# Predicting Long Non-coding RNA-disease Associations using Multiple Features and Deep Learning

Van Tinh Nguyen[1,2], Dang Hung Tran[2]

[1] Hanoi University of Industry, Hanoi, Vietnam
[2] Hanoi National University of Education, Hanoi, Vietnam

Correspondence: Dang Hung Tran. Email:hungtd@hnue.edu.vn
Communication: received 11 June 2022, revised 01 August 2022, accepted 24 August 2022
DOI: 10.32913/mic-ict-research.v2022.n2.1069

***Abstract*: Various long non-coding RNAs have been shown to play crucial roles in different biological processes including cell cycle control, transcription, translation, epigenetic regulation, splicing, differentiation, immune response and so forth in the human body. Discovering lncRNA-disease associations promotes the awareness of human complex disease at molecular level and support the diagnosis, treatment and prevention of complex diseases. It is costly, laboratory and time-consuming to discover and verify lncRNA-disease associations by biological experiments. Therefore, it is crucial to develop a computational method to predict lncRNA-disease associations to save time and resources. In this paper, we proposed a new method to predict lncRNA-disease associations using multiple features and deep learning. Our method uses a weighted $K$-nearest known neighbors algorithm as a pre-processing step to eliminate the impact of sparsity data problem. And it combines the linear and non-linear features extracted by singular value decomposition and deep learning techniques, respectively, to obtain better prediction performance. Our proposed method achieves a decisive performance with the best AUC and AUPR values of 0.9702 and 0.8814, respectively, under LOOCV experiments. It is superior to other state-of-the-art SDLDA and NCPLDA methods in both AUC and AUPR evaluation metrics. It could be considered as a powerful tool to predict lncRNA-disease associations.**

**Keywords:** *lncRNA-disease associations prediction, weighted K nearest known neighbors, singular value decomposition, feature extraction, deep learning.*

## I. INTRODUCTION

Long non-coding RNAs (lncRNAs) are a class of different non-coding RNAs (ncRNAs) which have more than 200 nucleotides in length [1]. For a long time, they were considered to be noise and have no biological function. However, a diversity of researches has shown that lncRNAs have crucial roles in different biological processes including cell cycle control, transcription, translation, epigenetic regulation, splicing, differentiation, immune response and so forth in the human body [2–6] It is not surprise that the abnormal expressions, mutations and dysregulations of lncRNAs can cause numerous complex human diseases [5, 7]. Certainly, inferring lncRNA-disease associations can contribute to the awareness of human complex diseases' mechanisms at lncRNA levels as well as to discover the prognosis markers for diagnosis, treatment and prevention of complex diseases [5–7]. In fact, the number of validated lncRNA-disease associations is very limited in comparison with the number of possible associations between lncRNAs and diseases. Furthermore, discovering lncRNA-disease associations by performing the biological wet-experiments is costly, laboratory and time-consuming [5, 7]. Therefore, it is urgent and important to develop effective and meaningful computational methods to predict relationships between lncRNAs and diseases.

Over the past few years, various computational methods have been developed to predict potential lncRNA-disease associations. They can roughly be divided into the following categories: similarity network-based, multiple data sources-based, machine learning-based and multi-model integration methods [8–11]. *Firstly*, the similarity network-based methods mainly rely on known lncRNA-disease associations information to infer potential associations between lncRNAs and diseases. They are built on the basic assumption that similar diseases are likely to be associated with functionally similar lncRNAs [9]. For examples, *Xiao et al.* [12] presented a BPLLDA method to predict relationships between lncRNAs and diseases relied on simple paths with limited lengths in a heterogeneous network. *Xiao et al.* [13] proposed a weight matrix of lncRNA-disease associations prediction (WLDAP) method which based on known lncRNA-disease associations to calculate Gaussian interaction profile kernel for lncRNAs

and diseases. Then, they inferred potential lncRNA-disease associations by using a weighting model to obtain similarity scores between lncRNAs and diseases. *Chen et al.* [14] inferred latent lncRNA-disease associations by a label-propagation algorithm and random projection on a heterogeneous network constructed from known lncRNA-disease associations network, disease semantic similarity network, lncRNA functional similariry network and Gaussian interaction profile kernel for lncRNAs and diseases. Generally, this kind's models achieve high prediction performance but they require the calculation of multiple similarities.

*Secondly*, the multiple data sources-based methods normally integrate multiple types of biological information to reveal latent associations between lncRNAs and diseases [8, 9]. For instances, *Ding et al.* [15] presented a model called TPGLDA based on a tripartite graph constructed from lncRNA-disease associations network, disease-gene associations networks to detect lncRNA-disease associations. *Liu et al.* [16] built a lncRNA prioritization model which integrates gene-disease interactions, gene expression profiles, and lncRNA expression profiles to prioritize novel disease-associated lncRNAs. *Nguyen* and *Tran* [17] proposed an improved computational method for prediction of lncRNA-disease associations based on collaborative filtering and resource allocation which used known lncRNA-disease associations, verified lncRNA-miRNA interactions and validated miRNA-disease associations to predict lncRNA-disease associations. These methods require multiple types of biological information which are not easy to identify in the heterogeneous data sources.

*Thirdly*, the machine learning-based methods employed machine learning algorithms to predict lncRNA-disease associations relied on training and unlabeled samples, verified and unknown associations, respectively. Formerly, numerous meachine learning algorithms have been implemented to forecast lncRNA-disease associations such as Naïve Bayesian classifier [18, 19], Random forest [20, 21], Bagging support vector machine (SVM) [22] and matrix factorization based methods [10, 23]. Each type of machine learning algorithms has its own strengths and weaknesses. For example, the Naïve Bayesian classifier is a simple probabilistic classifier, however, it requires a naïve independence assumption that any feature of a class is independent of the other features [18]. The Random forest algorithm requires obtaining the reliable negative samples that are difficult to be collected. Therefore, randomly selecting negative samples may influence the prediction performance [20]. The SVM in [22] bases on the assumption that positive unlabeled learning problems have a particular structure that leads to instability of classifiers. Among machine learning-based methods to predict lncRNA-disease associations, the matrix factorization methods can be considered as the most popular. Normally, traditional matrix factorization based methods only extract the linear features of lncRNA-disease associations by projecting the constructed matrix into sub-matrices in lower dimension spaces. However, in fact, the lncRNA-disease associations are very complex and non-linear. Thus, traditional matrix factorization-based methods can not model such complex and non-linear associations. Recently, deep learning methods which consitute a basic part of machine learning, are applied to gather the complex and non-linear features of lncRNA-disease associations by assuming that the linear and non-linear features can reinforce each other to obtain high quality features to improve prediction performance [8]. Up to now, various deep learning methods have been developed to predict lncRNA-disease associations such as SDLDA [8], LDAPM [24], DBNLDA [25] and so on.

