Random Forest with Self-paced Bootstrap Learning in Lung Cancer Prognosis

QINGYONG WANG and YUN ZHOU, Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology , China
WEIPING DING*, School of Computer Science and Technology, Nantong University, China
ZHIGUO ZHANG, School of Biomedical Engineering, Health Science Center, Shenzhen University, China
KHAN MUHAMMAD, Department of Digital Contents, Sejong University, Korea
ZEHONG CAO* , Discipline of ICT, School of Technology, Environment and Design, College of Science and Engineering, University of Tasmania, Australia

Training gene expression data with supervised learning approaches can provide an alarm sign for early treatment of lung cancer to decrease death rates. However, the samples of gene features involve lots of noises in a realistic environment. In this study, we present a random forest with self-paced learning bootstrap for improvement of lung cancer classification and prognosis based on gene expression data. To be specific, we proposed an ensemble learning with random forest approach to improving the model classification performance by selecting multi-classifiers. Then, we investigated the sampling strategy by gradually embedding from high- to low-quality samples by self-paced learning. The experimental results based on five public lung cancer datasets showed that our proposed method could select significant genes exactly, which improves classification performance compared to that in existing approaches. We believe that our proposed method has the potential to assist doctors for gene selections and lung cancer prognosis.

CCS Concepts: • **Life and medical sciences** → *Computational genomics*; • **Mathematics of computing** → *Resampling methods.*

Additional Key Words and Phrases: Lung Cancer, Random Forest, Self-paced Learning, Bootstrap, Classification

ACM Reference Format:
Qingyong Wang, Yun Zhou, Weiping Ding, Zhiguo Zhang, Khan Muhammad, and Zehong Cao. 2019. Random Forest with Self-paced Bootstrap Learning in Lung Cancer Prognosis . 37, 4, Article 111 (August 2019), 12 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Cancer is a disease of the body's cells. Typically cells grow and multiply in a controlled way, but this control may be lost if something causes a mistake to occur in the cells' genetic blueprints [1]. In terms of lung cancer, it is the leading cause of cancer death, and common cancer diagnosed all over the world [2]. Lung cancer symptoms include shortness of breath, wheezing, hoarseness, chest pain, coughing, or spitting up blood. To avoid this miserable situation, we need

*Corresponding Author

Authors' addresses: Qingyong Wang; Yun Zhou, Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology , China; Weiping Ding, School of Computer Science and Technology, Nantong University, China, dwp9988@163.com; Zhiguo Zhang, School of Biomedical Engineering, Health Science Center, Shenzhen University, China; Khan Muhammad, Department of Digital Contents, Sejong University, Korea; Zehong Cao, Discipline of ICT, School of Technology, Environment and Design, College of Science and Engineering, University of Tasmania, Australia, zehong.cao@utas.edu.au;zhcaonctu@gmail.com.

machine learning based models for the early detection and prevention of lung cancer. Taking medications during the early stages of lung cancer increases medication efficacy and reduces recurrence [3]. Therefore, pre-emptive detection and diagnosis of lung cancer may be clinically beneficial, especially for patients with undefined lung-related cancer. Biologically, the pathogenic genes of lung cancer are found and confirmed by experiments, but most of the genes have a low correlation with the disease. We need various experiments to prove if a gene is correlated with lung cancer, which would cost lots of resources. However, machine learning algorithms can prioritize the most important 1% of the disease-causing genes, and then prove through biological experiments, which can save a lot of resources.

The relationship between tumor diagnosis and disease-causing genes can be discovered by machine learning approaches, which can be used for predicting cancer in an early stage. To be specific, some effective machine learning algorithms have been reported in the literature, including the computer-aided diagnosis based on artificial neural networks [4], ensemble-based feature selection methods [5], AdaBoost algorithm with support vector machine (SVM) [6], enhanced probabilistic neural network [7], a hybrid approach combining the advantages of fuzzy sets, ant-based clustering and multilayer perceptron neural networks classifiers [8]. Since ensemble learning can improve the robustness and accuracy of the model by combining multiple weak classifiers [9], we assume that ensemble learning could be useful in this study. Particularly, ensemble learning involved in bagging and boosting, and random forest can be considered as a common bagging algorithm for the improvement of cancer classification.

Gene microarray technology has emerged as a prospective tool for diagnosis and classification of cancers, and statistical or machine learning based methods have been applied to extract reliable gene features as the inputs of cancer classification models. In terms of lung cancer data, the small sample size is a challenge for microarray data analysis and training. If samples involved in some noise, it causes negative influences on the performance of training models. Furthermore, the sampling strategy only randomly searches classifiers using the random forest, and a better choice that we assumed sampling examples in the training stage by gradually embedding from high to low-quality samples. For example, an alternative approach with self-paced learning as a new formulation is designed to identify high-quality samples [10]. That is to say, by gradually increasing the penalty of the SPL regularizer during the optimization, more samples are selected in the training stage from high- to low-quality modes. Of note, it has been enjoyed rapidly increasing adoption, such as multi-task learning [11], image classification [12], molecular descriptor selection [13]. Therefore, in this study, we propose a novel random forest with self-paced learning (RFSPL), allowing us to extract samples with high-quality effectively.

