IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 28, 2020, accepted February 12, 2020, date of publication February 17, 2020, date of current version February 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2974624

Logistic Regression Analysis for LncRNA-Disease Association Prediction Based on Random Forest and Clinical Stage Data

BO WANG[1,3] **AND JING ZHANG**[2,1]
[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]College of Computer and Control, Qiqihar University, Qiqihar 161006, China

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

This work was supported in part by the National Natural Science Foundation of China (NSFC) (2017–2020) under Grant 51679058, in part by the Shandong Natural Science Foundation in China (2020–2022) under Grant ZR2019LZH005, and in part by the Scientific Research Project of Qiqihar University (the Fundamental Research Funds in Heilongjiang Provincial Universities) under Grant 135109249 and Grant 135109241.

ABSTRACT An increasing amount of studies have found that LncRNA plays an important role in various life processes of the body. In current prediction research on lncRNA-disease associations, correlation analysis of disease prognosis is overlooked. In this study, a logistic regression prediction model based on tumor clinical stage data and the expression quantity of lncRNA transcript is constructed. The proposed model is based on unknown human lncRNA-disease associations combining with the clinical stage data. Firstly, the importance of the characteristic variable is calculated by the proposed CVS_gC-RF algorithm. Secondly, 95 lncRNAs, which are most closely related to prostate cancer, are calculated from 480 alternative lncRNAs by CASO and CVS_e-CS-CF. On the basis of the above 95 lncRNAs, the CSPA-PL algorithm is used to select a further 22 lncRNAs that are most closely related to the tumor clinical stage for prostate cancer. Finally, 22 lncRNAs are used to construct a logistic regression prediction model. Additionally, this method is applied to lung cancer data; 16 lncRNAs are selected to construct a logistic regression prediction model for lung cancer. Experimental results show that the best results for ROC Area, the accuracy and recall rate of the prediction model are achieved by the proposed method for prostate cancer and lung cancer, which provides a promising basis for subsequent prediction studies of lncRNA-disease associations.

INDEX TERMS LncRNA-disease association, random forest, logistic regression analysis, clinical stage data.

I. INTRODUCTION

Long non-coding RNA (lncRNA) is non-coding RNA with more than 200 nucleotides in length [1]. It has very important biological functions and is another important area in the bioinformatics field [2], [3]. Studies show that lncRNA is closely correlated with many diseases, such as lung cancer [4], [5], Alzheimer's disease [6], osteosarcoma [7], breast cancer [8], gastric cancer [9], colon cancer [10], prostate cancer [11], cervical cancer [12], etc. At present, more and more researchers are engaged in research in this area, which is an important molecular target in the diagnosis and treatment of disease. It is extremely important to study the relationship between lncRNA and the prognosis of cancer patients by utilizing clinical data. Current research in lncRNA is in the initial stages; people still know little about the deep mechanisms in the occurrence and development of cancer. Therefore, it is important to study lncRNA, which has a significant impact on the prognosis of cancer patients by using bioinformatics combined with clinical data.

Relevant research results in recent years are broadly divided into three types, as follows.

The first type is machine-learning-based methods and known disease-related lncRNAs. For instance, Yu *et al.* [13] proposed a new method called CFNBC based on the Naïve Bayes classifier to predict lncRNA-disease association. The novelty of CFNBC lies in the introduction of the item-based collaborative filtering algorithm and Naïve Bayes classifier,

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

which guarantee that CFNBC can be applied to predict potential lncRNA-disease associations efficiently without entirely relying on known miRNA-disease associations. Cui *et al.* [14] developed a novel model called BLM-NPAI for predicting lncRNA-disease associations. The main advantage of BLM-NPAI was that it could also make predictions using nearest neighbors for some lncRNAs and diseases without any association. Chen and Yan [15] used semi-supervised learning to predict the potential associations between lncRNAs and diseases, and proposed the first lncRNA-disease association prediction model (LRLSLDA) on the premise that similar functions of lncRNA tended to result in similar diseases. However, the model was too complex and exhibited high computational complexity. Meanwhile many parameters need to be selected in the calculation process. Huang *et al.* [16] improved the calculation of disease similarity based on the framework of LRLSLDA to further improve the prediction results, and presented a new method, ILNCSIM. This approach kept the general hierarchical structure information of disease DAGs and determined the disease similarity calculation based on an edge-based method. Finally, the prediction performance was improved to some extent, but there were still some limitations. For example, the similarity score in the model needs to be further optimized. The lack of unrecorded but real lncRNA-disease associations had a large impact on the model, and the integration of multiple types of data was lacking. Chen [17] built a new approach (KATZLDA) by integrating known lncRNA-disease associations, lncRNA expression profiles, lncRNA functional similarity, disease semantic similarity and Gaussian interaction profile kernel similarity to predict the potential LncRNA-disease associations. The biggest advantage of KATZLDA was that it could be effectively applied to new diseases and lncRNAs without any known associations. However, the learning network built by KATZLDA was based on the known correlation relationship, so this was limited by the known learning knowledge and had certain limitations in prediction. Zhao *et al.* [18] constructed a multi-source data set by integrating multidimensional data (genome, regulatory group and transcriptome), and proposed a Bayesian classification method using this multi-source data to predict the lncRNA–disease associations. Experimental results showed that this method successfully identified 707 lncRNAs related to human cancer. However, this method was a supervised classification algorithm which required a large number of negative cases, but these are difficult to obtain.

The second type is network-based methods. For instance, Li *et al.* [19] present a novel network consistency projection approach called NCPLDA for lncRNA-disease association prediction. The network was built by integrating the lncRNA-disease association probability matrix with the integrated disease similarity and lncRNA similarity. Zhou *et al.* [20] proposed a new method (RWRHLD) that built a heterogeneous network of lncRNA–disease associations, on which a random walk algorithm was executed. However, the limitation was that the incomplete coverage of the lncRNA crosstalk network and the lncRNA–disease associations could lead to inaccurate predictions. Liu *et al.* [21] established a bidirectional network of protein-coding genes (PCG) and lncRNA for the prostate cancer and protein interaction databases based on lncRNAs and PCG expression maps, and further realized lncRNA–disease association prediction based on this network. However, the method was limited by the incomplete protein interaction database, and its performance had some limitations.

The third type is RW-based methods (RW is short for random walk). For instance, Li *et al.* [22] proposed a prediction model called LRWHLDA for inferring LncRNA-disease association. LRWHLDA can be implemented in the case of lacking known lncRNA-disease associations by using an improved local random walk method. Yu *et al.* [23] used multidimensional heterogeneous data to construct lncRNA networks with similar functions and the disease ontology to construct disease networks. On this basis, BRWLDA was proposed to predict the lncRNA-disease associations. BRWLDA improved the random walk model and the prediction performance to some extent.

To summarize, the limitations of the current research were described by the review [24] and the aforementioned discussions. Current studies have ignored correlation analysis of clinical prognosis, concerns of the prediction model have been limited to a single lncRNA forecast. The clinical prognosis of the disease associated with lncRNA information is rarely involved, such as tumor clinical stage, tumor pathological stage, survival time, disease status, family history of genetic diseases, and so on.

In this study, a logistic regression prediction model of lncRNA-disease associations based on the tumor clinical stage data was constructed. Three kinds of circular allelism operations ($\Gamma_{center}(\Theta^{sub}_{[a,b]})$, $\Gamma_{X-axis}(\Theta^{sub}_{[a,b]})$, $\Gamma_{Y-axis}(\Theta^{sub}_{[a,b]})$) were proposed for the prediction model. The calculation of the significance of characteristic variables based on random forests was proposed, and the selection algorithm for the characteristic variables was given. Finally, the clinical stage prediction algorithm of cancer-associated lncRNA was implemented using the simplified characteristic variables. Experimental results showed that the proposed method had a higher predictive performance.

II. MATERIALS AND METHODS

A. LNCRNA DATA

The lncRNA expression data for prostate cancer was obtained from the lncRNAtor database [25]. A total of 220 samples were obtained (denoted by $S^{normal \cup tumor} = \{S_1, \cdots S_{220}\}$), including 44 normal samples (denoted by $S^{normal} = \{S_1, \cdots, S_{44}\}$) and 176 cancer samples (denoted by $S^{tumor} = \{S_{45}, \cdots, S_{220}\}$). Based on the differential expression P-value ($P \leq 0.001$) of lncRNA transcripts between S^{normal} and S^{tumor}, 480 lncRNA transcripts with significant differences (denoted by $Lr_1, Lr_2, \cdots, Lr_{480}$ in ascending order of P-value) were obtained. Of these, 480 lncRNA transcripts

TABLE 1. Aliquot barcode of clinical data(220).

status	aliquot barcode							
	project	tss	sample	vial	portion	analyte	plate	center
normal	TCGA	CH,EJ, G9,HC,	11	A,B	01,02	R	1580,1789, 1858,1965, 2118,2263	7
tumor	TCGA	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

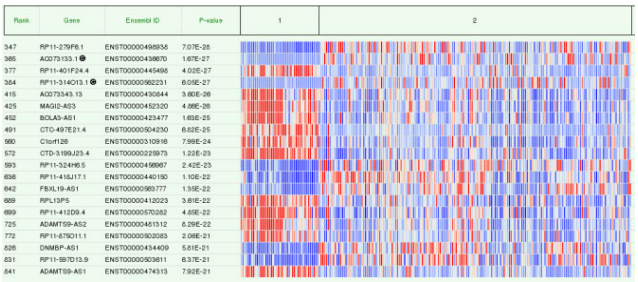


FIGURE 1. Differential expression of LncRNA genes (only the top 20 genes listed, 480 in total).

were denoted by $Lr = \{Lr_1, Lr_2, \cdots, Lr_i\ (1 \le i \le 480)\}$, Lr^{sub} was the subset of $Lr(Lr^{sub} \subseteq Lr)$ and the expression of Lr on S_i was denoted by $Lr^S = \left\{Lr_1^{S_i}, Lr_2^{S_i}, \cdots, Lr_{480}^{S_i}(1 \le i \le 220)\right\}$. Figure 1 shows the details of the top 20 lncRNA transcripts in the 480 transcripts, where each row represents one lncRNA transcript. The first column is the rank of the transcript, the second column is ensemble gene ID, the third column is the gene name, the fourth column is the ensemble transcript ID, the fifth column contains the P values of the differential expression between normal samples and cancer samples. Each column after the fifth column is the expression quantity of the transcript in the sample. These transcripts were the more pronounced differences between normal and cancer samples.

