chen X, Yan CC, Luo C, Ji W, Zhang Y, Dai Q (2015) Constructing IncRNA functional similarity network based on IncRNA-disease associations and disease semantic similarity. Sci Rep 5:11338
6. Chen X, Yan CC, Luo C, Ji W, Zhang Y, Dai Q (2015) Constructing IncRNA functional similarity network based on IncRNA-disease associations and disease semantic similarity. Sci Rep 5:11338
7. Chen X, Yan CC, Zhang X, Li Z, Deng L, Zhang Y, Dai Q (2015) RBMMMDA: Predicting multiple types of disease-microRNA associations. Sci Rep 5:13877
8. Chen X, You Z-H, Yan G-Y, Gong D-W (2016) IRWRRLDA: Improved random walk with restart for IncRNA-disease association prediction. Oncotarget 7(36):57919
9. Davis J, Goadrich M (2006) The relationship between precision- recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning, pp 233–240
10. Ding L, Wang M, Sun D, Li A (2018) TPGLDA Novel Prediction of associations between IncRNAs and diseases via IncRNA-disease-gene tripartite graph. Sci Rep 8(1):1–11
11. Du Y, Hao X, Liu X (2018) Low expression of long noncoding mircDNA-292 induces apoptosis and suppresses growth of hepatocellular carcinoma cells. Mol Med Rep 15(4):4912–4918
12. Fan Y, Siklenka K, Arora SK, Ribeiro P, Kimmins S, Xia J (2016) miRNet-dissecting miRNA-target interactions and functional associations through network-based visual analysis. Nucleic Acids Res 44(W1):W135–W141
13. Fu G, Wang J, Domeniconi C, Yu G (2018) Matrix factorization-based data fusion for the prediction of IncRNA–disease associations. Bioinformatics 34(9):1529–1537
14. Goff LA, Rinn JL (2015) Linking RNA biology to IncRNAs. Genome Res 25(10):1456–1465
15. Grover A, Leskovec J (2016) node2vec Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pp 855–864
16. Guo S, Zhang L, Zhang Y, Wu Z, He D, Li X, Wang Z (2019) Long non-coding RNA TUG1 enhances chemosensitivity in non-small cell lung cancer by impairing microRNA-221-dependent PTEN inhibition. Aging (Albany NY) 11(18):7553
17. Hinton GE (2009) Deep belief networks. Scholarpedia 4(5):5947
18. Huarte M (2015) The emerging role of IncRNAs in cancer. Nat Med 21(11):1253
19. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv:1609.02907
20. Lan W, Li M, Zhao K, Liu J, Wu F-X, Yi P, Wang J (2017) LDAP: A web server for IncRNA-disease association prediction. Bioinformatics 33(3):458–460
21. Lee JT (2012) Epigenetic regulation by long noncoding RNAs. Science 338(6113):1435–1439
22. Li JW, Gao C, Wang Y, Ma W, Tu J, Wang JP, Chen Z, Kong W, Cui Q (2014) A bioinformatics method for predicting long noncoding mRNAs associated with vascular disease. Sci China Life Sci 57(8):852–857
23. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-J (2014) starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res D1:D92–D97
24. Liang M, Li Z, Chen T, Zeng J (2014) Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. IEEE/ACM Trans Comput Biol Bioinform 12(4):928–937
25. L Liu XY, Zhou JQ, Zhang G, Wang J, He Y, Chen C, Huang L (2018) Li, and SQ Li. LncRNA HULC promotes non-small cell lung cancer cell proliferation and inhibits the apoptosis by up-regulating sphingosine kinase 1 (SphK1) and its downstream PI3/akt pathway. Eur Rev Med Pharmacol Sci 22:8722–8730
26. Liu M-X, Chen X, Chen G, Cui Q-H, Yan G-Y (2014) A computational framework to infer human disease-associated long noncoding mRNAs. PloS One 9(1):e84408
27. Lu C, Yang M, Luo F, Wu F-X, Li M, Yi P, Li Y, Wang J (2018) Prediction of IncRNA–disease associations based on inductive matrix completion. Bioinformatics 34(19):3357–3364
28. Luo P, Li Y, Tian L-P, Wu F-X (2019) Enhancing the prediction of disease–gene associations with multimodal deep learning. Bioinformatics 35(19):3735–3742
29. Luo Q, Chen Y (2016) Long noncoding RNAs and Alzheimer’s disease. Clin Intervent Aging 11:867
30. Ning S, Zhang J, Wang P, Zhi H, Wang J, Liu Y, Gao Y, Guo M, Yue M, Wang L et al (2016) Lnc2cancer: a manually curated database of experimentally supported IncRNAs associated with various human cancers. Nucleic Acids Res 44(D1):D980–D985
31. Paraskevopoulou MD, Hatziioannou AG (2016) Analyzing miRNA–lncRNA interactions. In: long non-coding RNAs. Springer, pp 271–286
32. Piao H-Y, Guo S, Wang Y, Zhang J (2019) Long noncoding RNA NALT1-induced gastric cancer invasion and metastasis via NOTCH signaling pathway. World J Gastroenterol 25(44):6508
33. Ping P, Wang L, Kuang L, Ye S, Iqbal MFB, Pei T (2018) A novel computational framework to infer human disease-associated long noncoding mRNAs. PloS One 9(1):e84408
34. Ren H, Yang X, Yang Y, Zhang X, Zhao R, Wei R, Zhang X, Yi Z (2018) Upregulation of IncRNA BCYRN1 promotes tumor progression and enhances epCAM expression in gastric carcinoma. Oncotarget 9(4):4851
35. Song J, Wang Q, Zong L (2020) LncRNA MIR155HG contributes to smoke-related chronic obstructive pulmonary disease by targeting miR-128-5p/BRD4 axis. Biosci Rep 40(3)
36. Sun J, Shi H, Wang Z, Zhang C, Liu L, Wang L, He W, Hao D, Liu S, Zhou M (2014) Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network. Mol BioSyst 10(8):2074–2081

