IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 26, 2020, accepted January 19, 2021, date of publication January 22, 2021, date of current version February 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3053839

Principal Component Regression Analysis for lncRNA-Disease Association Prediction Based on Pathological Stage Data

BO WANG[1,4] AND JING ZHANG[1,2,3]

[1]College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
[2]School of Information Science and Engineering, University of Jinan, Jinan 250022, China
[3]Shandong Provincial Key Laboratory of Network-Based Intelligent Computing, Jinan 250022, China
[4]College of Computer and Control, Qiqihar University, Qiqihar 161006, China

Corresponding author: Jing Zhang (ise_zhangjing@ujn.edu.cn)

The work was supported in part by the funding of National Natural Science Foundation of China (NSFC) (2017-2020, No. 51679058), in part by the funding of Shandong Natural Science Foundation in China (2020–2022, No. ZR2019LZH005), and in part by the funding of the Young Innovative Talents Project of Basic Scientific Research Business Expenses for Provincial Universities of Heilongjiang Province (No. 135509210).

ABSTRACT Accumulating researches have found that lncRNAs play a key role in many important biological processes, such as chromatin modification, transcription, and post-transcription regulation. Because lncRNAs play an important role in the life process, many important complex diseases have been linked to the variation and dysfunction of lncRNAs. In current prediction researches on lncRNA-disease association, clinical prognosis information of the disease (such as pathological stage, clinical stage and so on) is rarely mentioned. In this manuscript, we apply the pathological stage data into the lncRNA-disease association prediction. Firstly, coordinates reverse rotation in circular (CRRC) is proposed. 6 clusters are calculated by the proposed cluster generating algorithm (C_lG_eA) based on CRRC. Secondly, harmonic importance ranking (HIR) is put forward. 28 core variables are obtained by the proposed selection algorithm of core variables for cancer pathological stage (SA-CV-CPS) based on HIR and cluster. Finally, on the basis of the above 28 core variables, pathological stage prediction algorithm for lncRNA-disease association based on principal component regression analysis (PSPA-LA-PCRA) is developed. Through PSPA-LA-PCRA, principal component set (including 20 PCs) and prediction model are gained. The proposed prediction model is based on unknown human lncRNA-disease association combining with the pathological stage data. Experimental results show that better results for AUC, precision rate, recall rate and F1-score of the prediction model are achieved by PSPA-LA-PCRA, which provides a favorable research premise for subsequent prediction studies of lncRNA-disease associations.

INDEX TERMS lncRNA-disease association, core variable, principal component regression analysis, pathological stage.

I. INTRODUCTION

Long non-coding RNA (lncRNA) is a class of RNA molecules that are longer than 200nt and non-coding for proteins [1]–[3]. LncRNA plays an important role in the biological process of chromatin modification, transcription and post-transcription regulation and so on [4], and has a very important biological function. The variation or dysfunction of lncRNA can lead to the occurrence of many diseases, such as lung cancer [5]–[8], breast cancer [9], [10], prostate cancer [11], [12], osteosarcoma [13], colorectal cancer [14], gastric cancer [15], bladder cancer [16] and cervical cancer [17], etc. Therefore, the research on lncRNA-disease association prediction can not only deepen the understanding of the pathogenic mechanism for complex diseases at the molecular level, but also use lncRNA as disease diagnosis, predicted biological target, drug target for treatment and prevention. Through bioinformatics methods combined with clinical and pathological data, it is a meaningful work to study lncRNAs that have a significant impact on the prognosis of cancer patients, which has important research value and social significance.

The associate editor coordinating the review of this manuscript and approving it for publication was Arif Ur Rahman.

In recent years, the research results of lncRNA-disease association prediction can be divided into two categories: machine-learning-based method and network-based method.

The first category (machine-learning-based method) is as follows. For instance, Yu *et al.* [18] proposed a novel prediction method called CFNBC that was developed by the Naïve Bayes classifier. CFNBC didn't depend on any known lncRNA-disease associations and achieved effective prediction results in condition of scarce known lncRNA-disease associations. Yuan *et al.* [19] used the cluster correlation based method to evaluate the strength of the inner relationships between disease and gene. Then a new method was proposed to predict potential lncRNA-disease associations. This method obtained a total of 2320 potential gene-disease associations (1321 lncRNA-disease pairs and 999 protein coding gene-disease pairs). But its limitations were based on known associations between diseases and lncRNAs/protein coding genes. Yao *et al.* [20] built a new prediction model (RFLDA) to improve the ability of LDA prediction models. RFLDA applied the random forest method to train prediction model with the most useful features. RFLDA formed a random forest regression model to score potential lncRNA-disease associations. However, RFLDA required a lot of negative samples to complete the calculation. But negative samples were difficult to obtain. Zeng *et al.* [21] proposed a novel computational method (SDLDA in short) by blending singular value decomposition and deep learning to predict lncRNA-disease associations. SDLDA used singular value decomposition to get linear features of lncRNAs and diseases. SDLDA used deep learning to get non-linear features of lncRNAs and diseases. Finally, the prediction model based on linear features and non-linear features of lncRNAs and diseases was constructed. Tan *et al.* [22] adopted a multi-view consensus graph learning method to establish a novel learning framework for predicting lncRNA-disease associations. It learned a consensus graph from the multiple similarity matrices. The prediction model of this work was a multi-label learning framework. Lan *et al.* [23] proposed a new method of identifying lncRNA-disease associations by collaborative deep learning, which is called LDICDL. LDICDL utilized denoise technology and matrix decomposition algorithm to achieve the prediction. Moreover, LDICDL used the hybrid model to predict associations between new lncRNA (or disease) and diseases (or lncRNA). The experiment results showed that LDICDL was competitive. Chen *et al.* [24] built a novel prediction method for lncRNA-disease association based on the lncRNA similarities, disease similarities and the support vector machine (ILDMSF in short). The lncRNA similarities and disease similarities were integrated into ILDMSF. And the support vector machine was used to predict the potential lncRNA-disease associations. The experimental results showed that ILDMSF was effective. Lan *et al.* [25] integrated multiple biological data resources, and constructed a web server for lncRNA-disease association prediction (LDAP in short). LDAP used bagging SVM to predict lncRNA-disease associations by integrating lncRNA similarity and disease similarity. The test result showed that it was able to identify known and potential new lncRNA-disease associations.

The second category (network-based method) is as follows. For instance, Chen *et al.* [26] proposed a method of HGLDA by integrating with miRNA-disease associations and lncRNA-miRNA interactions. HGLDA used the information of MiRNA (LFSCM) to complete the calculation of lncRNA functional similarity. However, HGLDA was not be applied in the prediction without any known miRNA interaction partners. Li *et al.* [27] built a new heterogeneous network to predict the potential lncRNA-disease associations and proposed a new model called LRWHLDA. LRWHLDA used an improved local random walk to achieve high prediction precision. Meanwhile, it was suitable for lacking known lncRNA-disease associations. Wang *et al.* [28] constructed a data fusion strategy for predicting lncRNA-disease associations and put forward a prediction approach(named WMFLDA). WMFLDA used different kinds of data(including genes, lncRNAs, and Disease Ontology terms). WMFLDA exhibited a lncRNA-disease association matrix by the optimized matrices and weights on the heterogeneous network. Chen *et al.* [29] came up with IRWRLDA method, which improved the restart random walk algorithm and performance. However, the limitation of IRWRLDA was how to obtain integrated lncRNA similarity based on lncRNA functional similarity and lncRNA Gaussian interaction profile kernel similarity.

To sum up, deficiencies in the current research were as follows. The existing methods can merely predict whether lncRNA is associated with disease, in other words, it can only tell us whether it is related or not, but is not able to specify what aspects of the disease associated with lncRNA, such as clinical stage, pathological stage, survival time, disease status, family history of genetic diseases, etc. Obviously, they ignore the clinical prognosis information of the disease, but the predictive analysis of clinical prognostic information for lncRNA-disease association has more practical significance and value.

However, our method correlated the pathological stage data that was treated as a decision attribute, then a pathological stage prediction algorithm for lncRNA-disease association was built. So our method can not only predict the association between lncRNA and disease, but also predict the association between lncRNA and the pathological stage of disease, so as to predict the pathological stage of disease. Furthermore, two new methods (coordinates reverse rotation in circular, harmonic importance ranking) were proposed for the prediction model. Based on the above two methods, the concept of cluster generating and core variable was proposed. Finally, a pathological stage prediction model for lncRNA-disease association based on principal component regression analysis was constructed for the lncRNA-disease association prediction. Our method was a prediction model for lncRNA-cancer pathological stage, the input was polytomy variable that was lncRNAs, the output was decision variable that was pathological stage data. Experimental results showed

that the proposed method achieved better prediction results.