B. CLINICAL DATA

Clinical data associated with $S^{normal \cup tumor}$ were obtained from the TCGA database (https://cancergenome.nih.gov). The aliquot barcode of the clinical data in S^{normal} (size 44) and S^{tumor} (size 176) are presented in Table 1. Each S_i contains 70 clinical reference values. Some of these were retained as follows: barcode and sample type of $S^{normal \cup tumor}$ (denoted by $P^{normal \cup tumor} = \{P_1, \cdots, P_{220}\}$), tumor clinical stage of S^{tumor} (denoted by $CT^{tumor} = \{CT_1, \cdots, CT_{176}\}$). Of these, the barcode was used to correlate the clinical data with the lncRNA data, the sample type was used to select the characteristic variables of the lncRNA-disease associations, and the tumor clinical stage was used to predict lncRNA with significant impact on the prognosis of cancer patients combined with the clinical data. In this case, the clinical stage (CTNM) was performed using the TNM stage system, where T represents the tumor size, N represents lymph node metastasis, and M represents distant metastasis. The distribution of CT^{tumor} (size 176) associated with S^{tumor} is shown in Table 2. As can be seen from this table, 118 effective clinical stage data and 58 invalid clinical stage data were obtained. Variance analysis of $\{T_1 \cup T_2\}$ and $\{T_1 \cup T_2 \cup T_3\}$ showed that $\{T_1 \cup T_2\}$ has a better-balanced distribution and is conducive to machine classification learning. Finally, 104 available data (denoted by $CT^{\xi - tumor} = CT^{tumor} \cap \{T_1 \cup T_2\} = \left\{CT_1^{\xi}, CT_2^{\xi}, \cdots, CT_{104}^{\xi}\right\}$) were selected from 118 valid clinical stage data.

The following two matrices were constructed by combining lncRNA data and clinical data for the prediction study in this paper.

(a) The matrix M^{CV} of the characteristic variables is shown in (1), which contains 480 columns of characteristic variables, 1 categorical variable column, and 220 rows of sample data. After the selection algorithm, the λ lncRNAs that were most closely related to prostate cancer were screened from 480 characteristic variables.

$$\begin{aligned} M^{CV} &= Lr^S \bowtie P^{normal \cup tumor} \\ &= \begin{bmatrix} Lr_1^{S_1} & Lr_2^{S_1} & \cdots & Lr_n^{S_1} & P_1 \\ Lr_1^{S_2} & Lr_2^{S_2} & \cdots & Lr_n^{S_2} & P_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Lr_1^{S_m} & Lr_2^{S_m} & \cdots & Lr_n^{S_m} & P_m \end{bmatrix} \\ &\quad (n = 480, m = 220) \\ P_{1 \le i \le 44} &= [normal, tumor] \times \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \\ P_{45 \le i \le 220} &= [normal, tumor] \times \begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{aligned} \tag{1}$$

(b) The matrix M^{PP} of prognosis prediction is shown in (2), which contains λ columns of characteristic variables, 1 categorical variable column, and 104 rows of sample data.

$$\begin{aligned} M^{PP} &= Lr^{*S} \bowtie CT^{\xi - tumor} \\ &= \begin{bmatrix} Lr_1^{*S_1} & Lr_2^{*S_1} & \cdots & Lr_\lambda^{*S_1} & CT_1^{\xi} \\ Lr_1^{*S_2} & Lr_2^{*S_2} & \cdots & Lr_\lambda^{*S_2} & CT_2^{\xi} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Lr_1^{*S_m} & Lr_2^{*S_m} & \cdots & Lr_\lambda^{*S_m} & CT_m^{\xi} \end{bmatrix} \\ &\quad \left(Lr^{*S_i} = Lr^{S_{(i-45)}}, m = 104\right) \end{aligned} \tag{2}$$

TABLE 2. Distribution of CT^{tumor}.

stage type of CTNM	number	standard deviation	
		$T_1 \cup T_2$	$T_1 \cup T_2 \cup T_3$
T_1	56	4	18.20867
T_2	48		
T_3	14	-	
available	104	-	-
effective	118	-	-
Unknown	51	-	-
Not Available	7	-	-
invalid	58	-	-
total	176	-	-

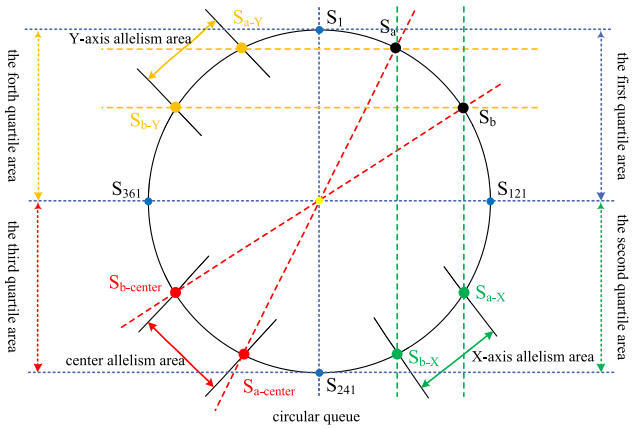


FIGURE 2. Schematic of allelism operation on the circular queue.

C. CASO METHOD

The circular allelism subarea operation (abbreviated to CASO) for characteristic variable selection is described below. The 480 $Lr_i s$ in M^{CV} formed a circular queue Θ. According to the importance of each Lr_i, the Lr_i in descending order are evenly clockwise distributed on the ring Θ. Θ contains 480 nodes in descending order of importance (denoted by $\Theta = \{S_1, S_2, \cdots, S_{480}\}$). The $Significance(Lr_i)$ is the importance of Lr_i. The descending rank of $Significance(Lr_i)$ is $j = D - rank(Significance(Lr_i))$. Here, the node S_j is Lr_i, and the subset of Θ is denoted by $\Theta^{sub}_{[a,b]} = \{S_a, \cdots, S_b\}(1 < a < b < 480)(\Theta^{sub}_{[a,b]} \subset \Theta)$. The descending queue formed with Lr^{sub} in descending order according to $Significance(Lr_i)$ is $Q^{Dec}(Lr^{sub})$.

The circular queue Θ is formed as shown in Figure 2. The descending order from S_1 to S_{480} is evenly clockwise distributed on Θ. The center symmetric points of S_a and S_b are $S_{a-center}$ and $S_{b-center}$. The X-axis symmetric points of S_a and S_b are S_{a-X} and S_{b-X}. The Y-axis symmetric points of S_a and S_b are S_{a-Y} and S_{b-Y}.

Figure 2 illustrates the following. The red line, area and data represent the center allelism operation. The green line, area and data represents the X-axis allelism operation. The yellow line, area and data represents the Y-axis allelism operation. The circular queue Θ is divided into four areas (the first quartile area, the second quartile area, the third quartile area, and the fourth quartile area).

Definition 1 (Center Allelism Operation $\Gamma_{center}(\Theta^{sub}_{[a,b]})$): $\Gamma_{center}(\Theta^{sub}_{[a,b]})$ is defined as $\Theta^{sub}_{[a,b]} \cup \Theta^{sub-center}_{[a,b]}$. $\Theta^{sub-center}_{[a,b]}$ is the center allelism area of $\Theta^{sub}_{[a,b]}$. The set of central symmetric points of $\Theta^{sub}_{[a,b]} = \{S_a, \cdots, S_b\}$ is denoted by $\Theta^{sub-center}_{[a,b]} = \{S_{a-center}, L, \cdots, S_{b-center}\}$.

Definition 2 (X-Axis Allelism Operation $\Gamma_{X-axis}(\Theta^{sub}_{[a,b]})$): $\Gamma_{X-axis}(\Theta^{sub}_{[a,b]})$ is defined as $\Theta^{sub}_{[a,b]} \cup \Theta^{sub-X-axis}_{[a,b]}$. $\Theta^{sub-X-axis}_{[a,b]}$ is the X-axis allelism area of $\Theta^{sub}_{[a,b]}$. The set of X-axis symmetric points of $\Theta^{sub}_{[a,b]} = \{S_a, \cdots, S_b\}$ is denoted by $\Theta^{sub-X-axis}_{[a,b]} = \{S_{a-X}, \cdots, S_{b-X}\}$.

Definition 3 (Y-Axis Allelism Operation $\Gamma_{Y-axis}(\Theta^{sub}_{[a,b]})$): $\Gamma_{Y-axis}(\Theta^{sub}_{[a,b]})$ is defined as $\Theta^{sub}_{[a,b]} \cup \Theta^{sub-Y-axis}_{[a,b]}$. $\Theta^{sub-Y-axis}_{[a,b]}$ is the Y-axis allelism area of $\Theta^{sub}_{[a,b]}$. The set of Y-axis symmetric points of $\Theta^{sub}_{[a,b]} = \{S_a, \cdots, S_b\}$ is denoted by $\Theta^{sub-Y-axis}_{[a,b]} = \{S_{a-Y}, \cdots, S_{b-Y}\}$.

The red area in Figure 2 is $\Gamma_{center}(\Theta^{sub}_{[a,b]})$, which is located in the first quartile area and the third quartile area. The green area in Figure 2 is $\Gamma_{X-axis}(\Theta^{sub}_{[a,b]})$, which is located in the first quartile area and the second quartile area.

The yellow area in Figure 2 is $\Gamma_{Y-axis}(\Theta^{sub}_{[a,b]})$, which is located in the first quartile area and the fourth quartile area.

Additionally, $\Theta^{sub-X-axis}_{[a,b]}$ is closer to $\Theta^{sub}_{[a,b]}$, $\Theta^{sub-center}_{[a,b]}$ is centered on $\Theta^{sub}_{[a,b]}$, and $\Theta^{sub-Y-axis}_{[a,b]}$ is far from $\Theta^{sub}_{[a,b]}$. The above $\Gamma_{center}(\Theta^{sub}_{[a,b]})$, $\Gamma_{X-axis}(\Theta^{sub}_{[a,b]})$, and $\Gamma_{Y-axis}(\Theta^{sub}_{[a,b]})$ constitute the circular subset Θ required by the next algorithm.