37. Wang G, Zheng X, Zheng Y, Cao R, Zhang M, Sun Y, Wu J (2019) Construction and analysis of the lncRNA-miRNA-mRNA network based on competitive endogenous RNA reveals functional genes in heart failure. Mol Med Rep 19(2):994–1003

38. Wang H, Shen Q, Zhang X, Yang C, Su C, Sun Y, Wang L, Fan X, Xu S (2017) The long non-coding RNA XIST controls non-small cell lung cancer proliferation and invasion by modulating miR-186-5p. Cell Physiol Biochem 41(6):2221–2229

39. Wang P, Guo Q, Gao Y, Zhi H, Zhang Y, Liu Y, Zhang J, Yue M, Guo M, Ning S et al (2017) Improved method for prioritization of disease associated lncRNAs based on ceRNA theory and functional genomics data. Oncotarget 8(3):4642

40. Wang S, Ke H, Zhang H, Ma Y, Ao L, Zou L, Yang Q, Zhu H, Nie J, Wu C et al (2018) LncRNA MIR100HG promotes cell proliferation in triple-negative breast cancer through triplex formation with p27 loci. Cell Death Disease 9(8):1–11

41. Wang Y, Zeng J (2013) Predicting drug-target interactions using restricted Boltzmann machines. Bioinformatics 29(13):i126–i134

42. Wu J, Chen H, Ye M, Wang B, Zhang Y, Sheng J, Meng T, Chen H (2019) Downregulation of long noncoding RNA HCP5 contributes to cisplatin resistance in human triple-negative breast cancer via regulation of PTEN expression. Biomed Pharmacother 115:108869

43. Wu X, Lan W, Chen Q, Dong Y, Liu J, Peng W (2020) Inferring LncRNA-disease Associations Based: On Graph Autoencoder Matrix Completion. Computational Biology and Chemistry, pp 107282

44. Xuan P, Cao Y, Zhang T, Kong R, Zhang Z (2019) Dual convolutional neural networks with attention mechanisms based method for predicting disease-related lncRNA genes. Front Gen 10:416

45. Xuan P, Pan S, Zhang T, Liu Y, Sun H (2019) Graph Convolutional Network and Convolutional Neural Network Based Method for Predicting lncRNA-disease Associations. Cells 8(9):1012

46. Yao D, Zhan X, Zhan X, Kwoh CK, Li P, Wang J (2020) A random forest based computational model for predicting novel lncRNA-disease associations. BMC Bioinform 21(1):1–18

47. Yu G, Fu G, Lu C, Ren Y, Wang J (2017) BRWLDA: Bi-random walks for predicting lncRNA-disease associations. Oncotarget 8(36):60429

48. Han J, Kamber M, Pei J (2011) Data mining concepts and techniques. The Morgan Kaufmann Series in Data Management Systems, pp 5.4

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

Manu Madhavan received his PhD in Computer Science and Engineering from National Institute of Technology Calicut (NITC), Kerala, India in 2021. Presently he is working as an Assistant Professor in the Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India. His research focus is on Long Non-coding RNA functional analysis and annotation. He received his M.Tech in Computational Linguistics from Calicut University. His areas of interests include Bioinformatics, Data-mining, Machine Learning Natural Language Processing.

G. Gopakumar received his PhD in Bioinformatics from the University of Kerala, India in 2013. Presently he is working as an Assistant Professor in the Department of Computer Science and Engineering, National Institute of Technology Calicut, India since July 2010. His research interests include RNA Bioinformatics, Biological Network Analysis and Data Mining. He is a member of the IEEE as well as ACM.