II. MATERIALS AND METHODS

A. lncRNA DATA

The lncRNA expression data of prostate cancer patients were acquired from the lncRNAtor [30] database. The data was obtained, including 44 normal samples (denoted by $N = \{N_1, N_2, \cdots, N_{44}\}$) and 176 prostate cancer samples (denoted by $S = \{S_1, S_2, \cdots, S_{176}\}$). According to the differential expression p-value (P≤0.001) between normal samples and prostate cancer samples for lncRNA transcripts, a total of 480 lncRNA transcripts (lr) with greater significant differences were obtained. According to the minimum and maximum normalization method [31], lr was normalized to Lr. This process is shown in (1). Further, lr_{max} was 0.002009, lr_{min} was 0, Nor_{max} was 1, Nor_{min} was 0 and θ was 10000.

$$Lr = \left(\frac{lr - lr_{min}}{lr_{max} - lr_{min}} \times (Nor_{max} - Nor_{min}) + Nor_{min}\right) \times \theta \tag{1}$$

The 480 Lrs were arranged in ascending order according to the differential expression P-value, denoted by $Lr = \{Lr_1, Lr_2, \cdots, Lr_{480}\}$.The 480 Lr_is in Lr had a significant effect on prostate cancer. Lr_i at the top of the Lr showed a far more significant effect. 176 prostate cancer samples S and 480 Lr_is constituted the lncRNA data matrix M^L required for the study.

$$M^L = \begin{bmatrix} Lr_1^{'S_1} & Lr_2^{'S_1} & \cdots & Lr_{480}^{'S_1} \\ Lr_1^{'S_2} & Lr_2^{'S_2} & \cdots & Lr_{480}^{'S_2} \\ \vdots & \vdots & \vdots & \vdots \\ Lr_1^{'S_{176}} & Lr_2^{'S_{176}} & \cdots & Lr_{480}^{'S_{176}} \end{bmatrix}$$

B. CLINICAL DATA

Download the clinical data associated with S from the TCGA database (https://cancergenome.nih.gov). The aliquot barcode for the clinical data associated with S is shown in Table 1. Each cancer patient has a unique aliquot barcode. The clinical reference information for each S_i included sample barcode, sample type, clinical stage, pathological stage, survival time, disease status and family history of disease, etc. In this study, sample barcode associated with M^L was screened and retained. Pathological stage data was also screened and retained, which was used to predict lncRNA that has a significant impact on the prognosis of cancer patients combined with clinical data. The pathological stage data of 176 S_i s is shown in Table 2. PTNM standard was used for pathological stage data. It can be seen from Table 2 that there were 174 valid data (The categories are T2, T3 and T4.) and 2 invalid data. The distribution standard deviation of $\{T_2 \cup T_3\}$ was 7, and the distribution standard deviation of $\{T_2 \cup T_3 \cup T_4\}$ was 38.60915. So as to known, the distribution of {T2∪T3} was more balanced, which was conducive to classification learning. If the distribution was not balanced, a class imbalance problem can arise. Finally, 170 pathological stage data in {T2∪T3} were selected as available data (denoted by $PT = \{PT_1, PT_2, \cdots, PT_{170}\}$). The M^L was projected on $\{T_2 \cup T_3\}$ to correlate with the available pathological stage data. Then we obtained the association matrix M^{LC} of lncRNA data and clinical data. In this paper, core variables for pathological stage (the variables most closely related to the cancer pathological stage) will be calculated in M^{LC}, and a prediction model will be built based on the core variables.

TABLE 1. Aliquot barcode of clinical data (176).

aliquot barcode						
tss	sample	vial	portion	analyte	plate	center
CH,EJ, FC,G9, H9,HC, HI,J4	01	A,B	01,02, 11,12, 13,21, 31	R	1580,1789, 1965,2118, 2263 ,2403	7

TABLE 2. Distribution of PT.

PTNM	count	available	effective	standard deviation	
				$T_2 \cup T_3$	$T_2 \cup T_3 \cup T_4$
T_2	78	170	174	7	38.60915
T_3	92				
T_4	4	-		-	
not available	1				
total	176	-	-	-	-

$$M^{LC} = \pi_{\{T_2 \cup T_3\}}(M^L) \bowtie PT = \begin{bmatrix} Lr_1^{S_1} & Lr_2^{S_1} & \cdots & Lr_{480}^{S_1} & PT_1 \\ Lr_1^{S_2} & Lr_2^{S_2} & \cdots & Lr_{480}^{S_2} & PT_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Lr_1^{S_{170}} & Lr_2^{S_{170}} & \cdots & Lr_{480}^{S_{170}} & PT_{170} \end{bmatrix}$$

C. CRRC METHOD

The method of coordinates reverse rotation in circular (CRRC in short) used to select core variables for pathological stage is described as follows. The importance of 480 Lr_is in M^{LC} (denoted by $Significance(Lr_i)$) was calculated separately. The ranking of $Significance(Lr_i)$ was denoted by $D - rank(Significance(Lr_i))$. According to $Significance(Lr_i)$, Lr_i was built into a circular descending queue($Q = \{Q_1, Q_2, \cdots, Q_{480}\}$). According to $Significance(Lr_i)$, 480 Lr_is were evenly clockwise distributed on Q in descending order. The position of the element in Q was $D - rank(Significance(Lr_i))$, and the corresponding relation between Q and Lr was $Q_{j=D-rank(Significance(Lr_i))} = Lr_i$. Moreover, 480 Lr_is are the global scope, the importance of seed is calculated in 480 Lr_is. The subset of 480 Lr_is is the local scope (But the subset must contain seed).

According to the quartile idea, Q in the coordinate system xoy was divided into the first quartile area $Quar_1^{xoy|0}$ (the superscript represented the coordinate system and the rotation angle, and since the initial coordinate system was

xoy, the rotation angle was 0), the second quartile area $Quar_2^{xoy|0}$, the third quartile area $Quar_3^{xoy|0}$ and the fourth quartile area $Quar_4^{xoy|0}$. The circular queue Q intersected the coordinate system *xoy* at four points $Q_{xoy}^{interset}$ that were $\{Q_{xl=361}, Q_{xr=121}, Q_{yu=1(481)}, Q_{yd=241}\}$. Q_{xl} was the left intersection point of the x axis, Q_{xr} was the right intersection point of the x axis, Q_{yu} was the up intersection point of the y axis, Q_{yd} was the down intersection point of the y axis. Since $Q_{xoy}^{interset}$ can only belong to one quartile area, and it was the boundary of the quartile area. In order to ensure the unique ownership between $Q_{xoy}^{interset}$ and the quartile area, $Q_{xoy}^{interset}$ met the principle of left close-right open (it was also applicable in other coordinate systems). In other words, if $Q_{xoy}^{interset}$ was the upper bound of the quartile area, it was equal relationship; if $Q_{xoy}^{interset}$ was the lower bound of the quartile area, it was less relationship (in this case, the lower bound was the previous node of $Q_{xoy}^{interset}$). As shown in Figure 1, the first quartile area of Q in the coordinate system *xoy* was $Quar_1^{xoy|0} = \{Q_i, i \in [yu, xr)\}$, the second quartile area of Q in the coordinate system *xoy* was $Quar_2^{xoy|0} = \{Q_i, i \in [xr, yd)\}$, the third quartile area of Q in the coordinate system *xoy* was $Quar_3^{xoy|0} = \{Q_i, i \in [yd, xl)\}$, and the fourth quartile area of Q in the coordinate system *xoy* was $Quar_4^{xoy|0} = \{Q_i, i \in [xl, yu)\}$. Take $Quar_4^{xoy|0}$ for example, the upper bound of $Quar_4^{xoy|0}$ was Q_{xl} (denoted by RQ^0) and the lower bound was the previous node of Q_{yu} (that was Q_{yu-1}, and denoted by Q^0). The nodes fell into four quartile areas of the coordinate system *xoy*, and their importance decreased from $Quar_1^{xoy|0}$ to $Quar_4^{xoy|0}$.

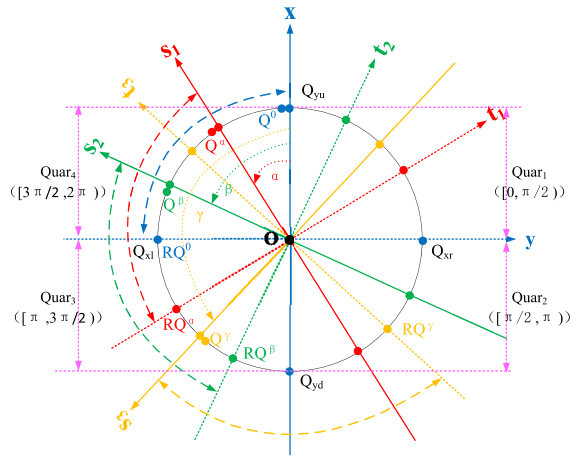

FIGURE 1. Schematic of CRRC method on the circular queue.

Definition 1 (*Operation of Coordinates Reverse Rotation in Circular,* $\Upsilon_{sot\leftarrow xoy}(Q\,|\theta)$):

The original coordinate system of $\Upsilon_{sot\leftarrow xoy}(Q\,|\theta)$ was *xoy*. By reverse rotation at $\theta \in [0, \pi]$, the coordinate system of $\Upsilon_{sot\leftarrow xoy}(Q\,|\theta)$ became *sot*. In the original coordinate system *xoy*, $Quar_4^{xoy|0}$ was $\{RQ^0, \cdots, Q^0\}$, Q^0 was $Q_{Low(Quar_4^{xoy|0})}$, RQ^0 was $Q_{Up(Quar_4^{xoy|0})}$, $Low(Quar_4^{xoy|0})$ was the subscript of the lower bound for $Quar_4^{xoy|0}$, $Up(Quar_4^{xoy|0})$ was the subscript of the upper bound for $Quar_4^{xoy|0}$. In the coordinate system *sot* by reverse rotation at θ, the fourth quartile area was denoted by $Quar_4^{sot|\theta}$, $Quar_4^{sot|\theta}$ was $\{RQ^\theta, \cdots, Q^\theta\}$, Q^θ was $Q_{Low(Quar_4^{sot|\theta})}$, RQ^θ was $Q_{Up(Quar_4^{sot|\theta})}$, $Low(Quar_4^{sot|\theta})$ was the subscript of the lower bound for $Quar_4^{sot|\theta}$, $Up(Quar_4^{sot|\theta})$ was the subscript of the upper bound for $Quar_4^{sot|\theta}$. The relation between the subscript of the lower bound and the subscript of the upper bound for a quartile area ($Quar_i$) is shown in (2). The relation between $Low(Quar_4^{xoy|0})$ and $Low(Quar_4^{sot|\theta})$ is shown in (3). $N(Q)$ was the total number of nodes in Q.