*Finally*, the multi model integration-based methods which integrated multi models into an unified one to reduce the issues of single model and increase the prediction performance [26]. For examples, the combination of matrix factorization technique and recommendation system ideas to predict lncRNA-disease associations was used in LDASR method developed *Guo et al.* [27]. *Shi et al.* [28] inferred lncRNA-disease associationsa by a VGAELDA hybrid method which integrates variational inference and graph autoencoders.

Although computational methods for lncRNA-disease association prediction have attained significant performance. However, the number of known lncRNA-disease associations between lncRNAs and diseases is very limited in comparison with the number of possible associations among them. Therefore, most of computational methods for lncRNA-disease association prediction have to deal with the problem of sparsity data.

As mentioned before, deep learning techniques can reinforce linear and non-linear features each other to improve prediction performance. Additionally, the similarity network based methods normally reach high accuracy and the problem of sparsity data needs to be solved. Hence, in this paper, we proposed a new method based on multiple features and deep learning to predict associations between lncRNAs and diseases. The proposed method based on lncRNA functional similarity, disease semantic similarity and known lncRNA-disease associations as input. Then, a weighted $K$-nearest known neighbors algorithm is employed as a pre-processing step to solve the problem of sparsity data. Next, singular value decomposition and deep learning techniques are used to extract linear and non-linear features of lncRNAs and diseases. The extracted linear and non-linear features are combined into a new vector before a final fully connected layer is used to forecast final prediction. The results showed that our proposed method

achieves a decisive performance with the best AUC and AUPR values of 0.9702 and 0.8814, respectively, under LOOCV experiments. It is superior to other state-of-the-art SDLDA [8] and NCPLDA [9] methods in both AUC and AUPR evaluation metrics. In case studies of Breast cancer and Hepatocellular Carcinoma, it recalled 8 and 7 known associations and it inferred 2 and 3 new associations out of top 10 predicted associations, respectively. Fortunately, all of the new predicted associations have been verified in biomedical literature. Therefore, our new proposed method could be considered as a forceful tool to predict lncRNA-disease associations.

## II. MATERIALS AND METHOD

### 1. Materials

#### *a) Human lncRNA-disease associations*

In this paper, the data sets used in experiments were collected from the LncRNADisease database including three versions (June-2012, January-2014 and June-2015 versions). After removing the associations with abnormal lncRNA or disease names as well as merging all duplicated records, we obtained three data sets as follows. Firstly, the June-2012 version data set (marked as DS1) contained 276 lncRNA-disease associations between 112 lncRNAs and 150 diseases. Secondly, the January-2014 version data set (marked as DS2) included 319 associations between 131 lncRNAs and 169 diseases. And finally, the June-2015 version data set (marked as DS3) consisted of 621 associations between 285 lncRNAs and 226 diseases. The association density of three data sets are 1.643%, 1.441% and 0.964% for DS1, DS2 and DS3, respectively. All of these data sets were already used in the work of *Li et al.* [9].

For accessibility, we used an adjacency matrix $A^{LD} \in R^{m \times n}$ to represent lncRNA-disease associations where m is the number of rows in the matrix $A^{LD}$ or the number of lncRNAs, n means the number of columns in the matrix $A^{LD}$ or the number of diseases, and $A^{LD}(i, j) = 1$ if the association between lncRNA $l_i$ and disease $d_j$ has been validated and $A^{LD}(i, j) = 0$ for other cases.

#### *b) Disease semantic similarity*

In this paper, disease semantic similarity was estimated according to Wang et al.'s method [29]. Concretely, we collected the relationships of different diseases based on the hierarchical directed acrylic graphs (DAGs) by downloading MeSH descriptors from the National Library of Medicine (http://www.ncbi.nlm.nih.gov/) to measure the similarity among diseases. Specifically, for a disease d, its directed acrylic graph is given by $DAG(d) = (d, A_d, E_d)$, where $A_d$ reflects the set of the disease d's ancestors and d itself, and $E_d$ means the set of edges which point to child nodes from parent nodes in the MeSH tree. Hence, the semantic contribution of disease k to disease d is as:

$$D_d(k) = \begin{cases} k & if\, k = d \\ max\{\Delta \times D_d(k') | k' \in \text{children of k}\} & if\, k \neq d \end{cases} \quad (1)$$

Where $\Delta$ indicates a predefined semantic contribution factor with values in range (0, 1). Similar to *Wang et al.* [29], in this paper, we set $\Delta$ equal to 0.5. We computed the semantic similarity among diseases relied on the assumption that two diseases having larger parts in their DAGs tend to have higher semantic similarity as:

$$DSS(d_i, d_j) = \frac{\sum_{k \in A_{d_i} \cap A_{d_j}} (D_{d_i}(k) + D_{d_j}(k))}{\sum_{k \in A_{d_i}} D_{d_i}(k) + \sum_{k \in A_{d_j}} D_{d_j}(k)} \quad (2)$$

#### *c) LncRNA functional similarity*

Based on the assumption that functionally similar lncRNAs are often associated with similar diseases, in this work, we calculated functional similarity between two lncRNAs by measuring the semantic similarity of two disease sets which are linked to these two lncRNAs [9, 30]. Explicitly, we assumed that lncRNA $l_i$ and lncRNA $l_j$ were associated with $n_1$ and $n_2$ diseases, respectively, then the similarity between lncRNA $l_i$ and lncRNA $l_j$ can be computed as in the following equations:

$$LFS(l_i, j_j) = \frac{\sum_{d \in D_{l_j}} SS(d, D(_{l_i})) + \sum_{d \in D_{l_i}} SS(d, D(_{l_j}))}{n_1 + n_2} \quad (3)$$

$$SS(d_k, D(l_i)) = max_{d \in D(l_i)}(DSS(d_k, d)) \quad (4)$$

where $LFS$ is the lncRNA functional similarity matrix and $D(l_i)$ means the disease set associated with lncRNA $l_i$.

It is important to remember that all of the known lncRNA-disease associations matrix $A^{LD}$, disease semantic similarity matrix $DSS$ and lncRNA functional similarity $LFS$ matrix are sparse. Therefore, in our proposed method, we have to solve this problem.