D. RANDOM FOREST

Random forest (abbreviated to RF) is an enhanced classifier constructed by multiple decision trees. In the process of building the decision tree, it is necessary to order the importance of variables. Since a random forest has a large number of decision trees, the importance obtained from each decision tree could be integrated to obtain the final importance rank of the variables. The selection of characteristic variables is carried out according to the order of the variables, which is more stable and reliable than a single decision tree. In the selection of M^{CV} characteristic variables based on RF, RF contained α trees (denoted by $T = \{T_1, \cdots, T_i, \cdots, T_\alpha\}$), 480 Lrs (denoted by $Lr = \{Lr_1, \cdots, Lr_i, \cdots, Lr_{480}\}$) in M^{CV} are the characteristic variable set, and 220 Ps (denoted by $P = \{P_1, \cdots, P_i, \cdots, P_{220}\}$) in M^{CV} are the classified variable set.

E. CVS_GC-RF ALGORITHM

The significance computing of the characteristic variables based on RF (abbreviated to CVS_gC-RF) is given in Algorithm 1. Here, OOB means out-of-bag. Records are extracted from the original data to construct the training set for decision tree learning. Because this process uses sampling with replacement, some samples are not included, termed out-of-bag. On average, 37% of the data is not selected in each

sampling with replacement, which is often used to validate the constructed decision tree model. If the characteristic variable Lr_i upseted on OOB has no effect on the result of the decision tree, then Lr_i is deemed to be not important. If the reverse is true, then Lr_i is very important.

Algorithm 1 CVS$_g$C-RF(Lr^{sub}, τ, η)

$RF(Lr^{sub}, \tau, \eta)$;
for $\forall T_k \in T$ **do**
  $OOB(T_k) = MeanDecreaseAccuracy^{St}(OOB \in T_k)$;
  for $\forall Lr_i \in Lr^{sub}$ **do**
    $T_k^{Lr_i} \leftarrow Random_{upset}(\{Lr_i^{S_1}, Lr_i^{S_2}, \cdots, Lr_i^{S_{220}}\})$;
    $OOB(T_k^{Lr_i}) = MeanDecreaseAccuracy^{St}(OOB \in T_k^{Lr_i})$;
    $Significance_{T_k}(Lr_i) = OOB(T_k^{Lr_i}) - OOB(T_k)$;
    $Lr^{sub} \leftarrow Lr^{sub} - Lr_i$;
  end for
  $T \leftarrow T - T_k$;
end for
$Significance(Lr_i) = \frac{\sum_{k=1}^{\tau} significance_{T_k}(Lr_i)}{\tau}$;
$Q^{Dec}(Lr^{sub}) \leftarrow Sort(Lr^{sub})^{\neg significance(Lr_i)}$;
return $Q^{Dec}(Lr^{sub})$;

The CVS$_g$C-RF algorithm is shown in Algorithm 1. In this algorithm, $RF(Lr^{sub}, \tau, \eta)$ is a random forest containing τ decision trees by being trained on Lr^{sub}. τ is the number of decision trees contained in a random forest. η is the number of random characteristic variables contained in each partition ($\eta = \left\lfloor \sqrt{RF_{cv-number}(Lr^{sub})} + 0.5 \right\rfloor$). $RF_{cv-number}(Lr^{sub})$ is the number of characteristic variables in Lr^{sub}. The relationship between τ and $RF_{cv-number}(Lr^{sub})$ is $\tau \propto RF_{cv-number}(Lr^{sub})$.

$MeanDecreaseAccuracy^{St}(OOB)$ is a standardized prediction error rate for OOB data. $Random_{upset}(\{Lr_i^{S_1}, Lr_i^{S_2}, \cdots, Lr_i^{S_{220}}\})$ is the random upset operation on a characteristic variable Lr_i. All characteristic variables in Lr^{sub} are arranged in descending order according to $Significance(Lr_i)$ (denoted by $Sort(Lr^{sub})^{\neg significance(Lr_i)}$). Further, a queue ($Q^{Dec}(Lr^{sub})$) is formed.

F. DISCUSSION OF CVS$_E$

A good characteristic variable selection (abbreviated to CVS$_e$) algorithm must possess both global selectivity and local stability. Global selectivity and local stability are mutually restricted. For example, the larger the selection range, the more complex the mutual relations among the characteristic variables, and bidirectional influence relations coexist. Local stability is needed for adjustment, but is limited by the selection range, which leads to insufficient coverage of the correlation between characteristic variables. At this point, an increase in global selectivity is required. Therefore, how to adjust the global selectivity and local stability of an algorithm is very important, and could affect the performance of CVS$_e$.

G. CVS$_E$-CS-CF ALGORITHM

CVS$_e$ combines CASO and CVS$_g$C-RF (named CVS$_e$-CS-CF). CVS$_e$-CS-CF algorithm for M^{CV} is provided in Algorithm 2. The Lr in M^{CV} is selected by the CVS$_e$-CS-CF algorithm. The λ $Lr_i s$ most closely related to categorical variables is selected in 480 $Lr_i s$. In order to possess both global selectivity and local stability, the CVS$_e$-CS-CF algorithm is divided into a primary stage, a stable stage and a run-off stage. The primary stage can guarantee global selectivity, the stable stage guarantees local stability, and the run-off stage guarantees the final selection result by combining global selectivity and local stability.

Algorithm 2 CVS$_e$-CS-CF(Lr, $\tau_{0\sim9}$, $\eta_{0\sim9}$, Δ, Ω, a, b)

$\Theta = CVS_gC - RF(Lr, \tau_0, \eta_0)$;
$Set_{penalty} \leftarrow \emptyset$, $Set_{shock} \leftarrow \emptyset$;
while $distance([a, \Delta]) \geq \Omega$ **do**
  if $[a, b]$ by Δ is dimidiate **then**
    $Q_{++f} = CVS_g$C-RF $(\{\Gamma_{center}(\Theta^{sub}_{[a,\Delta]}) \, or \, \Gamma_{center}(\Theta^{sub}_{[\Delta+1,b]})\}, \tau_{++f}, \eta_{++f})$;
    $Q_{++f} = CVS_g$C-RF $(\{\Gamma_{X-axis}(\Theta^{sub}_{[a,\Delta]} \, or \, \Gamma_{X-axis}(\Theta^{sub}_{[\Delta,b+1]})\}, \tau_{++f}, \eta_{++f})$;
    $Q_{++f} = CVS_g$C-RF $(\{\Gamma_{Y-axis}(\Theta^{sub}_{[a,\Delta]} \, or \, \Gamma_{Y-axis}(\Theta^{sub}_{[\Delta,b+1]})\}, \tau_{++f}, \eta_{++f})$;
  else
    $Q_{++f} = CVS_gC - RF(\Gamma_{center}(\Theta^{sub}_{[a,b]}), \tau_{++f}, \eta_{++f})$;
    $Q_{++f} = CVS_gC - RF(\Gamma_{X-axis}(\Theta^{sub}_{[a,b]}), \tau_{++f}, \eta_{++f})$;
    $Q_{++f} = CVS_gC - RF(\Gamma_{Y-axis}(\Theta^{sub}_{[a,b]}), \tau_{++f}, \eta_{++f})$;
  end if
  $\Delta = \lceil \frac{\Delta}{2} \rceil$;
end while
for each $Q_f (f \in [1, \max(f)])$ **do**
  for each Lr_i in $Q_f \cup \Theta_{[1,2\Omega]} - Q_f \cap \Theta_{[1,2\Omega]}$ **do**
    if $Lr_i \in \Theta_{[1,2\Omega]}$ and $Lr_i \notin Q_f$ **then**
      $Set_{penalty} \leftarrow Lr_i$;
    end if
    if $Lr_i \notin \Theta_{[1,2\Omega]}$ and $Lr_i \in Q_{f\leq\Omega}$ **then**
      $Set_{shock} \leftarrow Lr_i$;
    end if
  end for
end for
$\Theta^{select} \leftarrow \Theta_{[1,2\Omega]} \cup Set_{shock} - Set_{penalty}$;
$\lambda = number(\Theta^{select})$;
return Θ^{select}, λ;

Primary Stage: The 480 $Lr_i s$ are ranked in descending order of importance, with the top 2Ω $Lr_i s$ entering the candidate area and the remaining $480 - 2\Omega$ $Lr_i s$ entering the observation area. Since the Lr_i entering the candidate area is selected in the global range of 480, global selectivity is obtained. However, the larger the scope, the more complex the relationship of Lr_i will become, and the bidirectional influence relationship exists. In the following stable stage,

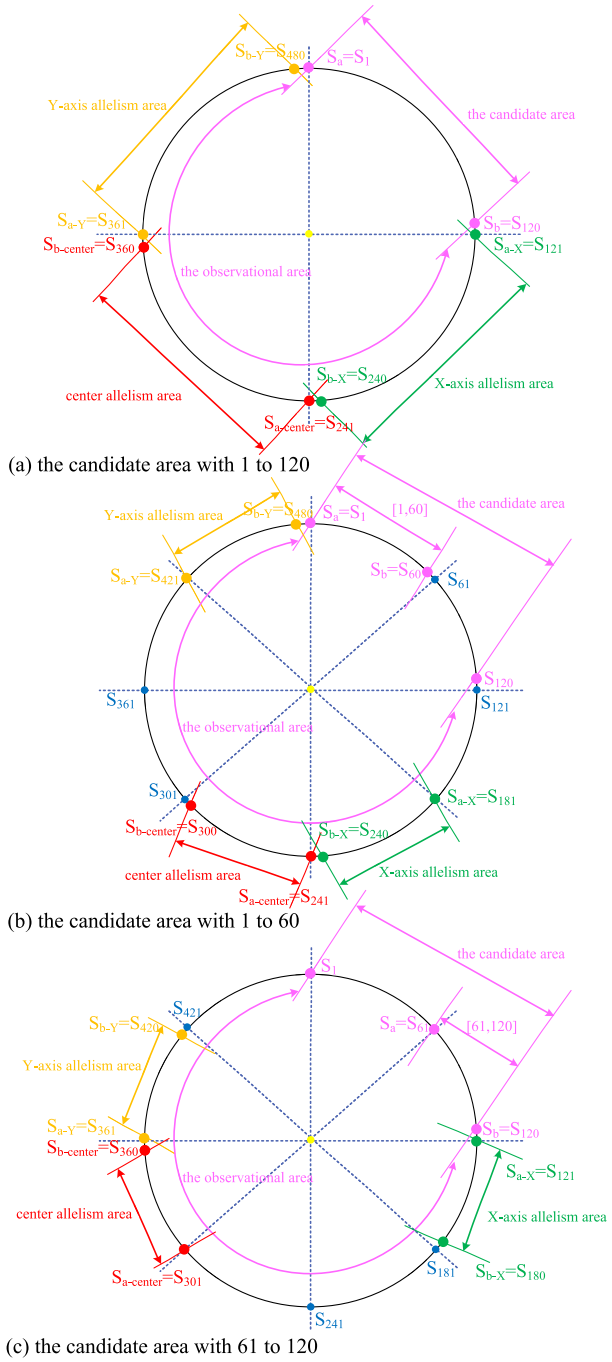


FIGURE 3. A CASO process of Lr_i in candidate area.

local stability is used to address this problem. The detailed process is described in steps 1-2 of Algorithm 2.