$$Up(Quar_i) = Low(Quar_i) - N(Q)/4 + 1 \tag{2}$$

$$Low(Quar_4^{sot|\theta}) = Low(Quar_4^{xoy|0}) - \frac{\theta \times N(Q)}{2\pi} \tag{3}$$

Definition 2 (*Local Incentive Area,* $Q_{sot\leftarrow xoy}^{L-incentive}$)

After $\Upsilon_{sot\leftarrow xoy}(Q\,|\theta)$ was implemented, the intersection point between the negative direction of t axis in the coordinate system *sot* and Q was Q_{tl} (that was, the left intersection point of t axis in $Q_{sot}^{interset}$). If Q_{tl} was located in the ith quartile area ($Quar_i^{xoy|0}$) in the coordinate system *xoy*, $Q_{sot\leftarrow xoy}^{L-incentive}$ was defined as $\bigcup_{j=i}^{4} Quar_j^{xoy|0}$,. The subscript of the upper bound for $Q_{sot\leftarrow xoy}^{L-incentive}$ was $Up(Q_{sot\leftarrow xoy}^{L-incentive})$, and the subscript of the lower bound for $Q_{sot\leftarrow xoy}^{L-incentive}$ was $Low(Q_{sot\leftarrow xoy}^{L-incentive})$. $Q_{sot\leftarrow xoy}^{L-incentive}$ was related to local exploitation.

Definition 3 (*Global Stability Area,* $Q_{sot\leftarrow xoy}^{G-stability}$): After $\Upsilon_{sot\leftarrow xoy}(Q\,|\theta)$ was implemented, local incentive area was generated as $Q_{sot\leftarrow xoy}^{L-incentive}$. By removing $Q_{sot\leftarrow xoy}^{L-incentive}$, the rest of Q was global stability area in the coordinate system *xoy*. Then $Q_{sot\leftarrow xoy}^{G-stability}$ was defined as $\bigcup_{j=2}^{i-1} Quar_j^{xoy|0} (i-1 \geq j)$.When $i-1$ was less than j, $Q_{sot\leftarrow xoy}^{G-stability}$ was null. And i in $Q_{sot\leftarrow xoy}^{G-stability} = \bigcup_{j=2}^{i-1} Quar_j^{xoy|0}$ was i in $Q_{sot\leftarrow xoy}^{L-incentive} = \bigcup_{j=i}^{4} Quar_j^{xoy|0}$. $Q_{sot\leftarrow xoy}^{G-stability}$ is related to global stability.

The current $Quar_1^{xoy|0}$ was got in global scope. There will be multicollinearity in global scope. To solve this problem, we put $Quar_1^{xoy|0}$ into local scope and constructed several local scopes to test the stability of Q in $Quar_1^{xoy|0}$. Stability here means that Q in $Quar_1^{xoy|0}$ ranked high both global scope and local scope. Each local scope was named as a cluster. And each local scope was a different cluster including the same $Quar_1^{xoy|0}$. Because each cluster contained $Quar_1^{xoy|0}$, $Quar_1^{xoy|0}$ was named as seed cluster(*seed* in short).

Definition 4 (*Area Truncated by Angle*): In the circular area Q, o was the center of the circle, the area truncated by acute angle($\angle xoy$) was defined as, and the node set of satisfied the principle of left close-right open. For example,

in Figure 1, the area truncated by $\angle s_1ox$ for Q was, the node set was $[Q^{\alpha}+1, Q_{yu})$.

Definition 5 (*Activity Block* $Q^{activity}_{sot\leftarrow xoy}$): After $\Upsilon_{sot\leftarrow xoy}(Q\,|\theta\,)$ was implemented for Q, activity block $Q^{activity}_{sot\leftarrow xoy}$ was generated in the local incentive area $Q^{L-incentive}_{sot\leftarrow xoy}$. $Q^{activity}_{sot\leftarrow xoy}$ was defined as $Quar_4^{sot|\theta}\cap Q^{L-incentive}_{sot\leftarrow xoy}$ (or).

Definition 6 (*Freeze Block* $Q^{freeze}_{sot\leftarrow xoy}$): After $\Upsilon_{sot\leftarrow xoy}(Q\,|\theta\,)$ was implemented for Q, freeze block $Q^{freeze}_{sot\leftarrow xoy}$ was generated in the local incentive area $Q^{L-incentive}_{sot\leftarrow xoy}$. $Q^{freeze}_{sot\leftarrow xoy}$ was defined as $Q^{L-incentive}_{sot\leftarrow xoy}-Q^{activity}_{sot\leftarrow xoy}$ (or $\cup$).

Activity block and freeze block were generated in the local incentive area ($Q^{L-incentive}_{sot\leftarrow xoy}$) after operation of coordinates reverse rotation in circular every time.

As shown in Figure 1, the blue coordinate system is xoy, the red coordinate system is s_1ot_1, the green coordinate system is s_2ot_2, and the yellow coordinate system is s_3ot_3.

It can be seen from Figure 1 as follows. After the coordinate system xoy rotates inversely at α to the coordinate system s_1ot_1 (in other words, $\Upsilon_{s_1ot_1\leftarrow xoy}(Q\,|\alpha\,)$ was performed), the fourth quartile area of the coordinate system s_1ot_1 was marked by a red dotted line ($Quar_4^{s_1ot_1|\alpha}=\{RQ^{\alpha},\cdots,Q^{\alpha}\}$). The intersection point between the negative direction of t_1 axis in the coordinate system s_1ot_1 and Q was Q_{t_1l}, which was in $Quar_3^{xoy|0}$. So local incentive area $Q^{L-incentive}_{s_1ot_1\leftarrow xoy}$ was $Quar_3^{xoy|0}\cup Quar_4^{xoy|0}$(including $Up(Q^{L-incentive}_{s_1ot_1\leftarrow xoy})=Q_{yd}$ and $Low(Q^{L-incentive}_{s_1ot_1\leftarrow xoy})=Q_{yu}-1$), activity block $Q^{activity}_{s_1ot_1\leftarrow xoy}$ was, freeze block $Q^{freeze}_{s_1ot_1\leftarrow xoy}$ was $\cup$, global stability area $Q^{G-stability}_{s_1ot_1\leftarrow xoy}$ was $Quar_2^{xoy|0}$ (including $Up(Q^{G-stability}_{s_1ot_1\leftarrow xoy})=Q_{xr}$ and $Low(Q^{G-stability}_{s_1ot_1\leftarrow xoy})=Q_{yd}-1$).

After the coordinate system xoy rotates inversely at β to the coordinate system s_2ot_2 (in other words, $\Upsilon_{s_2ot_2\leftarrow xoy}(Q\,|\beta\,)$ was performed), the fourth quartile area of the coordinate system s_2ot_2 was marked by a green dotted line ($Quar_4^{s_2ot_2|\beta}=\{RQ^{\beta},\cdots,Q^{\beta}\}$). The intersection point between the negative direction of t_2 axis in the coordinate system s_2ot_2 and Q was Q_{t_2l}, which was in $Quar_3^{xoy|0}$. So local incentive area $Q^{L-incentive}_{s_2ot_2\leftarrow xoy}$ was $Quar_3^{xoy|0}\cup Quar_4^{xoy|0}$ (including $Up(Q^{L-incentive}_{s_2ot_2\leftarrow xoy})=Q_{yd}$ and $Low(Q^{L-incentive}_{s_2ot_2\leftarrow xoy})=Q_{yu}-1$), activity block $Q^{activity}_{s_2ot_2\leftarrow xoy}$ was, freeze block $Q^{freeze}_{s_2ot_2\leftarrow xoy}$ was $\cup$, global stability area $Q^{G-stability}_{s_2ot_2\leftarrow xoy}$ was $Quar_2^{xoy|0}$ (including $Up(Q^{G-stability}_{s_2ot_2\leftarrow xoy})=Q_{xr}$ and $Low(Q^{G-stability}_{s_2ot_2\leftarrow xoy})=Q_{yd}-1$).

After the coordinate system xoy rotates inversely at γ to the coordinate system s_3ot_3 (in other words, $\Upsilon_{s_3ot_3\leftarrow xoy}(Q\,|\gamma\,)$ was performed), the fourth quartile area of the coordinate system s_3ot_3 was marked by a yellow dotted line ($Quar_4^{s_3ot_3|\gamma}=\{RQ^{\gamma},\cdots,Q^{\gamma}\}$). The intersection point between the negative direction of t_3 axis in the coordinate system s_3ot_3 and Q was Q_{t_3l}, which was in $Quar_2^{xoy|0}$. So local incentive area $Q^{L-incentive}_{s_3ot_3\leftarrow xoy}$ was $Quar_2^{xoy|0}\cup Quar_3^{xoy|0}\cup Quar_4^{xoy|0}$ (including $Up(Q^{L-incentive}_{s_3ot_3\leftarrow xoy})=Q_{xr}$ and $Low(Q^{L-incentive}_{s_3ot_3\leftarrow xoy})=Q_{yu}-1$), activity block $Q^{activity}_{s_3ot_3\leftarrow xoy}$ was, freeze block $Q^{freeze}_{s_3ot_3\leftarrow xoy}$ was $\cup$, global stability area $Q^{G-stability}_{s_3ot_3\leftarrow xoy}$ was null.