### 2. Method

#### *a) Method overview*

In overview, the proposed method contains following stages. At the first stage, we take disease semantic similarity ($DSS$) information, lncRNA functional similarity ($LFS$) information and known lncRNA-disease associations $A^{LD}$ as inputs. And we applied a weighted $K$-nearest known neighbors ($WKNKN$) algorithm as a pre-processing step to solve the sparsity data problem to obtain a new lncRNA-disease association matrix $A^{LD-new}$. At second stage, a singular value decomposition ($SVD$) technique is used to decompose matrix $A^{LD-new}$ into three matrices $U$, $\sum$, and

$V^T$ to extract the linear features of lncRNAs and diseases. The rows of $U$ indicate the linear features of lncRNAs where the columns of $V^T$ reflect the linear features of diseases. During the third stage, we extract non-linear features of lncRNAs and diseases with deep learning technique. In deep learning part, each row of $A^{LD-new}$ is considered as a lncRNA raw feature vector while each column of $A^{LD-new}$ is considered as a disease raw feature vector. For each interaction in $A^{LD-new}$ matrix, its corresponding raw feature vectors are fed into two fully connected layers with non-linear activation function to extract non-linear features. At fourth stage, the linear and non-linear features are concatenated to $Lnc_{new-fea-vec}$ and $Dis_{new-fea-vec}$ for lncRNA and disease, respectively. Next, a Hadamard product operation is employed to fuse two new feature vectors. And finally, the final prediction is obtained by applying a sigmoid activation function. The workflow of the proposed method is illustrated in Figure 1.

*b) Weighted K-nearest known neighbors algorithm*

As mentioned before, all of three matrices $A^{LD}$, $DSS$ and $LFS$ are sparse. To solve this problem, in this paper, we employed a weighted $K$-nearest known neighbors algorithm as a pre-processing step to reduce the impact of sparsity data problem and to eliminate the unknown associations in lncRNA-disease association set. It relied on the known neighbors' information of nodes by seeing the fact that many of the non-interacting lncRNA-disease pairs in $A^{LD}$ which are unknown cases but they could potentially be true associations. Specifically, the $WKNKN$ algorithm contains following steps. *Firstly*, for each lncRNA $l_i$, we took its functional similarities with $K$ known lncRNAs which are nearest to $l_i$ and their corresponding interaction profiles to estimate the interaction likelihood profile for lncRNA $l_i$. *Secondly*, for each disease $d_j$, we chose the semantic similarities with $K$ known diseases which are nearest to $d_j$ and their corresponding interaction profiles to evaluate the interaction likelihood profile for disease $d_j$. *Thirdly*, if $A^{LD}_{ij} = 0$, we changed it by averaging the two interaction likelihood profiles. And *finally*, we used a threshold to classify the interaction likelihood profiles into two classes labeled as 0 and 1. The Algorithm 1 contains the pseudocode that describes the above steps in detail in which $t$ is a decay term where $t \leq 1$, $\theta$ means a threshold, normally $\theta$ is set to 0.5, and $KNN()$ returns the $K$-nearest known neighbors in descending order based on their similarities to lncRNA $l_i$ or disease $d_j$.

*c) Obtaining linear features by decomposing $A^{LD-new}$ with singular value decomposition technique*

Singular value decomposition ($SVD$) is one of the most polular matrix factorization methods in recommendation systems [8]. It is used in our work to gather linear features

**Algorithm 1:** WKNKN algorithm.

**Inputs**: Matrices $A^{LD} \in R^{m \times n}$, $LFS \in R^{m \times m}$ and $DSS \in R^{n \times n}$, neighborhood size $K$, decay term $t$, threshold $\theta$.
**Outputs**: new adjacency association matrix $A^{LD-new}$.
**begin**
  $Al = Ad = 0$ # initialize two temporary mactrices ;
  # for lncRNAs
  **for** $l = 1$ to $m$ **do**
    lnn=KNN(l, LFS, K);
    **for** $i = 1$ to $K$ **do**
      $w_i = t^{i-1} LFS(l, lnn_i)$;
    **end**
    $Z_l = \sum_{i=1}^{K} LFS(l, lnn_i)$;
    $Al(l) = \frac{1}{Z_l} \sum_{i=1}^{K} w_i A^{LD}(lnn_i)$;
  **end**
  # for diseases
  **for** $d = 1$ to $n$ **do**
    dnn=KNN(d, DSS, K);
    **for** $i = 1$ to $K$ **do**
      $w_i = t^{i-1} DSS(d, dnn_i)$;
    **end**
    $Z_d = \sum_{i=1}^{K} DSS(d, dnn_i)$;
    $Ad(d) = \frac{1}{Z_d} \sum_{i=1}^{K} w_i A^{LD}(dnn_i)$;
  **end**
  $Ald = \frac{Al+Ad}{2}$;
  $A^{LD-new} = max(A^{LD}, Ald)$;
  # classify $A^{LD-new}$ by threshold $\theta$.
  **for** $i = 1$ to $m$ **do**
    **for** $j = 1$ to $n$ **do**
      **if** $A^{LD-new} \geq \theta$ **then**
        $A^{LD-new}[i, j] = 1$;
      **else**
        $A^{LD-new}[i, j] = 0$;
      **end**
    **end**
  **end**
  **return** $A^{LD-new}$;
**end**

of lncRNAs and diseases. The details of the process to obtain linear features of lncRNAs and diseases are described as follows. Let $A^{LD-new}$ be the new association matrix of lncRNAs and diseases, the $SVD$ of $A^{LD-new}$ matrix is a factorization of three matrices $U$, $\sum$ and $V^T$ as:

$$A^{LD-new} \approx U \sum V^T \tag{5}$$

where $U \in R^{m \times k}$ is a real matrix, $\sum \in R^{k \times k}$ is a diagonal matrix with non-negative square roots of the eigenvalues of the product $R^T R$ on the diagonal and $V^T \in R^{k \times n}$ is a real matrix. The $U_i$ represents a linear feature vector of lncRNA $l_i$ and the $V_j^T$ is a linear feature vector of disease $d_j$. Figure 2 illustrates the process of $SVD$ to obtain linear features of lncRNAs and diseases in our method.