Stable Stage: CASO is performed for each Lr_i that entered the candidate area during the primary stage. The detailed execution process is shown in Figure 3. In order to facilitate the execution of the algorithm, the endpoints of the three allelism areas are adjusted to $S_{b-center}-1$, $S_{b-X}-1$, and $S_{b-Y}-1$. The center allelism area, X-axis allelism area and Y-axis allelism area are located in the observation area. The candidate area and the three allelism areas of the above observation areas are rearranged in descending order of importance. (This is the CVS_gC-RF algorithm that was discussed previously.) Each Lr_i that is screened from the candidate area into the observation area is added to the penalty set. The top Ω $Lr_i s$ that are screened from the observation area into the candidate area are added to the shock set. Each Lr_i in $Set_{penalty}$ has a potential risk of poor stability. Each Lr_i in Set_{shock} has strong reactivation activity. The detailed process is described in steps 3-24 of Algorithm 2.

Run-Off Stage: $Set_{penalty}$ and Set_{shock} are used to update each Lr_i in the candidate area. Remove Lr_i in $Set_{penalty}$ from the candidate area and add it to Set_{shock} in the candidate area. The updated candidate area is the result of characteristic variable selection (denoted by Θ^{select}). The number of characteristic variables in Θ^{select} is λ. The detailed process is described in steps 25-27 of Algorithm 2.

In Figure 3, Ω is set to 60, the candidate area is $[S_1, S_{120}]$, and the observation area is $[S_{121}, S_{480}]$. Δ, a and b are set to 120, 1, and 120 in Figure 3(a). The area where the CASO is performed is $[S_1, S_{120}]$ (shown in pink). The center allelism area is $\Theta^{sub-center} = [S_{241}, S_{360}]$ (shown in red). The X-axis allelism area is $\Theta^{sub-X-axis} = [S_{121}, S_{240}]$ (shown in green). The Y-axis allelism area is $\Theta^{sub-Y-axis} = [S_{361}, S_{480}]$ (shown in yellow). Finally, the three allelism area combinations are $[S_1, S_{120}] \cup [S_{241}, S_{360}]$, $S_1, S_{120}] \cup [[S_{121}, S_{240}]$, and $[S_1, S_{120}] \cup [S_{361}, S_{480}]$. Δ, a and b are set to 60, 1 and 60 in Figure 3(b). The area where the CASO is performed is $[S_1, S_{60}]$ (shown in pink). The center allelism area is $\Theta^{sub-center} = [S_{241}, S_{300}]$ (shown in red). The X-axis allelism area is $\Theta^{sub-X-axis} = [S_{181}, S_{240}]$ (shown in green). The Y-axis allelism area is $\Theta^{sub-Y-axis} = [S_{421}, S_{480}]$ (shown in yellow). Finally, the three allelism area combinations are $[S_1, S_{60}] \cup [S_{241}, S_{300}]$, $[S_1, S_{60}] \cup [S_{181}, S_{240}]$, and $[S_1, S_{60}] \cup [S_{421}, S_{480}]$. Δ, a and b are set to 60, 61 and 120 in Figure 3(c). The area where the CASO is performed is $[S_{61}, S_{120}]$ (shown in pink). The center allelism area is $\Theta^{sub-center} = [S_{301}, S_{360}]$ (shown in red). The X-axis allelism area is $\Theta^{sub-X-axis} = [S_{121}, S_{180}]$ (shown in green). The Y-axis allelism area is $\Theta^{sub-Y-axis} = [S_{361}, S_{420}]$ (shown in yellow). Finally, the three allelism area combinations are $[S_{61}, S_{120}] \cup [S_{301}, S_{360}]$, $[S_{61}, S_{120}] \cup [S_{121}, S_{180}]$, and $[S_{61}, S_{120}] \cup [S_{361}, S_{420}]$.

H. CSPA-PL

The CVS_e-CS-CF algorithm selected λ lncRNAs that were most closely related to prostate cancer. Next, λ lncRNAs are correlated with the prognosis data of the tumor clinical stage, and a logistic regression model is adopted to propose a clinical stage prediction algorithm for cancer-associated lncRNA(abbreviated to CSPA-PL). Concerning the related operations of Θ^{select}, CSPA-PL is divided into an inspection stage and an optimization stage. CSPA-PL is described in Algorithm 3.

Algorithm 3 CSPA-PL($\Theta^{select}, \lambda, M^{PP}$)

1: $\Theta^{select}_{sub-pre} = \Theta^{select}_{\left[1,\left\lfloor\frac{\lambda}{2}\right\rfloor\right]}$, $\Theta^{select}_{sub-rear} = \Theta^{select}_{\left[\left\lfloor\frac{\lambda}{2}\right\rfloor+1,\lambda\right]}$;
2: $Z^{pre}_{[1,n]} \leftarrow Logistic - Step(\pi_{\Theta^{select}_{sub-pre}}(M^{PP}))$;
3: $\{\Upsilon^{pre}\} \leftarrow minimum(Z^{pre}_{[1,n]})\Big| AIC$;
4: $Z^{rear}_{[1,m]} \leftarrow Logistic - Step(\pi_{\Theta^{select}_{sub-rear}}(M^{PP}))$
5: $minimum(Z^{rear}_{[1,m]})\Big| AIC$;
6: **for** each Lr_i in $minimum(Z^{rear}_{[1,m]})\Big| AIC$ **do**
7: **if** $Value_P(Lr_i) < 0.01$ **then**
8: $\{\Upsilon^{rear}\} \leftarrow Lr_i$;
9: **end if**
10: **end for**
11: $\{Buffer - pool_i\} = \{\Upsilon^{pre}\} \cup \{\Upsilon^{pre}\} \times \{\Upsilon^{rear}\}$;
12: **for** each $Buffer - pool_i$ in $Buffer - pool$ **do**
13: $Buffer{-}pool^*_i \leftarrow Logistic{-}Step(\pi_{Buffer-pool_i}(M^{PP}))$
14: $\Phi_i \leftarrow Logistic(\pi_{Buffer-pool^*_i}(M^{PP}))$;
15: **end for**
16: $\Phi_{optimal} \leftarrow maximum(\Phi_i)|Accuracyrate$;
17: **return** $\Phi_{optimal}$;

Inspection Stage: Before being applied to the clinical stage prediction model, the λ $Lr_i s$ most closely associated with prostate cancer need to be inspected. The detailed process is described in steps 1-11 of Algorithm 3. In the inspection stage, Θ^{select} is divided into $\Theta^{select}_{\left[1,\left\lfloor\frac{\lambda}{2}\right\rfloor\right]}$ and $\Theta^{select}_{\left[\left\lfloor\frac{\lambda}{2}\right\rfloor+1,\lambda\right]}$, and then $Logistic - Step()$ is performed on $\Theta^{select}_{\left[1,\left\lfloor\frac{\lambda}{2}\right\rfloor\right]}$ and $\Theta^{select}_{\left[\left\lfloor\frac{\lambda}{2}\right\rfloor+1,\lambda\right]}$. $Logistic - Step()$ employs stepwise selection of variables for the logistic regression model. $\pi_{\Theta^{select}_{sub-pre}}(M^{PP})$ is the projection of the variable $\Theta^{select}_{sub-pre}$ on M^{PP}. $\pi_{\Theta^{select}_{sub-rear}}(M^{PP})$ is the projection of the variable $\Theta^{select}_{sub-rear}$ on M^{PP}. The process data of $Logistic - Step()$ on $\pi_{\Theta^{select}_{sub-pre}}(M^{PP})$ is recorded in $Z^{pre}_{[1,n]}$. The process data of $Logistic - Step()$ on $\pi_{\Theta^{select}_{sub-rear}}(M^{PP})$ is recorded in $Z^{rear}_{[1,m]}$. $\{\Upsilon^{pre}\} \leftarrow minimum(Z^{pre}_{[1,n]})\Big| AIC$ means that the set of Lr_i with the smallest AIC value in $Z^{pre}_{[1,n]}$ is put into $\{\Upsilon^{pre}\}$. Steps 5-10 indicate that the set of Lr_i with the smallest AIC values and significance less than 0.01 in $Z^{pre}_{[1,n]}$ is put into $\{\Upsilon^{rear}\}$. The Cartesian product of $\{\Upsilon^{pre}\}$ and $\{\Upsilon^{rear}\}$ is performed to form a buffer pool ($Buffer - pool$).

Optimization Stage (Steps 12-17 of Algorithm 3): Each element in $Buffer - pool$ constructs a logistic regression model, and the model with the highest accuracy (denoted by $maximum(\Phi_i)|Accuracyrate$) is selected as the optimal prediction model (denoted by $\Phi_{optimal}$).