D. HARMONIC IMPORTANCE RANKING

The criteria of harmonic importance ranking (HIR in short) on core variable selection for pathological stage is described as follows.

Definition 7 (*Importance Ranking Matrix of seed on n Clusters, IRm*):

IRm(*cluster*) was composed of the importance ranking of Q_i in *seed* on each n clusters of Q, and $N(seed)$ was the total number of *seed*. Further, the importance ranking of the same *seed* is separately computed on different clusters. N clusters of Q were $(cluster_1, cluster_2,\cdots, cluster_n)$. The relationship between $cluster_i$ and *seed* are constrained by (4).

$$IRm(cluster)=\begin{bmatrix} Q_1^{cluster_1} & Q_1^{cluster_2} & \cdots & Q_1^{cluster_n}\\ Q_2^{cluster_1} & Q_2^{cluster_2} & \cdots & Q_2^{cluster_n}\\ \vdots & \vdots & \vdots & \vdots\\ Q_{N(seed)}^{cluster_1} & Q_{N(seed)}^{cluster_2} & \cdots & Q_{N(seed)}^{cluster_n}\end{bmatrix}$$
$$cluster_i\cap seed=seed$$
$$cluster_i-seed\neq\emptyset \tag{4}$$

Each row in *IRm* was n importance ranking values that were produced by Q_i on *cluster*. The importance ranking of the $ith Q_i$ on the $jth cluster_j$ was denoted by $Q_i^{cluster_j}$. N importance ranking values of Q_i were denoted by $Q_i{\sim}n=\bigcup_{j=1}^{n}Q_i^{cluster_j}$. $Q_i\sim n$ was the intermediate calculation process which determined the final importance ranking of Q_i. The arithmetic means the value of $Q_i\sim n$ was denoted by $AMV(Q_i{\sim}n)$(that was $\frac{1}{n}\sum_{j=1}^{n}Q_i^{cluster_j}$). The standard deviation of $Q_i\sim n$ was denoted by $SD(Q_i\sim n)$ (that was $\sqrt{\frac{1}{n}\sum_{j=1}^{n}(Q_i^{cluster_j}-AMV(Q_i))^2}$). $AMV(Q_i\sim n)$ and $SD(Q_i\sim n)$ were two statistic variables that had a big impact on the final importance ranking of Q_i. $AMV(Q_i\sim n)$ reflected the general trend of importance ranking. The higher the ranking of $AMV(Q_i\sim n)$ was, the more important Q_i was. $SD(Q_i\sim n)$ reflected the stability of importance ranking. If the importance ranking deviation was larger for each time, it indicated that the importance ranking had poor stability and certain risk existed in the ranking. Furthermore, the smaller the value of $SD(Q_i\sim n)$ was, the better the performance was. In other words, the smaller the dispersion degree was, the better the performance was. To get a good balance between $AMV(Q_i{\sim}n)$ and $SD(Q_i\sim n)$, the following rules of harmonic importance ranking were defined.

Definition 8 (*Harmonic Importance Ranking, HIR*($Q_i\sim n$, $\frac{SD}{AMV}$)): The definition of harmonic importance ranking on $Q_i\sim n$ is shown in (5). $HIR(Q_i\sim n)$ was made up of three parts. They were $AMV(Q_i\sim n)$, $SD(Q_i\sim n)$ and harmonic index $\frac{SD}{AMV}$. The harmonic index $\frac{SD}{AMV}$ referred to the weight ratio of $AMV(Q_i\sim n)$ and $SD(Q_i\sim n)$ in the harmonic

process.

$$HIR(Q_i \sim n, \frac{SD}{AMV}) = AMV(Q_i \sim n) \times ((SD(Q_i \sim n))^{\frac{SD}{AMV}} + 1) \tag{5}$$

The harmonic importance ranking criteria of core variable selection for pathological stage was to determine the final importance ranking by the ranking of $HIR(Q_i \sim n, \frac{SD}{AMV})$.

Definition 9 (*Overflow Amplification, OVA*):

$OVA(Q_j^{cluster_i})$ was specific to $Q_j^{cluster_i}$ in *IRm*. If it was greater than $N(seed)$, $Q_j^{cluster_i}$ was judged to be in overflow state. This indicated that $Q_j^{cluster_i}$ had poor stability. So $Q_j^{cluster_i}$ was magnified, which was eliminated in the process of core variable selection for pathological stage. Conversely, if it was less than or equal to $N(seed)$, $Q_j^{cluster_i}$ did not change anything. Overflow amplification is shown in (6). When $Q_j^{cluster_i}$ was less than or equal to $N(seed)$, $Q_j^{cluster_i}$ was unchanged. Conversely, $Q_j^{cluster_i}$ was enlarged as $(Q_j^{cluster_i} - Q_j^{cluster_i}\%(N(seed)+1)) \times N(Q) + Q_j^{cluster_i}$.

$$\begin{aligned} &OVA(Q_j^{cluster_i}) \\ &= (Q_j^{cluster_i} - Q_j^{cluster_i}\%(N(seed)+1)) \times N(Q) + Q_j^{cluster_i} \end{aligned} \tag{6}$$

E. C_l G_eA

Cluster generating algorithm (C_lG_eA in short) used CRRC method to generate n clusters on Q. These n clusters were the basis for calculating the harmonic importance ranking. C_lG_eA executed the operation of coordinates reverse rotation in circular for each time, it took the seed cluster ($Quar_1^{xoy|0}$) as the original combining with the global stability area and the activity block of local incentive area, and a new cluster was generated. Because C_lG_eA correlated global stability area and local incentive area on the basis of the seed cluster, the cluster generated by C_lG_eA took into account both global stability and local exploitation [32]–[34]. Besides, it also had good diversity [35]–[37] among clusters.

Take $\theta = {}^{\pi}/_{6}$ as an example in Figure 2. C_lG_eA was used to divide Q into 6 clusters. In Figure 2 (a)-(f), the purple shaded area is local incentive area, the yellow shaded area is global stability area, the green shaded area is seed cluster, the red coordinate system is the rotated coordinate system (surrounded by a red enclosing rectangle), and the blue coordinate system is the initial coordinate system *xoy*. In Figure 2(a), the coordinate system s_1ot_1 is derived from the coordinate system *xoy* by rotation at ${}^{\pi}/_{6}$. Thus the first cluster was formed as $cluster_1 = \{Q_1, \cdots, Q_{240}, Q_{321}, \cdots, Q_{440}\}$. In Figure 2(b), the coordinate system s_2ot_2 is derived from the coordinate system *xoy* by rotation at ${}^{2\pi}/_{6}$. Thus the second cluster was formed as $cluster_2 = \{Q_1, \cdots, Q_{240}, Q_{281}, \cdots, Q_{400}\}$. In Figure 2(c), the coordinate system s_3ot_3 is derived from the coordinate system *xoy* by rotation at ${}^{3\pi}/_{6}$. Thus the third cluster was formed as $cluster_3 = \{Q_1, \cdots, Q_{360}\}$. In Figure 2(d), the coordinate system s_4ot_4 is derived from the coordinate system *xoy* by rotation at ${}^{4\pi}/_{6}$. Thus the fourth cluster was formed as $cluster_4 = \{Q_1, \cdots, Q_{120}, Q_{201}, \cdots, Q_{320}\}$. In Figure 2(e), the coordinate system s_5ot_5 is derived from the coordinate system *xoy* by rotation at ${}^{5\pi}/_{6}$. Thus the fifth cluster was formed as $cluster_5 = \{Q_1, \cdots, Q_{120}, Q_{161}, \cdots, Q_{280}\}$. In Figure 2(f), the coordinate system s_6ot_6 is derived from the coordinate system *xoy* by rotation at ${}^{6\pi}/_{6}$. Thus the sixth cluster was formed as $cluster_6 = \{Q_1, \cdots, Q_{240}\}$.

Algorithm 1 $C_lG_eA(Q, \theta_{init})$

1: $n = \left\lceil \frac{\pi}{\theta_{init}} \right\rceil$;
2: $seed = Quar_1^{xoy|0}$;
3: **for** $i = 1$ **to** n **do**
4: $\quad \theta = \theta_{init} \times i$;
5: $\quad \Upsilon_{s_iot_i \leftarrow xoy}(Q \,|\theta)$;
6: $\quad cluster_i = seed \cup Q_{s_iot_i \leftarrow xoy}^{G-stability} \cup Q_{s_iot_i \leftarrow xoy}^{activity}$;
7: **end for**
8: **for** $\forall cluster_i$ in $cluster$ **do**
9: $\quad$ **if** $cluster_i - seed \neq \emptyset$ **then**
10: $\quad\quad cluster = \bigcup_{i=1}^{n} cluster_i$;
11: $\quad$ **end if**
12: **end for**
13: **return** $cluster, seed$;

F. SA-CV-CPS

Selection algorithm of core variables for cancer pathological stage (SA-CV-CPS in short) was divided into four sections.

Section 1(step 1-3) Call C_lG_eA to initialize *cluster* and *seed*. The total number of nodes for *seed* was initialized, the total number of clusters for *cluster* was initialized and the core variable set of cancer pathological stage (Q^{core}) was initialized.

Section 2(step 4-8) Calculate n cluster importance ranking matrix (*IRm*) of *seed*.

Section 3(step 9-11) Calculate harmonic importance ranking ($HIR(Q_i \sim n, \frac{AMV}{SD})$) of $Q_i \sim n$, which was used to determine final importance ranking ($rank_{final}(Q_i)$) of Q_i.