*d) Obtaining non-linear features with deep learning*

As can be known, deep learning techniques are the basic and powerful techniques of machine learning. With multiple

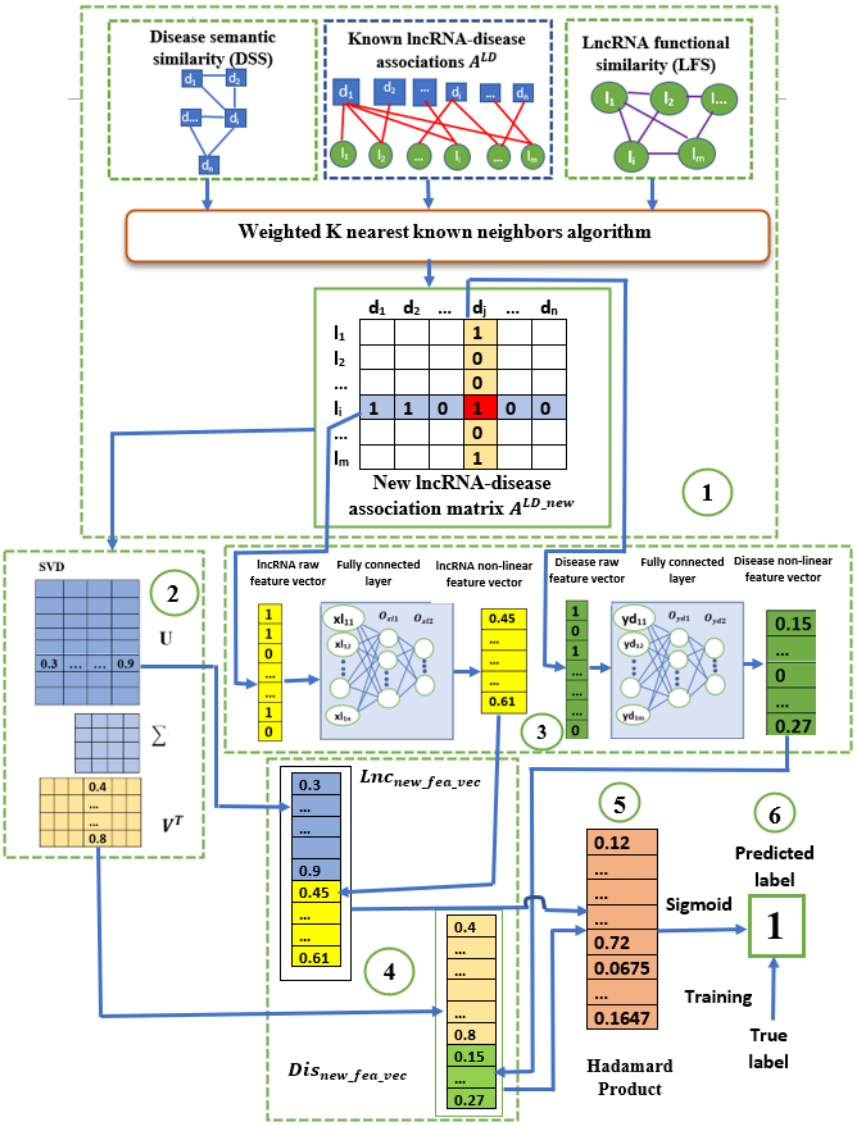


Figure 1. The workflow of the proposed method.

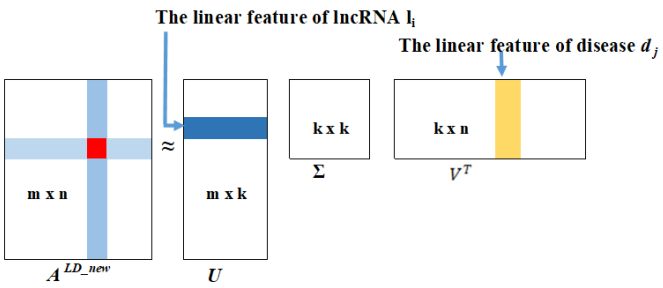


Figure 2. The process of SVD to obtain linear features of lncRNAs and diseases.

levels of representation, deep learning techniques can learn non-linear, complex and useful features. Therefore, in this work, we used a deep learning technique to obtain non-linear features of lncRNAs and diseases. Specifically, we considered the new association matrix $A^{LD-new}$ as input of deep learning process. Similar to the work of *Zeng et al.* [8], we considered each row of the $A^{LD-new}$ matrix as raw feature vector of lncRNA $l_i$ while each column of $A^{LD-new}$ matrix as raw feature vector of disease $d_j$. For each intersection of the $A^{LD-new}$ matrix, its corresponding raw feature vectors are fed into two fully connected layers with a non-linear activation function to gather non-linear features of lncRNA and disease. Formally, the lncRNA and disease raw feature vectors are denoted as $xl$ and $yd$, respectively. They are fed into the first fully connected layer to output $O_{xl_1}$ and $O_{yd_1}$ as:

$$O_{xl_1} = \delta(W_{xl_1} xl + b_{xl_1}) \quad (6)$$

$$O_{yd_1} = \delta(W_{yd_1} yd + b_{yd_1}) \quad (7)$$

where $\delta$ is the non-linear activation function, $W_{xl_1}$ and $W_{yd_1}$ are the first fully connected layer's weight matrices, $b_{xl_1}$ and $b_{yd_1}$ are bias terms of the first fully connected layer. Next, $O_{xl_1}$ and $O_{yd_1}$ are fed into the second fully connected layer to output the final non-linear features $O_{xl_2}$and $O_{yd_2}$as in the following equations:

$$O_{xl_2} = \delta(W_{xl_2}O_{xl_1} + b_{xl_2}) \tag{8}$$

$$O_{yd_2} = \delta(W_{yd_2}O_{yd_1} + b_{yd_2}) \tag{9}$$

where $W_{xl_2}$ and $W_{yd_2}$ are the second fully connected layer's weight matrices, $b_{xl_2}$ and $b_{yd_2}$ are bias terms of the second fully connected layer.

In each fully connected layer, we applied the Rectified Linear Unit ($ReLU$) activation function to obtain non-linear features as in the follow equation:

$$ReLU(x) = max(0,\ x) \tag{10}$$

*e) Combining linear features and non-linear features*

After obtaining linear and non-linear feature vectors by $SVD$ and deep learning technique, respectively, we firstly concatenated them into $Lnc_{new-fea-vec}$ and $Dis_{new-fea-vec}$ for lncRNAs and diseases, respectively.

$$Lnc_{new-fea-vec} = \begin{bmatrix} U_i \\ O_{xl_2} \end{bmatrix} \tag{11}$$

$$Dis_{new-fea-vec} = \begin{bmatrix} V_j^T \\ O_{yd_2} \end{bmatrix} \tag{12}$$

where [] indicates concatenation operation.
Then, we applied a Hadamard product operation to integrate two vectors $Lnc_{new-fea-vec}$ and $Dis_{new-fea-vec}$ into a new integrated vector:

$$Vec_{new-integrated} = HadamardProduct(Lnc_{new-fea-vec}, Dis_{new-fea-vec}) \tag{13}$$

*f) Obtaining final prediction*

Finally, the new integrated vector $Vec_{new-integrated}$ is fed into a fully connected layer with a sigmoid activation function to obtain final prediction.

$$Sigmoid(x) = \frac{1}{1 + \exp(-x)} \tag{14}$$

$$A^{LD-final}[i, j] = Sigmoid(Vec_{new-integrated}) \tag{15}$$

We used a binary cross-entropy function in our model as the loss function as follows:

$$Loss = \sum [A^{LD-new} \log(A^{LD-final}[i, j]) + (1 - A^{LD-new}) \log(1 - A^{LD-final}[i, j])] + \gamma(|\partial^2|) \tag{16}$$

where $\partial$ is the weight vector, $\gamma$ is a weight to balance between the experimental risk and regularized term.