III. RESULTS AND DISUSSION

A. PERFORMANCE EVALUATION OF CVS_GC-RF

When constructing RF in the CVS_gC-RF algorithm, the number of decision trees τ in the random forest has a large impact on the performance and efficiency of the algorithm. In order to determine the optimal value $\tau_{optimal}$, three groups of experiments were carried out under the premise of $\tau \propto RF_{cv-number}(Lr^{sub})$. Each group of experiments involved 10 randomized experiments for different τ, and a comparative analysis was given using the lost count and stability. The calculation of the lost count is shown in (3). The loss count of the j-th τ is denoted $Lost(\tau_j)$. The value of j is an integer between 1 and h (denoted by $Z[1,h]$). $T^{lost}_i\,|\tau_j$ is the lost count of the j-th randomized experiment for τ_j.

$$Lost(\tau_j) = \sum_{i=1}^{10} T^{lost}_i\,|\tau_j (j \in Z[1,h]) \tag{3}$$

Stability was investigated from two aspects: internal stability and external stability.

The calculation of internal stability is shown in (4). The internal stability of τ_j in the top d ranges is denoted $Internal - stability(\tau_j\,|pre-d\,)$ in (4). The union of the results for τ_j on 10 randomized experiments in the top d ranges is denoted $\bigcup_{j=1}^{10} Lr(\tau_j^{pre-d})$. The occurrence counts of Lr_i in the union are denoted $count(Lr_i \Big| \bigcup_{j=1}^{10} Lr(\tau_j^{pre-d}))$.

$$Internal - stability(\tau_j\,|pre-d\,) = \frac{\sum_{i=1}^{d} count(Lr_i \Big| \bigcup_{j=1}^{10} Lr(\tau_j^{pre-d}))}{d \times 10} \tag{4}$$

The calculation of external stability is shown in (5). The external stability of τ_j in the top d ranges is denoted $External - stability(\tau_j\,|pre-d\,)$ in (5). The union of the results for $\tau_{j\in[\tilde{h},h]}$ in the top d ranges is denoted $\bigcup_{j=\tilde{h}}^{h} Lr(\tau_j)$. The occurrence counts of Lr_i in the union are denoted $count(Lr_i \Big| \bigcup_{j=\tilde{h}}^{h} Lr(\tau_j))$.

$$External - stability(\tau_j\,|pre-d\,) = \frac{\sum_{i=1}^{d} count(Lr_i \Big| \bigcup_{j=\tilde{h}}^{h} Lr(\tau_j))}{d \times 10} \tag{5}$$

In the first group of experiments, $RF_{cv-number}(Lr^{sub})$ is set to 480, and 18 groups of data are taken in the interval [3000, 11500]($h = 18, \tilde{h} = 9$). The experimental results are shown in Table 3 and Table 4. It can be seen from Table 3 that the lost count gradually approaches 0 from $\tau_9 = 7000$, and there are two fluctuations of 1 loss in $\tau_{11} = 8000$ and $\tau_{12} = 8500$, and 0 loss is stable from $\tau_{13} = 9000$. It can be seen from Table 4 that from $\tau_{12} = 8500$, the internal stability of the top 60 and the top 80 both reached 100%, the top 100 reached more than 98.90%, and the top 120 reached more than 93.50%. From $\tau_{12} = 8500$, the external stability

TABLE 3. Loss count analysis of $RF_{cv-number}(Lr^{sub}) = 480$.

τ	h	serial number of 10 randomized experiments										lost count
		T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	
3000	1	5	3	2	1	5	3	3	8	4	7	41
3500	2	3	4	5	3	1	1	4	3	3	2	29
4000	3	3	3	3	2	1	2	3	1	2	1	21
4500	4	1	1	1	0	1	2	0	0	2	1	9
5000	5	0	1	1	1	1	1	1	1	1	1	9
5500	6	0	0	1	2	1	1	0	0	1	1	7
6000	7	0	0	0	0	0	0	0	1	0	0	1
6500	8	0	1	0	2	2	0	0	0	1	0	6
7000	$\tilde{h}$ 9	0	0	0	0	0	0	1	0	0	0	1
7500	10	0	0	0	0	0	0	0	0	0	0	0
8000	11	0	0	1	0	0	0	0	0	0	0	1
8500	12	0	0	0	0	0	0	0	0	1	0	1
9000	13	0	0	0	0	0	0	0	0	0	0	0
9500	14	0	0	0	0	0	0	0	0	0	0	0
10000	15	0	0	0	0	0	0	0	0	0	0	0
10500	16	0	0	0	0	0	0	0	0	0	0	0
11000	17	0	0	0	0	0	0	0	0	0	0	0
11500	h 18	0	0	0	0	0	0	0	0	0	0	0

TABLE 4. Stability analysis of $RF_{cv-number}(Lr^{sub}) = 480$.

τ	internal stability(%)				external stability(%)			
	top 60	top 80	top 100	top 120	top 60	top 80	top 100	top 120
7000	+	+	98.60	92.91	+	+	99.70	96.75
7500	+	+	99.00	93.58	+	+	99.80	95.83
8000	+	99.38	96.50	90.75	+	99.75	99.80	95.58
8500	+	+	+	97.33	+	+	99.90	95.75
9000	+	+	99.20	94.08	+	+	99.70	97.00
9500	+	+	99.70	94.17	+	+	99.40	96.42
10000	+	+	98.90	93.50	+	+	99.60	95.75
10500	+	+	99.60	94.58	+	+	99.60	96.08
11000	+	+	99.00	94.08	+	+	99.70	96.92
11500	+	+	99.60	94.67	+	+	99.90	95.75

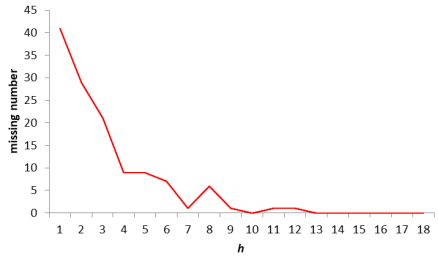


FIGURE 4. Loss count curve of $RF_{cv-number}(Lr^{sub}) = 480$.

of the top 60 and the top 80 both reached 100%, the top 100 reached more than 99.40%, and the top 120 reached more than 95.75%. Figure 5 shows that both the internal stability and external stability had a relatively high stability trend from τ_{13}. The left boundary ($\tau_{13} = 9000$) was moved two digits to the right. Finally, $\tau_{optimal}$ was set to $\tau_{15} = 10000$, and η was set to $\left\lfloor \sqrt{480} + 0.5 \right\rfloor = 22$.

In the second group of experiments, $RF_{cv-number}(Lr^{sub})$ is set to 240, and 13 groups of data are taken in the interval [1000, 7000] ($h = 13, \tilde{h} = 6$). The experimental results are shown in Table 5 and Table 6. It can be seen from Table 5 that the lost count gradually approaches 0 from $\tau_6 = 3500$, and there is a fluctuation of 1 loss in $\tau_7 = 4000$, and 0 loss is stable from $\tau_8 = 4500$.

It can be seen from Table 6 that from $\tau_6 = 3500$, the internal stability of the top 60 and the top 80 both reached 100%, the top 100 almost reached 100% (except a fluctuation of 99.80% for $\tau_{10} = 5500$), and the top 120 reached more than 95.33%. From $\tau_6 = 3500$, the external stability of the top 60, the top 80 and the top 100 all reached 100%, and the top 120 reached more than 95.56%. Figure 7 shows that both the internal stability and external stability had a relatively high stability trend from τ_{10} (except for a fluctuation in the top 100). The left boundary ($\tau_8 = 4500$) was moved three digits to the right. Finally, $\tau_{optimal}$ was set to $\tau_{11} = 6000$, and η was set to $\left\lfloor \sqrt{240} + 0.5 \right\rfloor = 15$.

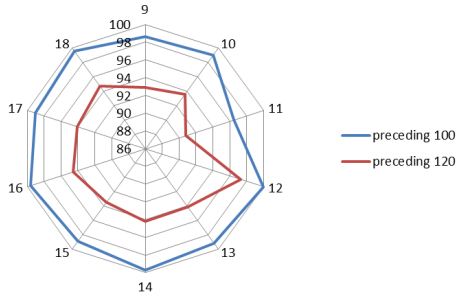
In the third group of experiments, $RF_{cv-number}(Lr^{sub})$ is set to 120, and 9 groups of data are taken in the interval [500, 2500] ($h = 9, \tilde{h} = 6$). The experimental results are shown in Table 7 and Table 8. It can be seen from Table 7 that 0 loss is stable from $\tau_6 = 1000$. It can be seen from Table 8 that from $\tau_6 = 1000$ the internal stability

TABLE 5. Loss count analysis of $RF_{cv-number}(Lr^{sub}) = 240$.

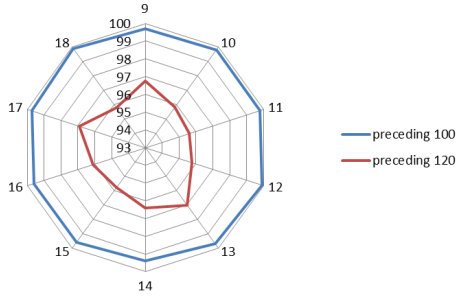
τ	h	serial number of 10 randomized experiments										lost count
		T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	
1000	1	7	9	11	11	6	10	12	3	6	12	87
1500	2	7	2	2	3	2	2	2	2	1	3	26
2000	3	4	1	3	0	1	1	2	1	2	2	17
2500	4	2	0	1	0	1	0	0	1	0	0	5
3000	5	1	0	0	0	0	0	1	0	0	0	2
3500	$\tilde{h}$ 6	0	0	0	0	0	0	0	0	0	0	0
4000	7	0	0	0	0	0	0	0	0	1	0	1
4500	8	0	0	0	0	0	0	0	0	0	0	0
5000	9	0	0	0	0	0	0	0	0	0	0	0
5500	10	0	0	0	0	0	0	0	0	0	0	0
6000	11	0	0	0	0	0	0	0	0	0	0	0
6500	12	0	0	0	0	0	0	0	0	0	0	0
7000	h 13	0	0	0	0	0	0	0	0	0	0	0

TABLE 6. Stability analysis of $RF_{cv-number}(Lr^{sub}) = 240$.