Section 4(step 12-16) Calculate Q^{core} according to $rank_{final}(Q_i)$ and break point of core variable selection (cut^{core}). $Top(rank_{final}(Q_i))$ represented Q_i with the highest value of $rank_{final}(Q_i)$ in *seed*. In addition, there might be overlaps in the primary results (Q^{*core}). Q^{core} was obtained by removing overlap operation ($R - overlap(Q^{*core})$).

Finally, the core variable set of cancer pathological stage (Q^{core}) was selected from Q after SA-CV-CPS. Q^{core} selected by SA-CV-CPS contained Q_i^{core} (the number of Q_i^{core} was $N(Q^{core})$) that were most closely related to the cancer pathological stage, which will be applied to follow-up correlation prediction studies. Moreover, Q^{core} was denoted by $\{Q_1^{core}, \cdots, Q_{N(Q^{core})}^{core}\}$.

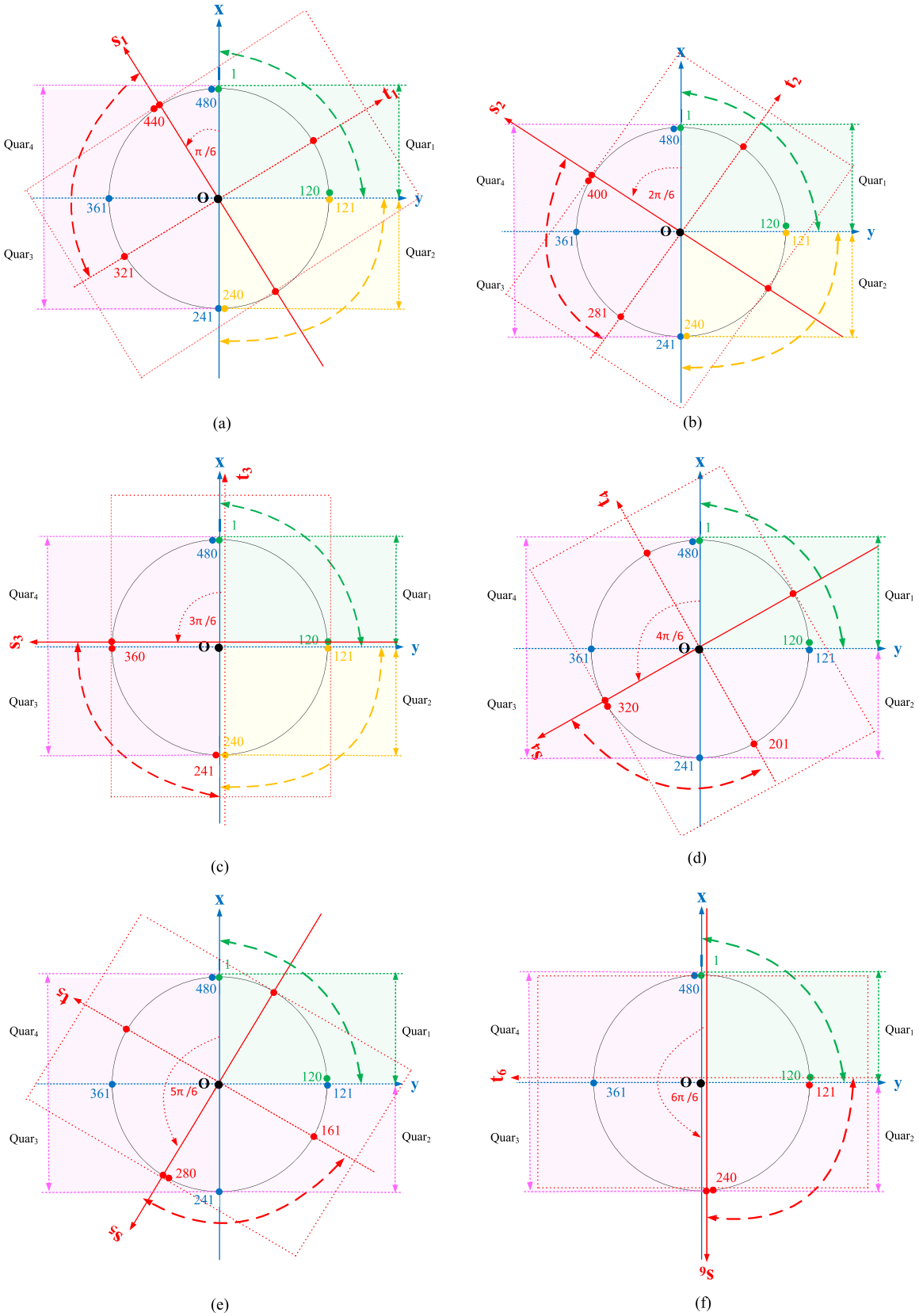

FIGURE 2. The executing process of C_lG_eA.

Algorithm 2 SA-CV-CPS($seed, cluster, \frac{SD}{AMV}, cut^{core}$)

$number_{seed} = N(seed)$;
$n = N(cluster)$;
$Q^{core} = \emptyset$;
for $i = 1$ **to** $number_{seed}$ **do**
  for $j = 1$ **to** n **do**
    $IRm(cluster) \leftarrow Q_i^{cluster_j} \times OVA(Q_i^{cluster_j})$;
  end for
end for
for $\forall Q_i$ in $seed$ **do**
  $rank_{final}(Q_i) = HIR(Q_i \sim n, \frac{SD}{AMV})$;
end for
while $(N(Q^{*core}) < number_{seed} \times cut^{core})$ **do**
  $Q^{*core} \leftarrow Top(rank_{final}(Q_i))$ in $seed$;
  $seed \leftarrow seed - Top(rank_{final}(Q_i))$;
end while
$Q^{core} = R - overlap(Q^{*core})$;
return Q^{core};

Algorithm 3 PSPA-LA-PCRA($M^{LC}, \theta, \frac{SD}{AMV}, cut^{core}, c^r$)

for $\forall Lr_i \in M^{LC}$ **do**
  $Significance(Lr_i) \leftarrow VIRA - RF(M^{LC})$;
end for
$Q \leftarrow Sort(Lr_{M^{LC}})^{\overrightarrow{significance(Lr_i)}})$
$(seed, cluster) = C_l G_e A(Q, \theta)$;
$Q^{core} = SA - CV - CPS$
  $(speed, cluster, \frac{SD}{AMV}, cut^{core})$;
$M^{LC-core} = \pi_{Q^{core}}(M^{LC})$;
$cor(MLC - core)$;
$PCA = primecomp(M^{LC-core}, cor = T)$;
$summary(PCA)$;
for each $comp.k$ in PCA **do**
  if $C_u P_r(comp.k) \leq c^r$ **then**
    $PCA^{final} \leftarrow comp.k$;
  else
    break;
  end if
end for
$P_r M_o = logistic - regression(PCA^{final})$;
return $P_r M_o$;

G. PSPA-LA-PCRA

Pathological stage prediction algorithm for lncRNA-disease association based on principal component regression analysis was named PSPA-LA-PCRA in short. PSPA-LA-PCRA was divided into five sections.

Section 1 (step 1-4) Variable importance ranking algorithm based on Random Forest was named VIRA-RF in short. By calling VIRA-RF, the importance of Lr_i in M^{LC}(that was $Significance(Lr_i)$) was calculated, and Q was built in descending order(that was $Sort(Lr_{M^{LC}})^{\overrightarrow{significance(Lr_i)}}$) of importance.

Section 2(step 5) *Cluster* and *seed* were calculated by calling C_l G_eA.

Section 3(step 6) Core variable set of cancer pathological stage(Q^{core}) was calculated by calling SA-CV-CPS. Besides, $Q_i^{cluster_j}$ was also calculated by calling VIRA-RF in SA-CV-CPS.

Section 4(step 7-14) In step 7, $\pi_{Q^{core}}(M^{LC})$ was obtained by projection operation of M^{LC} on Q^{core}, and then $M^{LC-core}$ was also obtained. In step 8-14, principal component analysis was completed on $M^{LC-core}$. In step 8, correlation matrix of Q^{core} on $M^{LC-core}$ was got. In step 9, correlation matrix eigenvalue and principal component load were computed, and then principal component candidate set PCA was acquired. PCA was denoted by $\{comp.1, \cdots comp.k, \cdots, comp.N(PCA)$. In step 10-14, the principal component set (PCA^{final}) was selected from PCA according to the cumulative variance contribution rate. The selection rule was less than or equal to the threshold of cumulative variance contribution (c^r).

Section 5(step 15) In step15, the prediction model ($P_r M_o$) was obtained by performing logistic regression on the principal component set (PCA^{final}).

The flowchart of the whole process for PSPA-LA-PCRA is shown in Figure 3.

III. RESULTS AND DISUSSION

A. PARAMETER SETTING

The parameter setting of the algorithm is shown in Table 3. It includes parameter θ_{init} used in CRRC, harmonic index $\frac{SD}{AMV}$ used in computing HIR, break point of core selection cut^{core} used in calculating core variable set for cancer pathological stage (Q^{core}) and the threshold of cumulative variance contribution (c^r) used in principal component analysis.

TABLE 3. Parameter setting.