## III. RESULTS

### 1. Performance evaluation

To evaluate the proposed method's prediction performance, in this paper, we already performed Leave-one-out-cross-validation ($LOOCV$) experiments on all three data sets as described in Materials section. In $LOOCV$ experiments, each entry value equal to 1 in $A^{LD}$ matrix is taken as test sample and the remaining associations are considered as training samples. After implementing the proposed model, we gathered $A^{LD-final}$ matrix. We repeatedly performed the process and then used the obtained results to calculate true positive rate ($TPR$), false positive rate ($FPR$), Precision and Recall as in the followings:

$$TPR = \frac{TP}{TP + FN} \tag{17}$$

$$FPR = \frac{FP}{FP + TN} \tag{18}$$

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

where $TP$ indicates the number of true positive samples, $FN$ means the number of false negative samples, $FP$ is the number of false positive samples and $TN$ reflects the number of true negative samples.

After that, we figured out Receiver Operating Characteristic (ROC) curve [31] and Precision-Recall (PR) curve [32] of proposed method on each data set. Parallely, we computed $AUC$ (Area under ROC curve) and $AUPR$ (Area under Precision-Recall curve) values of proposed method on each data set. Figure 3 illustrates ROC curves and $AUC$ values and Figure 4 shows PR curves and $AUPR$ values of proposed method on three data sets when the value of $K$ in $WKNKN$ algorithm is set to 8. As can be seen, $AUC$ values of proposed method on three data sets are similar and the best $AUPR$ value is achieved on the DS3 data set. Therefore, we chose the $AUC$ and $AUPR$ values of proposed method on DS3 to compare with the performance of other methods in the next section.

### 2. Performance in comparison with other related methods

To emphasize the performance of our proposed method in comparison with other previous related methods, we compared our proposed method prediction performances, on the same data set DS3, with SDLDA and NCPLDA methods [8, 9]. Figure 5 shows the ROC curves and $AUC$ values of the related methods and our proposed method on the same data set DS3. Figure 6 represents the PR curves and $AUPR$ values of the related methods and our

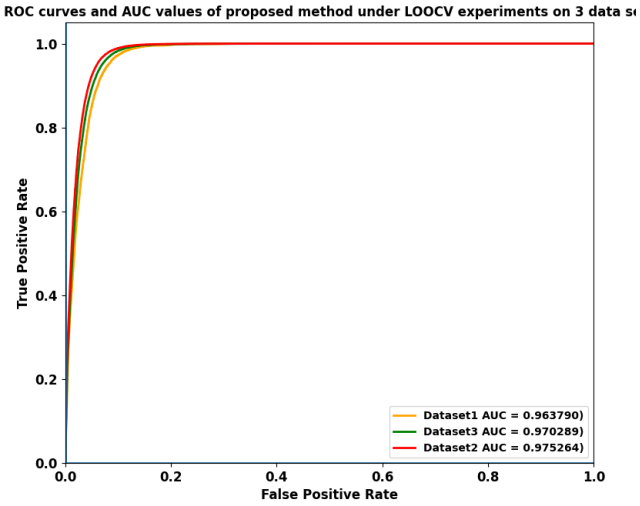


Figure 3. The ROC curves and AUC values of proposed method on three data sets when K in WKNKN is set to 8.

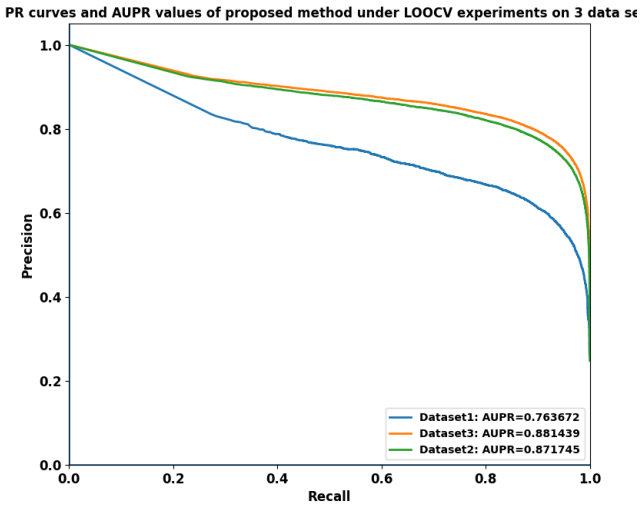


Figure 4. The PR curves and AUPR values of proposed method on three data sets when K in WKNKN is set to 8.

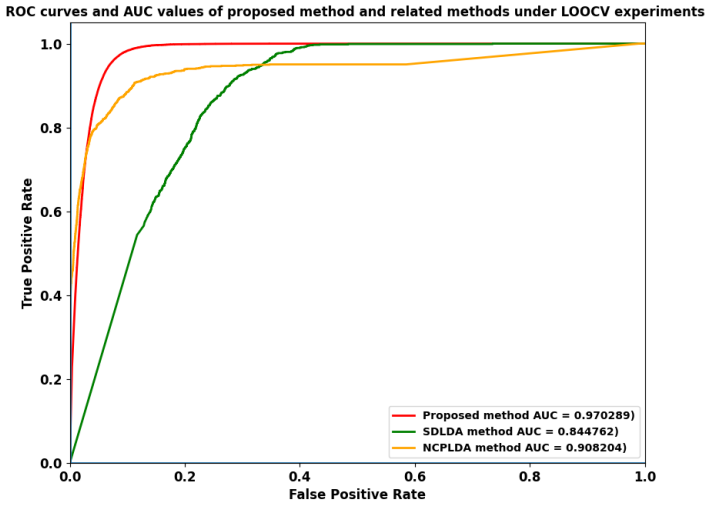


Figure 5. ROC curves and AUC values of the proposed method and related methods under LOOCV experiments.