The contributions of this paper are listed as follows. Firstly, we proposed the bootstrap based on self-paced learning with samples from high-confidence to low-confidence to train the model for cancer diagnosis and classification using DNA microarray technology. Secondly, we presented a random forest with self-paced learning framework for lung cancer diagnosis. Our proposed method is superior to existing classifiers in terms of accuracy, F1-score, and AUC. Additionally, our proposed method could select a small number ($< 1\%$) of highly relevant genes to facilitate the early prognosis of the disease.

The rest of this paper is organized as follows. Section 2 describes the related work, including lung cancer classification, ensemble learning, and random forest approaches. Then, a random forest with self-paced learning is proposed in section 3, to process the noisy samples for improving the classification performance. Furthermore, section 4 presents that our proposed method and comparable methods are tested based on simulation data and real experimental datasets. Finally, we discussed and concluded our study in section 5.

2 RELATED WORK

In terms of cancer classification, cancers are classified by the type of cell that the tumor cells resemble and is therefore presumed to be the origin of the tumor. The SVM, a discriminative classifier formally defined by a separating hyperplane, is applied in the field of cancer prognosis. The SVM classifier needs to decide different types of the kernel function in the training process. The previous experimental results showed the linear kernel, and radial basis function (RBF) kernel methods can be the better choices for a small scale dataset. For a large scale dataset, SVM with RBF kernel can be improved 10% of accuracy than the other classifiers [14]. In terms of logistic regression, it can develop a multi-parametric model suitable for prospectively identifying prostate cancer, and the result is not statistically significant (P=0.090) [15, 16].

Furthermore, three popular data mining techniques, such as decision trees, artificial neural networks (ANNs), and SVM, are used to develop prediction or classification models for cancer survivability [17]. For example, breast cancer survivability is evaluated by multiple data mining and statistical methods, and the results indicated that the decision tree (C5) is the best predictor than other approaches [18]. On the basis of statistical learning theory, the ensemble learning model can improve robustness, accuracy, and generalization of the classification models [19]. The ensemble learning includes Bagging [20], Adaboost [21], random forest [22], rotation forest [23] et al. Therefore, the ensemble learning models, including bagging, boosting and random forest measurement, are the potential to be applied in our study.

Primarily, the random forest was investigated in cancer classification, which is considered as a gene selection model for cancer classification. Prior studies showed that random forest could achieve better performance comparing to other classification methods, including k-nearest neighbors (KNN) and SVM. Due to the small sets of genes in the selection procedure [24], the random forest is only used in sub-sampling to train disease prediction in fully balanced samples. The result shows an average area under the curve (AUC) of a random forest could achieve better performance than that of SVM based on eight disease datasets [25].

To learn in a self-controlled pace, the self-paced learning (SPL), by mimicking the cognitive mechanism of humans and animals, was proposed to learn from easy to hard samples gradually. Due to the characteristic of generality, prior studies have been developed various types of SPL to embed curriculum design as a regularization term into the learning objectives [26]. The SPL aims to achieve a better weighting strategy by determining the minimizer functions, and a recent approach improved by artificially designing the specific form of SPL regularizers [27]. An example in [28], showed that the alternative search strategy method could measure the majorization-minimization in SPL, and deduce the underlying objectives of hard, linear, and mixture regularizers.

Early detection and diagnosis of lung cancers using a computed tomography (CT) may benefit the reduction of lung cancer mortality [29]. With the development of deep learning and convolutional neural networks (CNNs), it can be identified to analyze lung CTs for prognosis prediction and diagnosis [30]. Deep learning is accessible machine learning toolbox for image processing. The lung cancer computed tomography (CT), a special image, can be used to predict overall survival of non-small-cell lung cancer patients from CT data by deep learning network [31]. Furthermore, a trained convolutional neural network (CNN, or ConvNet) can extract deep features from CT images to predict short- and long-term survivors, leading to an enhancement of 12.5% than that of decision tree classifier [32]. Besides deep learning applied in lung cancer, other types of medical images recognition task are learned by a deep convolutional neural network (DCNN), such as breast cancer diagnosis [33].