$seed$	θ_{init}	SD/AMV	cut^{core}	c^r
$Q_{1\leq i\leq 120}$	$\pi/6$	$2/3$	0.25	0.9

B. PERFORMANCE EVALUATION OF HIR

The three categories of ranking values on Q_i that were involved in calculating HIR, which were the original ranking $Orig(Q_i)$, the mean ranking $AMV(Q_i \sim n)$ and the harmonic ranking $HIR(Q_i \sim n, \frac{SD}{AMV})$. $Orig(Q_i)$ was the ranking value in the global scope, and reflected global stability. $AMV(Q_i \sim n)$ was the mean ranking among different clusters, which reflected the local exploitation due to the local characteristics of cluster. Because there were unstable factors among different clusters, $HIR(Q_i \sim n, \frac{SD}{AMV})$ reflected the stability among clusters. On the premise of ensuring global stability, the algorithm in this paper took the local exploitation of clusters and the stability among clusters into full account. $AMV(Q_i \sim n)$ was calculated by $Orig(Q_i)$ among the different clusters. According to harmonic index,

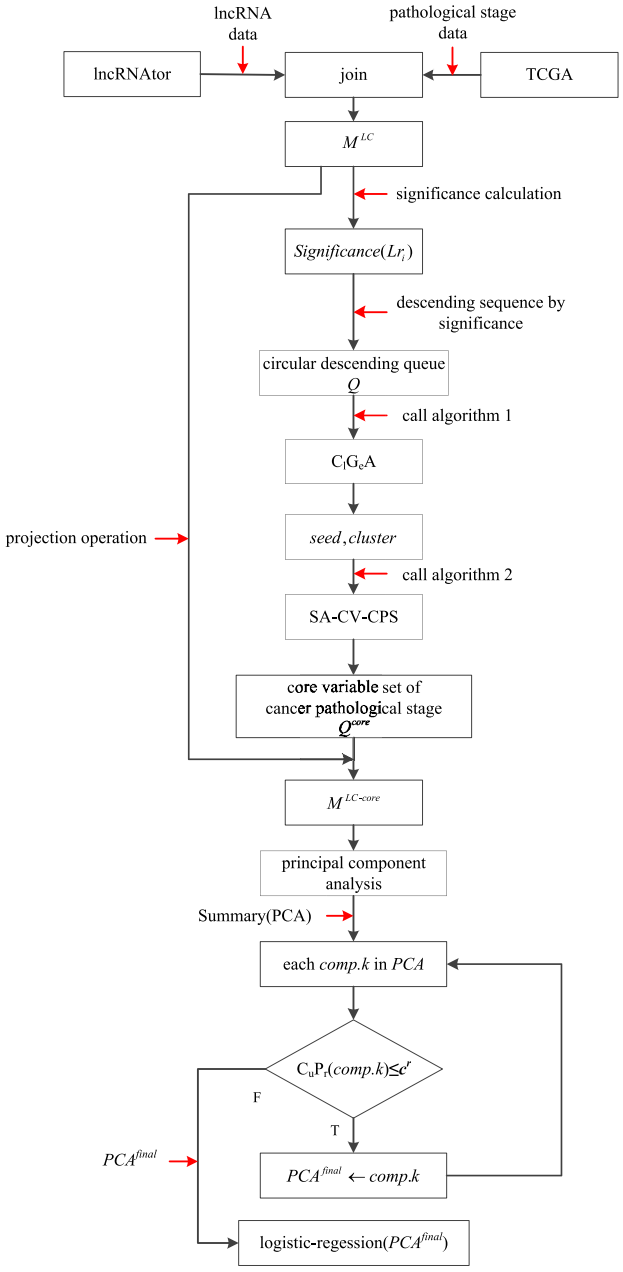

FIGURE 3. Flowchart of the whole process for PSPA-LA-PCRA.

$HIR(Q_i \sim n, \frac{SD}{AMV})$ was calculated by $AMV(Q_i \sim n)$. $HIR(Q_i \sim n, \frac{SD}{AMV})$ was the final ranking value. In order to investigate the relationship among ranking values of the three categories, the ranking values with large fluctuations were selected for comparative analysis. The standard deviation of the ranking values of the three categories was greater than 15, and the comparison curve was drawn, as shown in Figure 4. The overall trend of $HIR(Q_i \sim n, \frac{SD}{AMV})$ was between $Orig(Q_i)$ and $AMV(Q_i \sim n)$, and closer to $Orig(Q_i)$. Therefore, $HIR(Q_i \sim n, \frac{SD}{AMV})$ considered global stability, local exploitation and stability in a more comprehensive way.

TABLE 4. The value of bound point.

set	*upper-bound*	*lower-bound*
seed	1	120
$Rov\text{-}cluster_1$	321	440
$Rov\text{-}cluster_2$	281	400
$Rov\text{-}cluster_3$	241	360
$Rov\text{-}cluster_4$	201	320
$Rov\text{-}cluster_5$	161	280
$Rov\text{-}cluster_6$	121	240

C. PERFORMANCE EVALUATION OF CLUSTER

A total of 6 clusters($cluster_{1\leq i\leq 6}$) was obtained by the calculation of C_lG_eA, the distribution quality of these 6 clusters will affect the local exploitation of the algorithm. Figure 5 shows the distribution of the boundary points for $Rov-cluster_{1\leq i\leq 6}$ and $seed$. $Rov-cluster_{1\leq i\leq 6}$ in Figure 5 is cluster without internal overlap and $seed$, the specific calculation of which is in (7). At this point, $cluster_i^{not}$ was a set of the other clusters except for $cluster_i$. Table 4 shows the distribution of bound point in $Rov - cluster_{1\leq i\leq 6}$. Figure 5 shows that the upper bound and the lower bound of the cluster are more evenly distributed in the global scope. Therefore, this is conducive to comprehensively complete the local exploitation in the global scope.

$$Rov - cluster_i = cluster_i - seed - (cluster_i \cap \forall cluster_i^{not}) \tag{7}$$

D. PERFORMANCE EVALUATION OF SA-CV-CPS

28 LncRNAs that were most closely related to the pathological stage of prostate cancer were calculated by SA-CV-CPS. They constituted core variable set (Q^{core}) used for pathological stage prediction in Table 5. The calculation of Q^{core} was realized according to HIR adjusted by OVA. Figure 6 shows a comparison between used OVA and unused OVA. IRm with used OVA was denoted by $IRm-OVA$. IRm with unused OVA was denoted by IRm. If the number of Q_i in $Cluster_{1\leq i\leq 6}$ that was greater than $N(seed) - 10$ from the data set was not greater than $N(seed)$ in AMV, this Q_i will be selected. Next, the number of the selected Q_i was denoted by $N(*cluster_i)$ which showed in (8). Then the greater $N(*cluster_i)$ was, the poorer the stability performance was. As can be seen from Figure 6, $N(*cluster_i)$ of $IRm - OVA$ was significantly lower than IRm. In the experiment, if the threshold value was changed from $N(seed) - 10$ to $N(seed)$, $IRm - OVA$ would be all 0. All of these indicate that OVA has improved the stability of the algorithm.

$$N(*cluster_i) = \sum_{j=1}^{AMV \leq N(seed)} count(Q_j^{cluster_i} > N(seed) - 10) \tag{8}$$

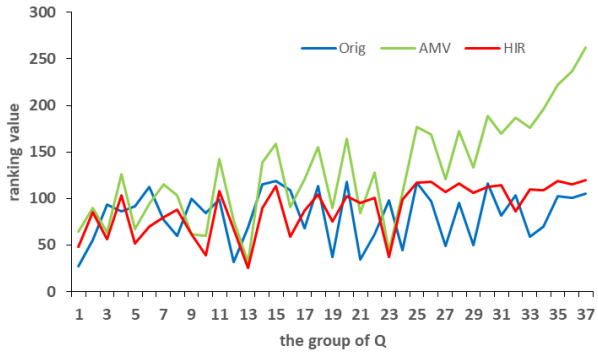

FIGURE 4. Relationship of three ranking values.

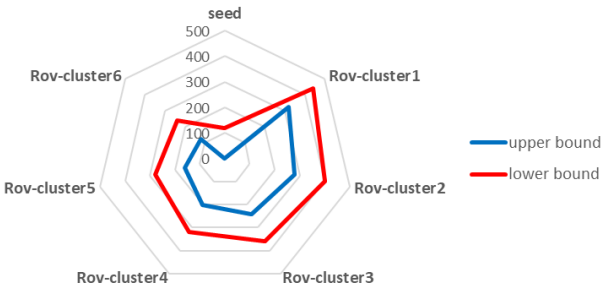

FIGURE 5. Distribution of bound point in cluster.

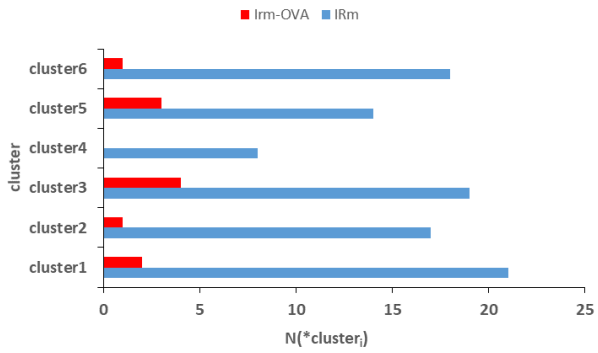

FIGURE 6. Performance comparison of OVA.

E. PERFORMANCE EVALUATION OF PSPA-LA-PCRA

PSPA-LA-PCRA mainly performed principal component analysis on Q^{core} containing 28 LncRNAs. The distribution diagram on cumulative contribution rate of $comp.i$ in the principal component candidate set (PCA) was obtained (as shown in Figure 7). The threshold of cumulative variance contribution (c^r) was set to 0.9. As shown in Figure 7, the cumulative variance contribution rate of $comp.20$ reaches 0.9. Therefore, Q^{core} obtained the principal component set (PCA^{final}) through principal component analysis, which contains 20 principal components in total(denoted by $\{comp.1, \cdots, comp.20\}$).