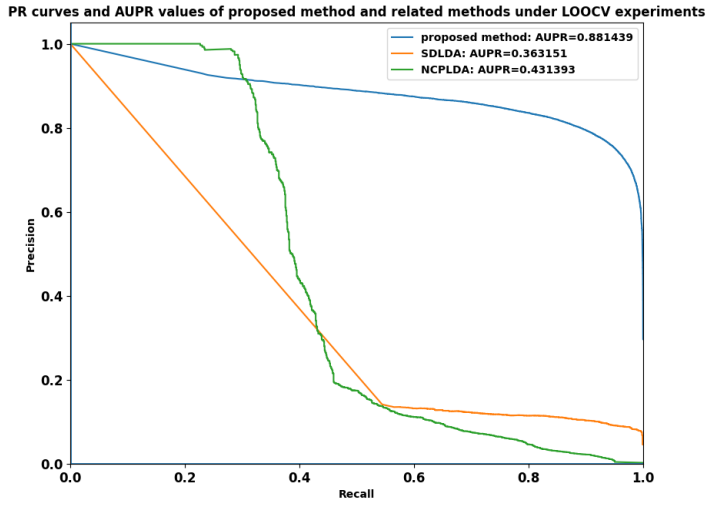


Figure 6. PR curves and AUPR values of the proposed method and related methods under LOOCV experiments.

proposed method on the same data set DS3.

As can be seen in Figure 5 and Figure 6, our new proposed method is superior to previous related methods in both AUC and AUPR terms. The highest AUPR value indicates that our proposed method can recall more known associations by reducing the impact of sparsity data problem.

## 3. Effects of parameters

### *a) Parameters of WKNKN algorithm*

Considering the fact that there are some unknown lncRNA-disease associations in the matrix $A^{LD}$. We used the *WKNKN* algorithm as a pre-processing step to eliminate unknown values in lncRNA-disease association set relied on their known neighbors. In this algorithm, the $K$ parameter means the number of nearest known neighbors, $t$ reflects a decay term where $t \leq 1$, $\theta$ is a threshold. In this study, we mainly targeted on the impact of number

of nearest known neighbors to scale down the effects of sparsity data problem. It is true that when the more nearest known neighbors were selected, the more associations between lncRNAs and diseases would be added into the heterogeneous network and then the impact of sparity data problem would be reduced. However, the imbalanced data problem would again appears when the number of added associations becomes too big. Therefore, in our experiments, we repeatedly changed the value of $K$ and $t$ to choose the optimal result. And it showed that $AUC$ and $AUPR$ of our proposed method reached the best values when $K = 8$ and $t = 0.7$ while $\theta$ is fixed to 0.5 as be mentioned before.

*b) Implementation details and hyper-parameters*

In this paper, the $SVD$ is executed with Scipy library, the lncRNA and disease linear features are represented by two 64-dimensional vectors. The deep learning framework use Tensor flow [33]. Non-linear features of lncRNAs and diseases are extracted by ReLU activation function.

For hyper-parameters, after repeatedly performing the experiments, the optimal values of parameters are as follows. The first and second fully connected layers used 48 and 32 neurons, respectively. The dropout rate and regularization parameter $\gamma$ are set to 0.05 and 0.003, respectively. The optimizer used the adaptive moment estimation (Adam). The initinal learning rate is set to 0.001. Finally, the number of neurons in the last fully connected layer is set to 48.

## 4. Case studies

To encourage our method's prediction reliability by evaluating $AUC$ and $AUPR$ values under $LOOCV$ experiments, we also tested some case studies to observe the most predicted lncRNAs for any given disease. In the following sections, we showed top 10 predicted lncRNAs for case studies' results of Breast Cancer and Hepatocellular Carcinoma, respectively.

*a) Breast Cancer's case study*

Breast Cancer is also known as Breast Neoplasms. It is a malignant tumor that has a high mortality rate and mostly occurs in women [34]. Over the past few years, different lncRNAs have been reported to be associated with Breast Cancer. For examples, lncRNA H19 has been indicated to be associated with poor prognosis in breast cancer patients and promotes cancer stemness [35]. High expression of lncRNA MALAT1 is related to breast cancer relapse and may play a role in tumor progression [36]. The case study of Breast Cancer on data set DS3 showed that our proposed method recalled 8 known associations and it inferred 2 new associations out of top 10 predicted Breast Cancer-related lncRNAs. The Table I showed top 10 predicted Breast cancer-related lncRNAs and all predicted associations have been known or verified by biomedical literature.

TABLE I
TOP 10 PREDICTED BREAST CANCER-RELATED LNCRNAS

| LncRNA | Rank | Known before | Evidences (PMID) |
|---|---|---|---|
| CDKN2B-AS1 | 1 | 1 | Known association |
| GAS5 | 2 | 1 | Known association |
| H19 | 3 | 1 | Known association |
| HOTAIR | 4 | 1 | Known association |
| PVT1 | 5 | 1 | Known association |
| MALAT1 | 6 | 1 | Known association |
| MEG3 | 7 | 1 | Known association |
| MIAT | 8 | 0 | 29100300 |
| NEAT1 | 9 | 0 | 32273717 |
| XIST | 10 | 1 | Known association |

TABLE II
TOP 10 PREDICTED HEPATOCELLULAR CARCINOMA-RELATED LNCRNAS

| LncRNA | Rank | Known before | Evidences (PMID) |
|---|---|---|---|
| HOTAIR | 1 | 1 | Known association |
| MALAT1 | 2 | 1 | Known association |
| H19 | 3 | 1 | Known association |
| CDKN2B-AS1 | 4 | 0 | 32801906 |
| PMEG3 | 5 | 1 | Known association |
| UCA1 | 6 | 1 | Known association |
| CCND1 | 7 | 0 | 32443384 |
| PCAT29 | 8 | 1 | Known association |
| GAS5 | 9 | 0 | 25120813; 26163879 |
| AIR | 10 | 1 | Known association |

*b) Hepatocellular Carcinoma's case study*

Hepatocellular carcinoma ($HCC$) is the most common primary liver malignancy and it is a leading cause of cancer-related death over worldwide [37]. Recently, many lncRNAs have been indicated to be related to hepatocellular carcinoma. For instances, lncRNA MALAT1 promotes hepatocellular carcinoma development by SRSF1 up-regulation and mTOR activation [38]. Upregulated lncRNA-UCA1 promotes to progression of hepatocellular carcinoma through inhibition of miR-216b and activation of FGFR1/ERK signaling pathway [39]. The case study of Hepatocellular carcinoma indicated that our proposed method recalled 7 known associations and derived 3 new associations out of top 10 predicted Hepatocellular carcinoma-related lncRNAs. Table II demonstrated top 10 predicted Hepatocellular carcinoma-related lncRNAs and all predicted associations have been known or verified by biomedical literature.