τ	internal stability(%)				external stability(%)			
	top 60	top 80	top 100	top 120	top 60	top 80	top 100	top 120
3500	+	+	+	95.67	+	+	+	96.98
4000	+	+	+	95.33	+	+	+	96.67
4500	+	+	+	95.67	+	+	+	98.12
5000	+	+	+	96.33	+	+	+	97.19
5500	+	+	99.80	95.50	+	+	+	96.56
6000	+	+	+	96.75	+	+	+	97.81
6500	+	+	+	96.41	+	+	+	97.92
7000	+	+	+	96.83	+	+	+	96.67



(a) internal stability



(b) external stability

FIGURE 5. Stability distribution of $RF_{cv-number}(Lr^{sub}) = 480$.

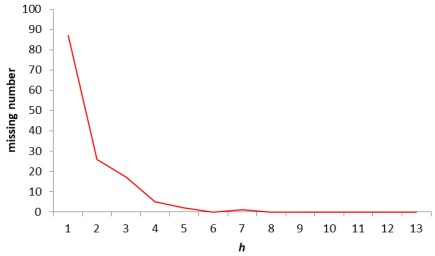


FIGURE 6. Loss count curve of $RF_{cv-number}(Lr^{sub}) = 240$.

of the top 60 reached 100%, the top 80 reached more than 98.25%, and the top 100 reached more than 92.60%. From $\tau_8 = 2000$, the external stability of the top 60 and the top 80 both reached 100%, and the top 100 reached more than 94%. Figure 9 shows that both internal stability and external stability had a relatively high stability trend from τ_8. The left boundary ($\tau_6 = 1000$) was moved two digits to the right. Finally, $\tau_{optimal}$ was set to $\tau_8 = 2000$, and η was set to $\lfloor \sqrt{120} + 0.5 \rfloor = 11$.

Figure 4, Figure 6 and Figure 8 show the following: The lost count for τ_j was more in the early stage and decreased greatly in the middle stage, but was unstable. It decreased to 0 in the later stage and tended to be stable. Note: "+" in Tables 4, 6 and 8 represents 100.

B. PERFORMANCE EVALUATION OF CVS_E-CS-CF

The parameter settings for the CVS_e-CS-CF algorithm are shown in Table 9. The $Set_{penalty}$ obtained by the CVS_e-CS-CF algorithm is denoted $Set^i_{penalty}(i \in [1, 25])$

TABLE 7. Loss count analysis of $RF_{cv-number}(Lr^{sub}) = 120$.

τ	h	serial number of 10 randomized experiments										lost count
		T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	
500	1	2	3	2	1	1	2	4	2	2	3	22
600	2	2	0	0	1	0	0	1	1	0	2	7
700	3	1	2	2	1	0	1	0	0	1	0	8
800	4	0	0	0	0	1	1	1	1	1	0	5
900	5	1	1	0	1	0	1	0	0	0	0	4
1000	$\tilde{h}$ 6	0	0	0	0	0	0	0	0	0	0	0
1500	7	0	0	0	0	0	0	0	0	0	0	0
2000	8	0	0	0	0	0	0	0	0	0	0	0
2500	h 9	0	0	0	0	0	0	0	0	0	0	0

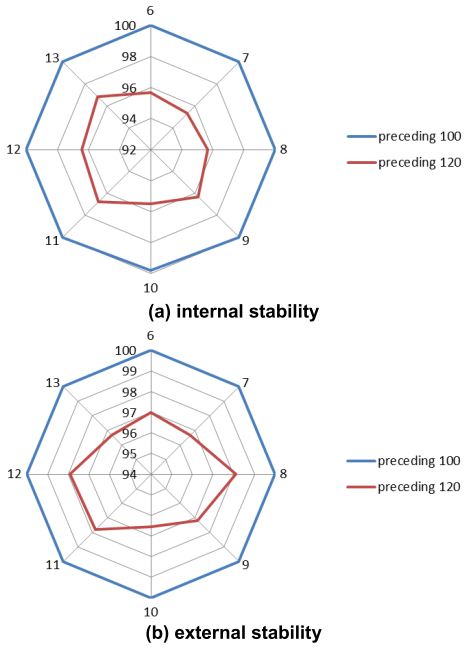


FIGURE 7. Stability distribution of $RF_{cv-number}(Lr^{sub}) = 240$.

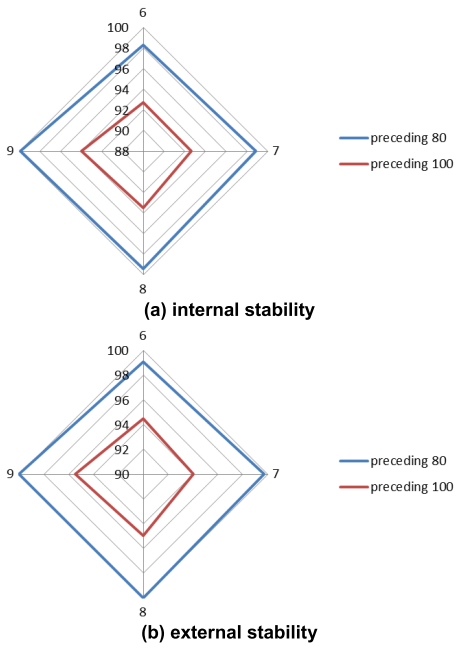


FIGURE 9. Stability distribution of $RF_{cv-number}(Lr^{sub}) = 120$.

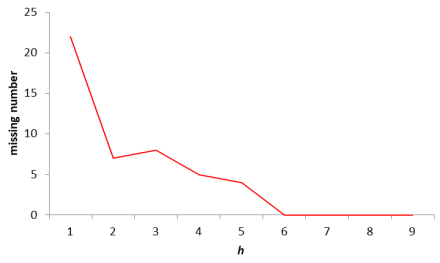


FIGURE 8. Loss count curve of $RF_{cv-number}(Lr^{sub}) = 120$.

(as shown in Table 10). The location of $Set^i_{penalty}$ in Θ is denoted $position(Set^i_{penalty})^{\Theta}$. The relative position coefficient of $Set^i_{penalty}(RPC(Set^i_{penalty}))$ is calculated via (6). The Set_{shock} containing 30 elements is shown in Table 11.

$$RPC(Set^i_{penalty}) = \frac{position(Set^i_{penalty})^{\Theta}}{\Delta} \tag{6}$$

RPC is a value between 0 and 1. The smaller it is, the more important the Lr_i is. This indicates that the location of $Set^i_{penalty}$ in Θ is at the front. The larger the RPC value, the less important the Lr_i is. This indicates that the location of $Set^i_{penalty}$ in Θ is later in the queue. If $RPC(Set^i_{penalty})$ is larger, $Set^i_{penalty}$ will be punished. It indicates that the algorithm could protect those Lr_i with higher importance, so the algorithm has better stability. According to Figure 10, 96% of $RPC(Set^i_{penalty})$ in $Set_{penalty}$ are above 0.49 (except 0.44), the mean of which is 0.79. This indicates that the top 60 $Lr_i s$ in Θ are stably protected and the algorithm has good stability. Ultimately, λ lncRNAs most closely related to prostate cancer are selected by the CVS_e-CS-CF algorithm from 480 lncRNAs ($\lambda = 95$). The results are shown in Table 12.

C. PERFORMANCE EVALUATION OF CSPA-PL

After implementing $Logistic - Step(\pi_{\Theta^{select}_{sub-pre}}(M^{PP}))$ in the inspection stage of the CSPA-PL algorithm, a total

TABLE 8. Stability analysis of $RF_{cv-number}(Lr^{sub}) = 120$.

τ	internal stability(%)			external stability(%)		
	top 60	top 80	top 100	top 60	top 80	top 100
1000	+	98.25	92.70	+	99.06	94.50
1500	+	98.88	92.60	+	99.69	94.00
2000	+	99.38	93.50	+	+	95.00
2500	+	99.88	94.00	+	+	95.50

TABLE 9. Parameter setting of CVS $_E$-CS-CF algorithm.

Lr	τ_0	η_0	$\tau_{1\text{-}3}$	$\eta_{1\text{-}3}$	$\tau_{4\text{-}9}$	$\eta_{4\text{-}9}$	Δ	Ω	a	b	f
M^{CV}	10000	22	6000	15	2000	11	120	60	1	121	0

TABLE 10. Result of $Set_{penalty}$.

Lr_i	gene name	Lr_i	gene name	Lr_i	gene name
41	FAM201A	99	MATN1-AS1	193	TSPAN10
43	RP11-262H14.1	101	RP11-244F12.3	204	RP11-696N14.1
45	RP11-713P17.3	104	LINC00476	236	AC139666.1
61	AC093627.10	112	CTC-308K20.1	239	RP11-834C11.3
66	CTC-504A5.1	119	POLR2J4	251	RP11-558F24.4
67	CTC-774J1.2	120	AMZ2P1	331	MSR1
73	RP11-573I11.2	122	A2M-AS1	382	RP11-446H18.3
90	RP11-206L10.11	151	LINC00086		
97	LINC00338	169	CTBP1-AS1		

TABLE 11. Result of Set_{shock}.