3 METHODOLOGY

3.1 Bootstrap with self-paced learning

The bootstrap involves resampling from one's samples with replacement. Giving training dataset $D = (x_i, y_i)_{i=1}^m$ with m samples, where $x_i \in R^d$ is the $i\text{-}th$ sample, y_i represents its label, let $L(y_i, f(x_i, w))$ denote the loss function, it calculates the cost between ground truth label y_i and estimated one $f(x_i, w)$. Here w represents the model parameter inside the decision function g. The SPL model includes a weighted loss term on all samples and a general self-paced regularizer imposed on sample weights, expressed as:

$$\min_{w, v\in[0,1]^m} E(w, v; \lambda) = \sum_{i=1}^{m} v_i L(y_i, f(X_i, w)) + g(v_i; \lambda), \tag{1}$$

where λ is a age parameter for controling the learning pace, and $g(\lambda, v_i)$ is self-paced regularizer, whose intrinsic conditions have been theoretically abstracted by [34]. By jointly learn the model parameter w and the latent weight $v = [v_1, v_2, ..., v_m]^T$ by alternative search strategy algorithm with gradually increasing age parameter, more samples can be automatically included into training form easy to comlex in a purely self-paced way. Specifically, giving sample weights v, the minimization over w is a weighted loss minimization problem, independent on regularizer $g(v_i; \lambda)$; Giving model parameter w, the optimal weight of $i\text{-}th$ sample is determined in Equation (2),

$$v_i^* = min_{v_i} v_i L(y_i, f(x_i, w)) + g(v_i; \lambda), \tag{2}$$

where $l_i = L(y_i, f(X_i, w))$ denotes the loss of samples x_i.

To solve v is Equation (2), the self-paced function $g(v_i; \lambda)$ needs to be specified. Reference [35] has summarized the general properties of a self-pace function in $g(v_i; \lambda)$, it is convex w.r.t. $v_i \in [0, 1]$ to guarantee the uniqueness of v_i^*. $v^*(L_i; \lambda$ is monotonically decreasing w.r.t. L_i, which guides the model to select easy samples with small losses in favor of complex samples with larger losses. Similarly, it is monotonically increasing w.r.t. λ, which means that a larger λ has a higher tolerance to the losses and can incorporate more complex samples. We specify the self-paced function as the one for mixture weighting, due to its overall better performance in the experiments:

$$g(v_i; \lambda, \zeta) = -\zeta In(v_i + \zeta/\lambda), \lambda, \zeta > 0, \tag{3}$$

where an extra SPL parameter ζ is introduced in addition to λ. The corresponding optimal v_i^* is given by:

$$v_i^* = \begin{cases} 1 & l_i \le \zeta\lambda/(\zeta + \lambda) \\ 0 & l_i \ge \lambda \\ \zeta/l_i - \zeta/\lambda & otherwise \end{cases} \tag{4}$$

which is a mixture of a hard 0-1 weighting and a soft real-valued weighting.

The Bootstrap self-paced learning algorithm is shown in Algorithm 1.

3.2 Bagging with BSPL

Ensemble learning consists of building and combining multiple learners for a predictive task, it is usually used to improve overall robustness and accuracy of the problem of regression or classification [36]. Giving a dataset $D = (x_i, y_i)(|D| = m, x_i \in R^n), y_i \in R$, K additive functions are used in a ensemble model to predict the output, as shown in Equation (5).

Algorithm 1: Bootstrap with self-paced learning (BSPL) algorithm

Input: Training set $(x_i, y_i)_{i=1}^{m}$; step size $\mu > 1$.
Output: Model parameter w, v
Process: Initialize sample weights v^* and parameter λ;
while *not convergence* **do**
  Update $(w^*, v^*) = argmin_{w,v} E(w, v; \lambda)$ for finding optimal pseudo label in each selected instance randomly;
  Augment(increase or decrease) λ by step-size μ.
end
Return: w, v

$$\hat{y_i} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{5}$$

where $F = f(x) = w_{q(x)}(q : R^m \rightarrow T, w \in R^T)$ represents regression trees. Meanwhile, the minimize objective function is shown in Equation (6),

$$L(\phi) = \sum_i L(\hat{y_i}, y_i) + \sum_k \omega(f_k), \tag{6}$$

where $\omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2$, $L()$ represents the loss function, ω is complexity of the model.

Bagging (short for bootstrap aggregating) is a representative in parallel ensemble learning methods [37]. Bootstrapped replicas obtain a diversity of bagging in the training data, i.e., different subsets are randomly drawn with replacement from the whole training data. Bagging is particularly appealing when the available data is fewer. The relatively large portions of the samples (75-100%) are drawn into each subset to ensure sufficient training samples in each subset. In this case, the individual training subset has some noise instances. The SPL learners, especially the base learners, can combine sampling with SPL strategy for bagging, to reduce variance. The pseudo-code of Bagging with BSPL algorithm shows in Algorithm 2.