## IV. CONCLUSION AND DISCUSSIONS

Various lncRNAs have been shown to play crucial roles in different biological processes including cell cycle control, transcription, translation, epigenetic regulation, splicing, differentiation, immune response and so forth in the

human body. Discovering associations between lncRNAs and diseases promotes the awareness of human complex diseases' mechanisms at molecular levels and support the diagnosis, treatment and prevention of complex diseases [5–7] However, the process of finding and verifying lncRNA-disease associations is costly, laboratory and time-consuming. Therefore, developing a forceful and meaningful computational method to forecast potential lncRNA-disease associations is become urgent in some recent years. In this paper, we proposed a new method to predict lncRNA-disease associations using multiple features and deep learning. It achieves a decisive performance with the best *AUC* and *AUPR* values of 0.9702 and 0.8814, respectively, under LOOCV experiments on the data set DS3 as mentioned before. It outperforms other state-of-the-art methods in both *AUC* and *AUPR* evaluation metrics.

The desirable prediction performance of our proposed method is contributed by the following factors. *Firstly*, by using multiple features, we already take into account the high accuracy advantage of the similarity network-based methods. *Secondly*, by using a *WKNKN* algorithm as a pre-processing step, we already reduced the impact of sparsity data problem. And *finally*, the combination of linear features and non-linear features extracted by singular value decomposition technique and deep learning, respectively, can reinforce each other to better predict potential lncRNA-disease associations [8]. However, there are too many parameters in deep learning model. Therefore, it is difficult to tune all hyper-parameters and try to do all possible cases of experiments.

## ACKNOWLEDGMENTS

This research was supported by the Vietnam Ministry of Education and Training, Project No. B2022-SPH-04.

## REFERENCES

[1] E. S. Lander, "Ribosome profiling provides evidence that large non-coding RNAs do not encode proteins," *Cell*, vol. 154, no. 1, pp. 240–251, 2014, doi: 10.1016/j.cell.2013.06.009.Ribosome.

[2] K. C. Wang and H. Y. Chang, "Molecular mechanisms of long noncoding RNAs," *Molecular*, vol. 43, no. 6, pp. 904–914, 2011, doi: 10.1016/j.molcel.2011.08.018.

[3] M. Guttman and J. L. Rinn, "Modular regulatory principles of large non-coding RNAs," *Nature*, vol. 482, no. 7385, pp. 339–346, 2012, doi: 10.1038/nature10887.

[4] V. Mohanty, Y. Gökmen-Polar, S. Badve, and S. C. Janga, "Role of lncRNAs in health and disease-size and shape matter," *Brief. Funct. Genomics*, vol. 14, no. 2, pp. 115–129, 2015, doi: 10.1093/bfgp/elu034.

[5] X. Chen, C. C. Yan, X. Zhang, and Z. H. You, "Long non-coding RNAs and complex diseases: From experimental results to computational models," *Brief. Bioinform.*, vol. 18, no. 4, pp. 558–576, 2017, doi: 10.1093/bib/bbw060.

[6] X. Li, J. Xu, Y. Xiao, and S. Ning, Non-coding RNAs in Complex Diseases. Springer Singapore, 2018.

[7] X. Lei et al., "A comprehensive survey on computational methods of non-coding RNA and disease association prediction," *Brief. Bioinform.*, vol. 00, no. August, pp. 1–31, 2020, doi: 10.1093/bib/bbaa350.

[8] M. Zeng et al., "SDLLDA: lncRNA-disease association prediction based on singular value decomposition and deep learning," *Methods*, vol. 179, no. February, pp. 73–80, 2020, doi: 10.1016/j.ymeth.2020.05.002.

[9] G. Li et al., "Prediction of LncRNA-Disease Associations Based on Network Consistency Projection," *IEEE Access*, vol. 7, pp. 58849–58856, 2019, doi: 10.1109/ACCESS.2019.2914533.

[10] Z. Xuan, J. Li, J. Yu, X. Feng, B. Zhao, and L. Wang, "A probabilistic matrix factorization method for identifying lncRNA-disease associations," *Genes (Basel).*, vol. 10, no. 2, 2019, doi: 10.3390/genes10020126.

[11] G. Xie, Z. Huang, Z. Liu, Z. Lin, and L. Ma, "NCPHLDA: A novel method for human lncRNA-disease association prediction based on network consistency projection," *Mol. Omi.*, vol. 15, no. 6, pp. 442–450, 2019, doi: 10.1039/c9mo00092e.

[12] X. Xiao et al., "BPLLDA: Predicting lncRNA-disease associations based on simple paths with limited lengths in a heterogeneous network," *Front. Genet.*, vol. 9, no. OCT, pp. 1–11, 2018, doi: 10.3389/fgene.2018.00411.

[13] G. Xie, L. Wu, Z. Lin, and J. Cui, "WLDAP: A computational model of weighted lncRNA-disease associations prediction," *Phys. A Stat. Mech. its Appl.*, vol. 558, p. 124765, 2020, doi: 10.1016/j.physa.2020.124765.

[14] M. Chen, Y. Deng, A. Li, and Y. Tan, "Inferring Latent Disease-lncRNA Associations by Label-Propagation Algorithm and Random Projection on a Heterogeneous Network," *Front. Genet.*, vol. 13, no. February, pp. 1–11, 2022, doi: 10.3389/fgene.2022.798632.

[15] L. Ding, M. Wang, D. Sun, and A. Li, "TPGLDA: Novel prediction of associations between lncRNAs and diseases via lncRNA-disease-gene tripartite graph," *Sci. Rep.*, vol. 8, no. 1, pp. 1–11, 2018, doi: 10.1038/s41598-018-19357-3.

[16] M. X. Liu, X. Chen, G. Chen, Q. H. Cui, and G. Y. Yan, "A computational framework to infer human disease-associated long noncoding RNAs," *PLoS One*, vol. 9, no. 1, 2014, doi: 10.1371/journal.pone.0084408.

[17] V. T. Nguyen and D. H. Tran, "An improved computational method for prediction of lncRNA-disease associations based on collaborative filtering and resource allocation," *2021 13th Int. Conf. Knowl. Syst. Eng.*, pp. 1–6, 2021, doi: 10.1109/kse53942.2021.9648632.