Lr_i	gene name	Lr_i	gene name	Lr_i	gene name	Lr_i	gene name
19	AC003090.1	135	KB-431C1.4	223	PVT1	272	PRKY
34	RPL13P5	160	UHRF1	225	SNHG15	310	ZNFX1-AS1
63	RNF126P1	170	RP11-72M17.1	230	AP001372.2	369	MTMR9LP
70	TTTY15	185	KTN1-AS1	234	SNHG7	383	RP11-488C13.5
87	RHPN1-AS1	197	RP11-294O2.2	243	NDUFB2-AS1	395	LINC00402
102	RP5-916L7.1	211	DLX6-AS1	253	CTD-2195M18.1	408	MLLT4-AS1
114	AC002055.4	214	RP4-659J6.2	258	WHAMMP2		
118	ATG9B	219	RP11-500G10.1	267	FLNB-AS1		

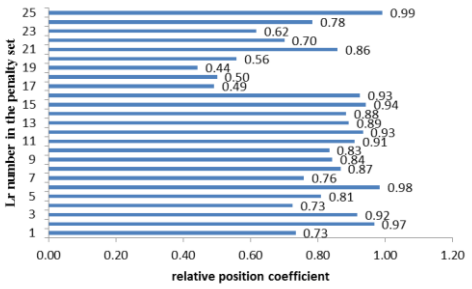


FIGURE 10. Comparison of Relative position coefficient distribution on $Set_{penalty}$.

of 21 groups ($Z^{pre}_{[1,21]}$) were obtained. The AIC distribution of $Z^{pre}_{[1,21]}$ is shown in Figure 11. This shows that the value of Z^{pre}_{8} is the minimum (157.66). So, $\{\Upsilon^{pre}\} = \{Lr_{i(1)}\}$ is obtained. After implementing $Logistic - Step(\pi_{\Theta^{select}_{sub-rear}}(M^{PP}))$, a total of 39 groups ($Z^{rear}_{[1,39]}$) were obtained. In this, the value of Z^{rear}_{26} is the minimum (172.29). There are three $Lr_i s$ in Z^{rear}_{26}, the P-values of which were less than 0.01. So, $\{\Upsilon^{rear}\} = \{Lr_{i(2)}\}$ is obtained. The values of $i(1)$ and $i(2)$ in the above $Lr_{i(1)}$ and $Lr_{i(2)}$ are shown in Table 13. There were three alternative sets in $Buffer - pool$, which were $Buffer - pool_1 = \{\{\Upsilon^{pre}\} \cup Lr_{162}\}$, $Buffer - pool_2 = \{\{\Upsilon^{pre}\} \cup Lr_{79}\}$ and $Buffer - pool_3 = \{\{\Upsilon^{pre}\} \cup Lr_{110}\}$. Next, $Buffer - pool^*_i$ was obtained by executing $Logistic - Step(\pi_{Buffer-pool_i}(M^{PP}))$ on $Buffer - pool_i$, which produced $Buffer - pool^*_1 = \{Lr_{j(1)}\}$, $Buffer - pool^*_2 = \{Lr_{j(2)}\}$ and $Buffer - pool^*_3 = \{Lr_{j(3)}\}$. The values of $j(1)$, $j(2)$ and $j(3)$ in the above $Lr_{j(1)}$, $Lr_{j(2)}$, and $Lr_{j(3)}$ are shown in Table 14. The accuracy comparison experiment for $Buffer - pool^*_i$ for the logistic regression model shows that the accuracy rate of $Buffer - pool_2 = \{\{\Upsilon^{pre}\} \cup Lr_{79}\}$ is the highest. Finally, the optimal logistic regression model for tumor clinical stage (denoted by $\Phi_{optimal}$) is obtained, which contained 22 $Lr_i s$ (as shown in Table 15).

TABLE 12. Result of $\Theta^{select}(\lambda = 95)$.

λ	Lr_i	gene name	λ	Lr_i	gene name	λ	Lr_i	gene name	λ	Lr_i	gene name
1	6	LINC00665	25	62	SRD5A2	49	79	NEURL3	73	57	RP11-166D19.1
2	3	RP11-342C23.4	26	9	FAM222A-AS1	50	46	RP5-1121A15.1	74	159	RP9P
3	7	RP11-368I7.2	27	27	BOLA3-AS1	51	55	RP11-166D19.1	75	180	RP11-392A22.2
4	30	CTD-3199J23.4	28	29	C1orf126	52	140	DANCR	76	39	RP11-597D13.9
5	20	AP001626.1	29	25	AC073343.13	53	186	RP11-48O20.4	77	77	AP000662.4
6	15	RP11-429J17.6	30	21	RP11-279F6.1	54	348	LINC00261	78	60	LGALS8-AS1
7	28	CTC-497E21.4	31	23	RP11-401F24.4	55	246	MORC2-AS1	79	162	RP11-182J1.9
8	12	CTD-2527I21.11	32	26	MAGI2-AS3	56	54	RP11-379F4.4	80	24	RP11-314O13.1
9	53	RP3-467K16.2	33	47	RP11-316P17.2	57	33	FBXL19-AS1	81	76	RP13-15E13.1
10	14	RP11-279F6.1	34	5	RP11-231P20.2	58	32	RP11-418J17.1	82	1	AC017048.3
11	64	SNHG16	35	86	RP11-108M9.4	59	121	BX284650.1	83	110	HOXA-AS2
12	8	MSL3P1	36	98	PGM5-AS1	60	139	MIR22HG	84	215	NOP14-AS1
13	51	LINC00087	37	82	AL078621.4	61	116	CD27-AS1	85	85	C17orf76-AS1
14	10	PCA3	38	22	AC073133.1	62	115	AC007743.1	86	37	RP11-875O11.1
15	65	LINC00092	39	137	RP11-66B24.4	63	36	ADAMTS9-AS2	87	168	AC073871.2
16	40	ADAMTS9-AS1	40	50	SNHG3	64	38	DNMBP-AS1	88	78	RP11-705C15.2
17	42	RP4-647C14.2	41	17	ERVH48-1	65	74	WWC2-AS2	89	68	PLEKHM1P
18	4	LINC00675	42	105	RP11-168G16.1	66	48	HLA-F-AS1	90	49	LINC00641
19	198	MIR205HG	43	156	XXbac-B135H6.15	67	59	RP11-57A19.2	91	189	RP4-564F22.2
20	13	RP11-627G23.1	44	31	RP11-324H6.5	68	58	RP11-17A19.1	92	107	RP11-519G16.3
21	16	RP11-279F6.1	45	56	SENP3-EIF4A1	69	117	NPY6R	93	125	RP11-399O19.5
22	52	MIR31HG	46	75	RP11-24M17.7	70	427	AC058791.2	94	232	ELOVL2-AS1
23	84	LINC00085	47	166	LINC00605	71	72	RP11-265M18.2	95	196	AC004540.5
24	35	RP11-412D9.4	48	91	SERTAD4-AS1	72	106	ZNF300P1			

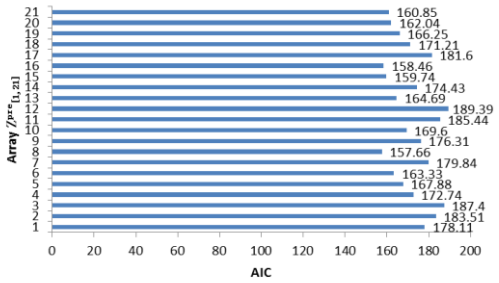


FIGURE 11. Contrastive distribution on *AIC* value of $Z^{pre}_{[1,21]}$.

TABLE 13. The $i(1)$ value of $Lr_{i(1)}$ in $\{\Upsilon^{pre}\}$, the $i(2)$ value of $Lr_{i(2)}$ in $\{\Upsilon^{rear}\}$.

Lr_i	value of i
$Lr_{i(1)}$	6,15,12,53,14,64,8,10,40,42,4,16,35,62,27,29,25,21,23,26,47,5,82,137,50,17,166
$Lr_{i(2)}$	162,79,110

It can be seen that the CSPA-PL algorithm could further select 22 $Lr_i s$ from the 95 $Lr_i s$ most closely related to prostate cancer, which are most closely related to the tumor clinical stage of prostate cancer.

In order to verify the universality of the work presented in this paper, we also chose the lung cancer data set in the lncRNAtor database and TCGA database for experimentation. A total of 290 samples were obtained, including 46 normal samples and 245 cancer samples. First, the importance of the characteristic variable was calculated using the CVS_gC-RF algorithm. Second, 120 lncRNAs which were most closely related to prostate cancer, were calculated from the 480 alternative lncRNAs by CASO and CVS_e-CS-CF. On the basis of the above 120 lncRNAs, the CSPA-PL algorithm was adopted to further select 16 lncRNAs that were most closely related to the tumor clinical stage of lung cancer. Finally, 16 lncRNAs were used to construct a logistic regression prediction model.

TABLE 14. The value of $j(1), j(2)$ and $j(3)$ of $Lr_{j(1)}, Lr_{j(2)}$ and $Lr_{j(3)}$ in $Buffer - pool^*_i$.

Lr_j	value of j
$Lr_{j(1)}$	6,15,12,14,64,8,40,42,4,16,62,27,29,25,21,47,82,137,50,17,166,162
$Lr_{j(2)}$	6,15,12,14,64,8,40,42,4,16,35,62,27,29,25,21,47,137,50,17,166,79
$Lr_{j(3)}$	6,15,12,53,14,64,8,10,40,42,4,16,35,62,27,29,25,21,23,26,47,5,82,137,50,17,166

D. PREDICTION RESULT

Three state-of-the-art methods (MlrLDAcp [26], REPTree [27], NaïveBayes [28]) were selected to compare with CSPA-PL by 10-fold cross validation. The comparison experiments were carried out from three aspects: ROC area, prediction accuracy and recall rate. The results of ROC area for prostate cancer are shown in Figure 12. The mean ROC area of three compared methods for prostate cancer was 0.673, and the ROC area of the CSPA-PL method in this paper was 0.857, which was the largest and 1.27 times that of the other methods. The results of ROC area for lung cancer are shown in Figure 13. The mean ROC area of the compared methods

TABLE 15. Result of Lr_i in $\Phi_{optimal}$.