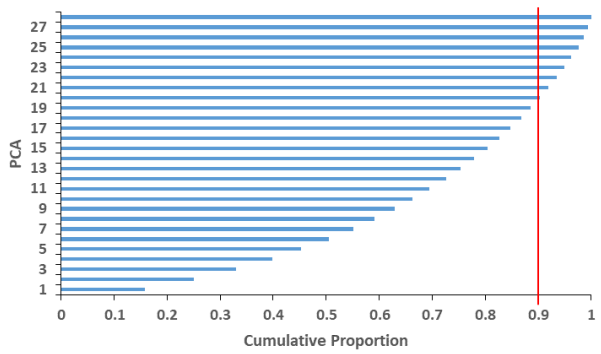

FIGURE 7. Cumulative Proportion of principal component.

The prediction model P_rM_o based on PCA^{final} was the final result of PSPA-LA-PCRA.

Three methods (REPTree [38], NaïveBayes [39] and SMO [40])were selected to compare with PSPA-LA-PCRA by 10-fold LOOCV(Leave-One-Out cross validation). The comparison experiments were carried out from four aspects: AUC, prediction rate, recall rate and F1-score. The comparison results of ROC curve and AUC are shown in Figure 8. The mean AUC of the four methods was 0.727. Figure 8 shows that the AUC of PSPA-LA-PCRA is the highest (0.822), which is 9.5 percentage points higher than the mean AUC, 16.4 percentage points higher than REPTree, 12.7 percentage points higher than NaïveBayes, 8.8 percentage points higher than SMO.

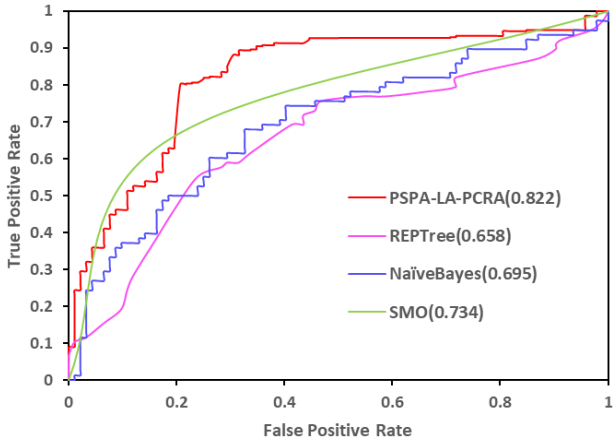

FIGURE 8. ROC comparison curve.

Figure 9 provides the comparison results of precision rate, recall rate and F1-score on the four methods. It can be seen that the precision rate of PSPA-LA-PCRA is the highest (0.811), which is 14.4 percentage points higher than REPTree, 13 percentage points higher than NaïveBayes and 5.3 percentage points higher than SMO. Obviously, the recall rate of PSPA-LA-PCRA is the highest (0.801), which is 8.4 percentage points higher than REPTree, 10.5 percentage points higher than NaïveBayes and 16 percentage points higher than

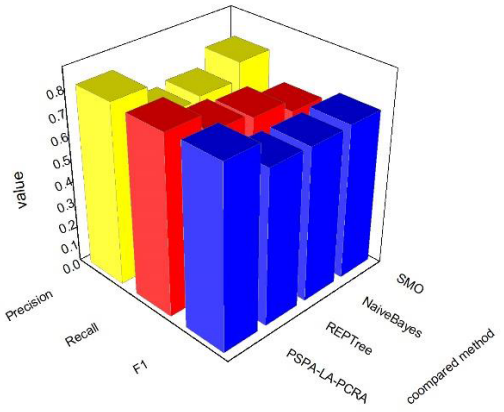

FIGURE 9. Comparison of prediction accuracy, recall rate and F1-score.

TABLE 5. The core variables Q^{core}.

rank	gene name	rank	gene name
1	AC009501.4	15	LINC00402
2	RP5-1096D14.2	16	RP5-1121A15.1
3	C1QTNF9B-AS1	17	RP11-573I11.2
4	RP11-161M6.2	18	RP11-693N9.2
5	LINC00152	19	PCA3
6	MRVI1-AS1	20	ZNFX1-AS1
7	ERICH1-AS1	21	ZNF300P1
8	PART1	22	RP11-438E5.1
9	RP5-916L7.1(7313)	23	PLK1S1
10	AC079776.2	24	RP11-398J10.1
11	RNF144A-AS1	25	RP5-1007M22.2
12	RP11-627G23.1	26	RP5-916L7.1
13	RP1-207H1.3	27	AC124944.2
14	RP11-17A19.1	28	CIDECP

SMO. In conclusion, it can be seen that both the precision rate and recall rate of PSPA-LA-PCRA achieved good results.

In order to further investigate the comprehensive situation of precision rate and recall rate, the index of F1-score was compared and analyzed. F1-score was the harmonic mean of precision rate and recall rate. Obviously, the F1-score of PSPA-LA-PCRA is the highest (0.806), which is 11.5 percentage points higher than REPTree, 11.8 percentage points higher than NaïveBayes and 11.2 percentage points higher than SMO.

These results indicate that the AUC, precision rate, recall rate and F1-score for PSPA-LA-PCRA were all good. The main reason why PSPA-LA-PCRA achieved good performance on the AUC, prediction precision, recall rate and F1-score was that global stability and local exploitation was both considered. The core variable set of cancer pathological stage in PSPA-LA-PCRA was initially obtained in global scope, then was trained for its stability in local scope. Several clusters were constructed to test the stability. In local exploitation stage, data with global stability was inspected by *HIR*. Eventually, the unstable

data was removed and the stable data was retained. Data with better stability ranked high in both global scope and local scope. In other words, PSPA-LA-PCRA considered global stability and local exploitation in a more comprehensive way. This was the key to the overall improvement of algorithm performance.

IV. CONCLUSION AND DISCUSSION

In this manuscript, we correlated pathological stage data to predict lncRNA-disease association, and constructed a pathological stage prediction algorithm for lncRNA-disease association based on principal component regression analysis (PSPA-LA-PCRA). PSPA-LA-PCRA was based on unknown human lncRNA-disease associations. The core modules of PSPA-LA-PCRA included CRRC method, HIR method, C_lG_eA algorithm and SA-CV-CPS algorithm. For C_lG_eA, a learning mode performed a CRRC operation every time, a new cluster will be generated by taking the seed cluster as the original, combining the global stability area and the activity block of the local incentive area. SA-CV-CPS selected the lncRNAs most closely associated with the cancer pathological stage for subsequent association prediction studies. The lncRNAs most closely associated with the cancer pathological stage were called core variables. Then core variables performed principal component analysis by PSPA-LA-PCRA. Finally, a novel prediction model was obtained by principal component logistic regression. Experimental results showed that the proposed method in this manuscript has good predictive result in these aspects of AUC, prediction precision, recall rate and F1-score.

REFERENCES

[1] E. Mathieu, M. Belhocine, L. T. Dao, D. Puthier, and S. Spicuglia, "Functions of lncRNA in development and diseases," *Medecine Sci.*, vol. 30, nos. 8–9, pp. 790–796, Sep. 2014.

[2] W. Sun, Y. Shi, Z. Wang, J. Zhang, H. Cai, J. Zhang, and D. Huang, "Interaction of long-chain non-coding RNAs and important signaling pathways on human cancers," *Int. J. Oncol.*, vol. 53, no. 6, pp. 2343–2355, Sep. 2018.

[3] H. Xie, B. Ma, Q. Gao, H. Zhan, Y. Liu, Z. Chen, S. Ye, J. Li, L. Yao, and W. Huang, "Long non-coding RNA CRNDE in cancer prognosis: Review and meta-analysis," *Clinica Chim. Acta*, vol. 485, pp. 262–271, Oct. 2018.

[4] R. J. Taft, K. C. Pang, T. R. Mercer, M. Dinger, and J. S. Mattick, "Non-coding RNAs: Regulators of disease," *J. Pathol.*, vol. 220, no. 2, pp. 126–139, Jan. 2010.

[5] Z. Chen, Z. Chen, T. Lei, X. Chen, J. Gu, J. Huang, B. Lu, and Z. Wang, "Long non-coding RNA in lung cancer," *Clinica Chim. Acta*, vol. 504, pp. 190–200, Nov. 2019.

[6] C. Fang, L. Wang, C. Gong, W. Wu, C. Yao, and S. Zhu, "Long non-coding RNAs: How to regulate the metastasis of non–small-cell lung cancer," *J. Cellular Mol. Med.*, vol. 24, no. 6, pp. 3282–3291, Feb. 2020.

[7] S.-P. Dai, J. Jin, and W.-M. Li, "Diagnostic efficacy of long non-coding RNA in lung cancer: A systematic review and meta-analysis," *Postgraduate Med. J.*, vol. 94, no. 1116, pp. 578–587, Oct. 2018.

[8] T. Li, R. He, J. Ma, Z. Li, X. Hu, and G. Chen, "Long non-coding RNAs in small cell lung cancer: A potential opening to combat the disease," *Oncol. Rep.*, vol. 40, no. 4, pp. 1831–1842, Aug. 2018.

[9] M. Cantile, M. Di Bonito, M. Cerrone, F. Collina, M. De Laurentiis, and G. Botti, "Long non-coding RNA HOTAIR in breast cancer therapy," *Cancers*, vol. 12, no. 5, p. 1197, May 2020, doi: 10.3390/cancers12051197.

[10] Y.-L. Wang, L.-C. Liu, Y. Hung, C.-J. Chen, Y.-Z. Lin, W.-R. Wu, and S.-C. Wang, "Long non-coding RNA HOTAIR in circulatory exosomes is correlated with ErbB2/HER2 positivity in breast cancer," *Breast*, vol. 46, pp. 64–69, Aug. 2019.