[18] J. Yu, P. Ping, L. Wang, L. Kuang, X. Li, and Z. Wu, "A novel probability model for lncRNA–disease association prediction based on the naïve bayesian classifier," *Genes (Basel).*, vol. 9, no. 7, 2018, doi: 10.3390/genes9070345.

[19] J. Yu, Z. Xuan, X. Feng, Q. Zou, and L. Wang, "A novel collaborative filtering model for LncRNA-disease association prediction based on the Naïve Bayesian classifier," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–13, 2019, doi: 10.1186/s12859-019-2985-0.

[20] D. Yao, X. Zhan, X. Zhan, C. K. Kwoh, P. Li, and J. Wang, "A random forest based computational model for predicting novel lncRNA-disease associations," *BMC Bioinformatics*, vol. 21, no. 1, pp. 1–18, 2020, doi: 10.1186/s12859-020-3458-1.

[21] R. Zhu, Y. Wang, J. X. Liu, and L. Y. Dai, "IPCARF: improving lncRNA-disease association prediction using incremental principal component analysis feature selection and a random forest classifier," *BMC Bioinformatics*, vol. 22, no. 1. 2021, doi: 10.1186/s12859-021-04104-9.

[22] W.Lan et al., "LDAP: a web server for lncRNA-disease association prediction," *Bioinformatics*, vol.33, pp458-460,

2017, doi: 10.1093/bioinformatics/btw639.
[23] G. Fu, J. Wang, C. Domeniconi, and G. Yu, "Matrix factorization-based data fusion for the prediction of lncRNA-disease associations," *Bioinformatics*, vol. 34, no. 9, pp. 1529–1537, 2018, doi: 10.1093/bioinformatics/btx794.
[24] N. Fraidouni and G. Zaruba, "A Matrix Completion Approach for Predicting lncRNA-disease association," *Int'l Conf. Bioinforma. Comput. Biol.*, pp. 1–6, 2019.
[25] M. Madhavan and G. Gopakumar, "DBNLDA: Deep Belief Network based representation learning for lncRNA-disease association prediction," *Appl. Intell.*, vol. 52, no. 5, pp. 5342–5352, 2022, doi: 10.1007/s10489-021-02675-x.
[26] C. Yan et al., "Computational Methods and Applications for Identifying Disease-Associated lncRNAs as Potential Biomarkers and Therapeutic Targets," *Mol. Ther. - Nucleic Acids*, vol. 21, pp. 156–171, 2020, doi: 10.1016/j.omtn.2020.05.018.
[27] Z. H. Guo, Z. H. You, Y. Bin Wang, H. C. Yi, and Z. H. Chen, "A Learning-Based Method for LncRNA-Disease Association Identification Combing Similarity Information and Rotation Forest," *iScience*, vol. 19, pp. 786–795, 2019, doi: 10.1016/j.isci.2019.08.030.
[28] Z. Shi, H. Zhang, C. Jin, X. Quan, and Y. Yin, "A representation learning model based on variational inference and graph autoencoder for predicting lncRNA-disease associations," *BMC Bioinformatics*, vol. 22, no. 1, pp. 1–20, 2021, doi: 10.1186/s12859-021-04073-z.
[29] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, 2010, doi: 10.1093/bioinformatics/btq241.
[30] J. Sun et al., "Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network," *Mol. Biosyst.*, vol. 10, no. 8, pp. 2074–2081, 2014, doi: 10.1039/c3mb70608g.
[31] H.-T. K., "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," *Casp. J Intern Med*, vol. 4(2), pp. 627–635, 2013.
[32] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One*, vol. 10, no. 3, p. e0118432., 2015, doi: 10.1371/journal.pone.0118432.
[33] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," *12th USENIX Symp. Oper. Syst. Des. Implement.*, pp. 265–283, 2016, doi: 10.1007/978-1-4842-6418-8-2.
[34] H. Jin et al., "lncRNA and breast cancer: Progress from identifying mechanisms to challenges and opportunities of clinical treatment," *Mol. Ther. - Nucleic Acids*, vol. 25, no. September, pp. 613–637, 2021, doi: 10.1016/j.omtn.2021.08.005.
[35] H. Shima et al., "Lnc RNA H19 is associated with poor prognosis in breast cancer patients and promotes cancer stemness," *Breast Cancer Res. Treat.*, vol. 170, no. 3, pp. 507–516, 2018, doi: 10.1007/s10549-018-4793-z.
[36] Z. Wang et al., "High expression of long non-coding RNA MALAT1 in breast cancer is associated with poor relapse-free survival," *Breast Cancer Res Trea*, vol. 171, no. 2, pp. 261–271, 2018, doi: 10.1007/s10549-018-4839-2.High.
[37] X. Xu et al., "The role of MicroRNAs in hepatocellular carcinoma," *J. Cancer*, vol. 9, no. 19, pp. 3557–3569, 2018, doi: 10.7150/jca.26350.
[38] P. Malakar et al., "Long Noncoding RNA MALAT1 Promotes Hepatocellular Carcinoma Development by SRSF1 Upregulation and mTOR Activation," *Cancer Res*, vol. 77, no. 5, pp. 1155–1167, 2018, doi: 10.1158/0008-5472.CAN-16-1508.Long.
[39] F. Wang et al., "Upregulated lncRNA-UCA1 contributes to progression of hepatocellular carcinoma through inhibition of miR-216b and activation of FGFR1/ERK signaling pathway," *Oncotarget*, vol. 6, no. 10, pp. 7899–7917, 2015, doi: 10.18632/oncotarget.3219.

**Van Tinh Nguyen** received B.S. and M.S degree in Information Technology from Hanoi University of Science and Technology, Hanoi, Vietnam in 2001 and 2006, respectively. He is a lecturer in the Falculty of Information technology, Hanoi University of Industry. Curently, he is a Ph.D. student of Hanoi National University of Education. His research includes bioinformatics, data mining, data representation and machine learning.
Email: tinhkh@gmail.com, nguyenvantinh_cntt@haui.edu.vn

**Dang Hung Tran** received the B.S. from Faculty of Mathematic and Informatics, Hanoi National University of Education in 2001, M.S. in Computer Science from University of Engineering and Technology, VNU in 2006, and Ph.D. degree in Knowledge Science from Japan Advance Institute of Science and Technology (JAIST) in 2009. He is currently an associate professor and dean of Faculty of Information Technology, Hanoi National University of Education (HNUE). His research includes bioinformatics, data mining and machine learning.
Email: hungtd@hnue.edu.vn