Lr_i	gene name	Lr_i	gene name	Lr_i	gene name
x6	LINC00665	x4	LINC00675	x47	RP11-316P17.2
x15	RP11-429J17.6	x16	RP11-279F6.1	x137	RP11-66B24.4
x12	CTD-2527I21.11	x35	RP11-412D9.4	x50	SNHG3
x14	RP11-279F6.1	x62	SRD5A2	x17	ERVH48-1
x64	SNHG16	x27	BOLA3-AS1	x166	LINC00605
x8	MSL3P1	x29	C1orf126	x79	NEURL3
x40	ADAMTS9-AS1	x25	AC073343.13		
x42	RP4-647C14.2	x21	RP11-279F6.1		

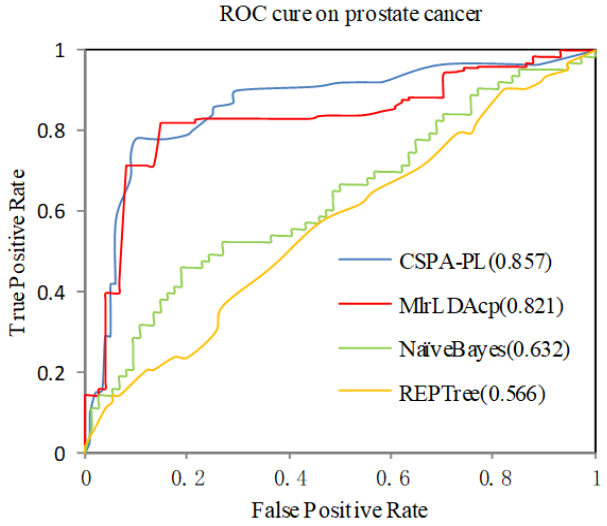


FIGURE 12. ROC area comparison on prostate cancer.

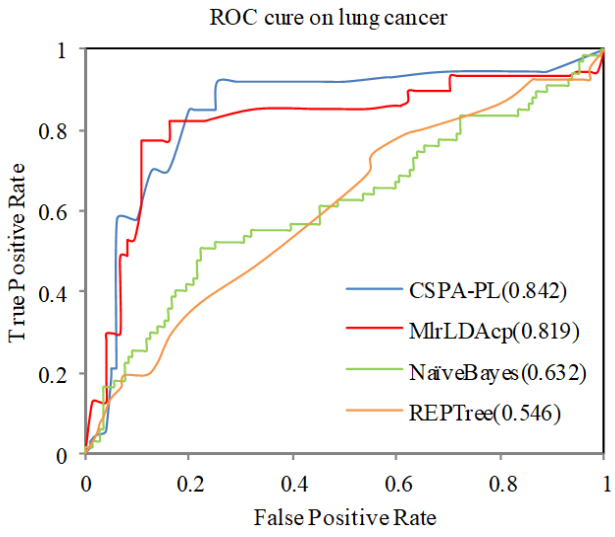


FIGURE 13. ROC area comparison on lung cancer.

for lung cancer was 0.666, and the ROC area of the CSPA-PL method in this paper was 0.842, which was the largest and 1.26 times that of the other methods. In the comparison experiment for recall rate, the accuracy and recall rate were often mutually restricted and offset each other. Therefore, AV-PR was implemented and used in this paper as the mean

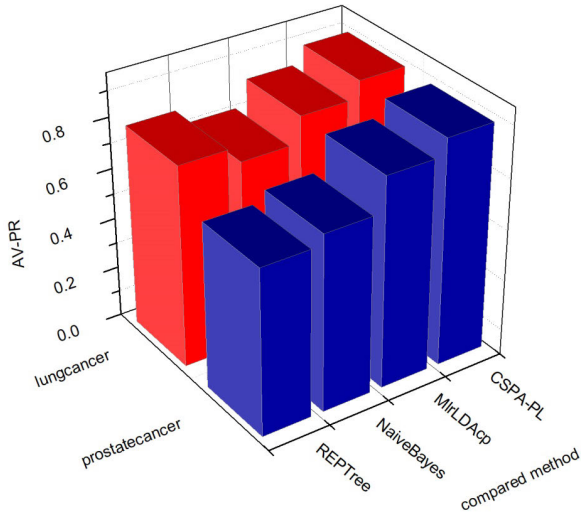


FIGURE 14. AV-PR comparison.

of the prediction accuracy and recall rate. The results for AV-PR are shown in Figure 14(The prostate cancer is shown by blue, the lung cancer is shown by red.). The mean AV-PR of the compared methods for prostate cancer was 0.723, and the AV-PR of CSPA-PL with the maximum was 0.889, which was about 1.231 times that of the other methods. The mean AV-PR of the compared methods for lung cancer was 0.784, and the AV-PR of CSPA-PL with the maximum was 0.896, which was about 1.142 times that of the other methods. These results indicate that the accuracy rate and ROC area for CSPA-PL were both good.

IV. CONCLUSION AND DISCUSSION

Although some methods have been applied to lncRNA-disease association prediction, clinical prognostic data were rarely involved. In this study, we constructed a clinical stage prediction algorithm for cancer-associated lncRNA (CSPA-PL), which utilized cancer clinical stage data. CSPA-PL was based on unknown human lncRNA-disease associations combining with the clinical stage data. The core modules of CSPA-PL included CASO, the CVS_g-RF algorithm, and the CVS_e-CS-CF algorithm. For CASO, a learning mode was formed in which the first quartile area was a defensive area and the other quartile areas were offensive. Three symmetric ideas were adopted, which were center allelism,

X-axis allelism and Y-axis allelism. The CVS_g-RF algorithm employed a variable selection algorithm based on random forests as the core to calculate the importance of the characteristic variables. This method exhibited good robustness. Experimental results showed that the proposed method in this study has good predictive performance.

The value of this model lies in the following:

(a) It provides a strong research foundation for the prediction of prognosis information for cancer patients by lncRNA-disease association.

(b) The simplified lncRNAs in the model were the closest to predicting the relationship between lncRNA and the disease, which provides a favorable research premise for subsequent studies of this association.

REFERENCES

[1] M. Guttman, P. Russell, N. T. Ingolia, J. S. Weissman, and E. S. Lander, "Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins," *Cell*, vol. 154, no. 1, pp. 240–251, Jul. 2013.

[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 (review)," *Int. J. Oncol.*, 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] 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.

[5] 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 (review)," *Oncol. Reports*, vol. 40, pp. 1831–1842, Aug. 2018.

[6] L. Chen, X. Guo, Z. Li, and Y. He, "Relationship between long non-coding RNAs and alzheimer's disease: A systematic review," *Pathol.-Res. Pract.*, vol. 215, no. 1, pp. 12–20, Jan. 2019.

[7] 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.

[8] 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.

[9] R. Zheng, J. Liang, J. Lu, S. Li, G. Zhang, X. Wang, M. Liu, W. Wang, H. Chu, G. Tao, Q. Zhao, M. Wang, M. Du, F. Qiang, and Z. Zhang, "Genome-wide long non-coding RNAs identified a panel of novel plasma biomarkers for gastric cancer diagnosis," *Gastric Cancer*, vol. 22, no. 4, pp. 731–741, Jul. 2019.

[10] S. Saijo, Y. Kuwano, S. Tange, K. Rokutan, and K. Nishida, "A novel long non-coding RNa from the HOXa6-HOXa5 locus facilitates colon cancer cell growth," *BMC Cancer*, vol. 19, no. 1, p. 532, Dec. 2019.

[11] 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.

[12] 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," *Molecular Med. Reports*, pp. 7845–7858, Apr. 2018.

[13] 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, Dec. 2019, doi: 10.1186/s12859-019-2985-0.

[14] Z. Cui, J.-X. Liu, Y.-L. Gao, R. Zhu, and S.-S. Yuan, "LncRNA-disease associations prediction using bipartite local model with nearest profile-based association inferring," *IEEE J. Biomed. Health Informat.*, to be published, doi: 10.1109/JBHI.2019.2937827.

[15] X. Chen and G.-Y. Yan, "Novel human lncRNA–disease association inference based on lncRNA expression profiles," *Bioinformatics*, vol. 29, no. 20, pp. 2617–2624, Oct. 2013.

[16] Y.-A. Huang, X. Chen, Z.-H. You, D.-S. Huang, and K. C. C. Chan, "ILNCSIM: Improved lncRNAs functional similarity calculation model," *Oncotarget*, vol. 7, no. 18, pp. 25902–25914, May 2016.

[17] X. Chen, "KATZLDA: KATZ measure for the lncRNA-disease association prediction," *Sci. Rep.*, vol. 5, no. 1, Dec. 2015, Art. no. 16840.

[18] T. Zhao, J. Xu, and L. Liu, "Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features," *Mol. Biosyst.*, vol. 11, no. 1, pp. 126–136,2015.

[19] G. Li, J. Luo, C. Liang, Q. Xiao, P. Ding, and Y. Zhang, "Prediction of LncRNA-disease associations based on network consistency projection," *IEEE Access*, vol. 7, pp. 58849–58856, 2019.

[20] M. Zhou, X. Wang, J. Li, D. Hao, Z. Wang, H. Shi, L. Han, H. Zhou, and J. Sun, "Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network," *Mol. BioSyst.*, vol. 11, no. 3, pp. 760–769, Dec. 2014.

[21] Y. Liu, R. Zhang, F. Qiu, K. Li, Y. Zhou, D. Shang, and Y. Xu, "Construction of a lncRNa–PCG bipartite network and identification of cancer-related lncRNas: A case study in prostate cancer," *Mol. BioSyst.*, vol. 11, no. 2, pp. 384–393, Oct. 2014.

[22] 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.*, to be published, doi: 10.1109/TCBB.2019.2934958.

[23] G. Yu, G. Fu, C. Lu, Y. Ren, and J. Wang, "BRWLDA: Bi-random walks for predicting lncRNA-disease associations," *Oncotarget*, vol. 8, no. 36, pp. 60429–60446, Sep. 2017.

[24] X. Chen, C. C. Yan, X. Zhang, and Z.-H. You, "Long non-coding RNAs and complex diseases: From experimental results to computational models," *Briefings Bioinf.*, vol. 18, no. 4, pp. 558–576, 2017.

[25] 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.

[26] B. Wang and J. Zhang, "Multiple linear regression analysis of lncRNA–disease association prediction based on clinical prognosis data," *BioMed Res. Int.*, vol. 2018, pp. 1–10, Dec. 2018, doi: 10.1155/2018/3823082.

[27] 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.

[28] 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.

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 was 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 the University of Melbourne. From 2005 to 2007, she was a Postdoctoral Researcher in 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 is currently a Professor with the School of Information Science and Engineering, University of Jinan. She is also a Doctoral Supervisor with the College of Computer Science and Technology, Harbin Engineering University. Her research interests include bioinformatics, image processing, and virtual reality.