[11] M. Xu, S. Gong, Y. Li, J. Zhou, J. Du, C. Yang, M. Yang, F. Zhang, C. Liang, and Z. Tong, "Identifying long non-coding RNA of prostate cancer associated with radioresponse by comprehensive bioinformatics analysis," *Frontiers Oncol.*, vol. 10, p. 498, Apr. 2020, doi: 10.3389/fonc.2020.00498.

[12] Y. Yan, Z. Chen, Y. Xiao, X. Wang, and K. Qian, "Long non-coding RNA SNHG6 is upregulated in prostate cancer and predicts poor prognosis," *Mol. Biol. Rep.*, vol. 46, no. 3, pp. 2771–2778, Jun. 2019.
[13] D. Chen, H. Wang, M. Zhang, S. Jiang, C. Zhou, B. Fang, and P. Chen, "Abnormally expressed long non-coding RNAs in prognosis of osteosarcoma: A systematic review and meta-analysis," *J. Bone Oncol.*, vol. 13, pp. 76–90, Nov. 2018.
[14] S. J. O'Brien, C. Bishop, J. Hallion, C. Fiechter, K. Scheurlen, M. Paas, J. Burton, and S. Galandiuk, "Long non-coding RNA (lncRNA) and epithelial-mesenchymal transition (EMT) in colorectal cancer: A systematic review," *Cancer Biol. Therapy*, vol. 21, no. 9, pp. 769–781, Sep. 2020, doi: 10.1080/15384047.2020.1794239.
[15] Y. Li, D. Ma, T. Li, and Y. Yin, "Identification of functional long non-coding RNAs in gastric cancer by bioinformatics analysis," *Int. J. Exp. Pathol.*, vol. 101, nos. 3–4, pp. 96–105, Jul. 2020.
[16] Y. Zhan, L. Zhang, S. Yu, J. Wen, Y. Liu, and X. Zhang, "Long non-coding RNA CASC9 promotes tumor growth and metastasis via modulating FZD6/Wnt/β-catenin signaling pathway in bladder cancer," *J. Exp. Clin. Cancer Res.*, vol. 39, no. 1, Jul. 2020, doi: 10.1186/s13046-020-01624-9.
[17] W. Wu, Y. Shen, J. Sui, C. Li, S. Yang, S. Xu, M. Zhang, L. Yin, Y. Pu, and G. Liang, "Integrated analysis of long non-coding RNA competing interactions revealed potential biomarkers in cervical cancer: Based on a public database," *Mol. Med. Rep.*, vol. 17, no. 6, pp. 7845–7858, Apr. 2018.
[18] 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 Bioinf.*, vol. 20, no. 1, Jul. 2019, doi: 10.1186/s12859-019-2985-0.
[19] Q. Yuan, X. Guo, Y. Ren, X. Wen, and L. Gao, "Cluster correlation based method for lncRNA-disease association prediction," *BMC Bioinf.*, vol. 21, no. 1, pp. 1–14, May 2020, doi: 10.1186/s12859-020-3496-8.
[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 Bioinf.*, vol. 21, no. 1, pp. 1–18, Mar. 2020, doi: 10.1186/s12859-020-3458-1.
[21] M. Zeng, C. Lu, F. Zhang, Y. Li, F.-X. Wu, Y. Li, and M. Li, "SDLDA: LncRNA-disease association prediction based on singular value decomposition and deep learning," *Methods*, vol. 179, pp. 73–80, Jul. 2020.
[22] H. Tan, Q. Sun, G. Li, Q. Xiao, P. Ding, J. Luo, and C. Liang, "Multiview consensus graph learning for lncRNA–disease association prediction," *Frontiers Genet.*, vol. 11, p. 89, Feb. 2020, doi: 10.3389/fgene.2020.00089.
[23] W. Lan, D. Lai, Q. Chen, X. Wu, B. Chen, J. Liu, J. Wang, and Y.-P.-P. Chen, "LDICDL: LncRNA-disease association identification based on collaborative deep learning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Oct. 30, 2020, doi: 10.1109/TCBB.2020.3034910.
[24] Q. Chen, D. Lai, W. Lan, X. Wu, B. Chen, Y.-P.-P. Chen, and J. Wang, "ILDMSF: Inferring associations between long non-coding RNA and disease based on multi-similarity fusion," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Aug. 20, 2019, doi: 10.1109/TCBB.2019.2936476.
[25] W. Lan, M. Li, K. Zhao, J. Liu, F. X. Wu, Y. Pan, and J. Wang, "LDAP: A Web server for lncRNA-disease association prediction," *Bioinformatics*, vol. 33, no. 3, pp. 458–460, Feb. 2017.
[26] X. Chen, "Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA," *Sci. Rep.*, vol. 5, no. 1, p. 13186, Aug. 2015, doi: 10.1038/srep13186.
[27] J. Li, H. Zhao, Z. Xuan, J. Yu, X. Feng, B. Liao, and L. Wang, "A novel approach for potential human LncRNA-disease association prediction based on local random walk," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Aug. 14, 2019, doi: 10.1109/TCBB.2019.2934958.
[28] Y. Wang, G. Yu, J. Wang, G. Fu, M. Guo, and C. Domeniconi, "Weighted matrix factorization on multi-relational data for LncRNA-disease association prediction," *Methods*, vol. 173, pp. 32–43, Feb. 2020.
[29] X. Chen, Z.-H. You, G.-Y. Yan, and D.-W. Gong, "IRWRLDA: Improved random walk with restart for lncRNA-disease association prediction," *Oncotarget*, vol. 7, no. 36, pp. 57919–57931, Sep. 2016.
[30] C. Park, N. Yu, I. Choi, W. Kim, and S. Lee, "LncRNAtor: A comprehensive resource for functional investigation of long non-coding RNAs," *Bioinformatics*, vol. 30, no. 17, pp. 2480–2485, Sep. 2014.
[31] S. Jain, S. Shukla, and R. Wadhvani, "Dynamic selection of normalization techniques using data complexity measures," *Expert Syst. Appl.*, vol. 106, pp. 252–262, Sep. 2018, doi: 10.1016/j.eswa.2018.04.008.
[32] K. Fang, Y. Zhou, and P. Ma, "An adaptive sequential experiment design method for model validation," *Chin. J. Aeronaut.*, vol. 33, no. 6, pp. 1661–1672, Jun. 2020, doi: 10.1016/j.cja.2019.12.026.
[33] A. Liu, X. Deng, L. Ren, Y. Liu, and B. Liu, "An inverse power generation mechanism based fruit fly algorithm for function optimization," *J. Syst. Sci. Complex.*, vol. 32, no. 2, pp. 634–656, Apr. 2019, doi: 10.1007/s11424-018-7250-5.
[34] Z. Zhang, C. Huang, H. Huang, S. Tang, and K. Dong, "An optimization method: Hummingbirds optimization algorithm," *J. Syst. Eng. Electron.*, vol. 29, no. 2, pp. 386–404, Apr. 2018, doi: 10.21629/JSEE.2018.02.19.
[35] C. Wang and W. Song, "A modified particle swarm optimization algorithm based on velocity updating mechanism," *Ain Shams Eng. J.*, vol. 10, no. 4, pp. 847–866, Dec. 2019, doi: 10.1016/j.asej.2019.02.006.
[36] E.-U. Haq, I. Ahmad, A. Hussain, and I. M. Almanjahie, "A novel selection approach for genetic algorithms for global optimization of multimodal continuous functions," *Comput. Intell. Neurosci.*, vol. 2019, pp. 1–14, Dec. 2019, doi: 10.1155/2019/8640218.
[37] S. Li and Y. Sun, "A novel numerical optimization algorithm inspired from garden balsam," *Neural Comput. Appl.*, vol. 32, no. 22, pp. 16783–16794, Nov. 2020, doi: 10.1007/s00521-018-3905-3.
[38] T. D. C. Negri, W. A. L. Alves, P. H. Bugatti, P. T. M. Saito, D. S. Domingues, and A. R. Paschoal, "Pattern recognition analysis on long noncoding RNAs: A tool for prediction in plants," *Briefings Bioinf.*, vol. 20, no. 2, pp. 682–689, Mar. 2019.
[39] 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*, vol. 9, no. 7, p. 345, Jul. 2018, doi: 10.3390/genes9070345.
[40] S. Peng, Q. Hu, J. Dang, and W. Wang, "Optimal feasible step-size based working set selection for large scale SVMs training," *Neurocomputing*, vol. 407, pp. 366–375, Sep. 2020, doi: 10.1016/j.neucom.2020.05.054.

BO WANG received the M.S. degree in computer application technology from the Institute of Computer and Control Engineering, Qiqihar University, Qiqihar, China, in 2004. He is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Harbin Engineering University, Harbin, China. He is also a Visiting Scholar with the Aerospace Software Engineering Center, Harbin Institute of Technology, Harbin, in 2013. His research interests include bioinformatics and multivariate statistical analysis.

JING ZHANG received the B.S., M.S., and Ph.D. degrees in computer science and technology from Harbin Engineering University, in 1987, 1998, and 2005, respectively. From 2003 to 2004, she was a Visiting Scholar with Melbourne University. From 2005 to 2007, she was a Postdoctoral Researcher in the area of image processing with the Harbin Institute of Technology, Harbin. From 2015 to 2016, she was an Academic Visitor with the University of York, where she was involved in research, under the guidance of Prof. E. Hancock. She has published 100 papers in journals, edited books, and refereed conferences. Her research interests include image processing and virtual reality.