Algorithm 2: The Bagging with BSPL algorithm

Input: $D = (x_1, y_1), (x_2, y_2), ..., (x_m, y_m)$;
Based Learning algorithm L;
Number of learning rounds T.
Output: H,D^*
Process:
for $t \leftarrow 1$ *to* T **do**
  $\{w, v\}$=BSPL(D);
  $D_t = min_{\hat{x},\hat{y}} \sum_{i=1}^{m} v_i L(y_i, f(x_i, w)) + g(v_i, \lambda)$
  % Generate a bootstrap sample with high-confidence from D
  $h_t = \psi(D_t)$;
  % Train a base learner h_t from the BSPL sample
end
Return H(x) $= argmax_{y \in Y} I(y = h_t(x))$
% the value of I(α) is 1 if α is true, otherwise 0;
$D^* = D_t$

3.3 Random forests with BSPL

Random forest is easy to turn in the training process, due to only adjusting two parameters (the number of variables and the number of trees in the forest). The fundamental structure of random forests adds one layer of randomness to bagging. Each subset is constructed in the random forest using a bootstrap method. Each node is split using the best split among all variables in standard classification trees of random forest. In our study, each node is trained with a high-confidence subset by the BSPL method. This somewhat counter-intuitive strategy turns out to perform well compared to many other classifiers and is robust against overfitting [38]. The random forest with BSPL is shown in Algorithm 3.

Algorithm 3: Random forest with BSPL algorithm

Input: D as a training data; Depth of decision tree d.
Output: $\hat{y}$
Process:
for $k \leftarrow 1 to K$ **do**
  Obtained training data D_k^* *according Algorithm 2 input D;*
  Create a decision tree $T_k(x)$*;*
  while *not converge* $\|$ *tree_depth* $> d_{min}$ **do**
    Randomly selected n feature from N feature in all; ;
    Selected the best suitable feature and the optimal splitting point from n feature;
  end
end
Random forest $\{T_k(x)\}_1^K$ *;*
Vote $\hat{y} = \phi(X) = \sum_{k=1}^{K} T_k(x)$.

Of note, 70% of sample are regarded as the training set, and 30% of sample are regarded as the testing set. In order to assess the performance of model, $Accuracy$, $Recall$, $Precision$, and $F1âĹŠScore$ are used in this paper, among partially defined as follows,

$$Recall = \frac{TP}{TP + FN}, \quad Precision = \frac{TP}{TP + FP}, F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall},$$

where are $true\ positive(TP)$, $true\ negative(TN)$, $false\ positive(FP)$, $false\ negative(FN)$. The area under the curve (ROC-AUC) and p-value as an assess index was also used in this work [39].

4 EXPERIMENT RESULTS

In this section, we conducted the experiment analysis to test the classification performance and determine the features influenced on lung cancer to evaluate our proposed RFSPL.

4.1 Lung cancer datasets

The testing process can assess the quality of a classification model by calculating the percentage of correct predictions for a given data set. In this study, we used five lung cancer datasets to examine the validity of the proposed method. These datasets contain the expression profiles, including GSE4115, GSE33356, GSE3141, GSE8894, and GSE40419 from US National Library of Medicine National Institutes of Health. The details are shown in Table 1.

These datasets contain the expression profiles in the below Table 1. For instance, the GSE4115 dataset includes 187 samples, consisting of 97 tumor samples and 90 normal samples with 22283 genes. The results of classification

Table 1. Five public cancer datasets

Dataset	Samples	tumour samples	normal samples	No. of genes
GSE4115	187	97	90	22283
GSE33356	120	60	60	54675
GSE3141	111	58	53	54675
GSE8894	138	69	69	54675
GSE40419	164	87	77	22401

performance based on five lung cancer datasets are shown in Table 2. We demonstrate that RFSPL can achieve the best classification performances with the least absolute error, which are much smaller than those of the other models.

The performance showed in Table 2. The average accuracy with RFSPL model achieves the highest accuracy based on the five datasets, i.e., the proposed RFSPL is better than the traditional models, such as logistics regression or SVM approaches. Similarly, the best outcomes of F1-Score belong to the RFSPL model as well. Furthermore, Fig. 1 shows that comparison on Receiver Operator characteristic curves (AUC-ROC) using each method on the five datasets. In terms of AUC, results from Fig. 1 show that the proposed RFSPL model is higher than that of the competitors. In terms of sensitivity and specificity, the value of sensitivity less than specificity, suggesting that the tumor samples predict normal samples, and the practicability of the model is so poor that impact patient treatment. The value of sensitivity by RFSPL improve 18.92%, and the value of specificity improves 6.19%. In summary, in terms of accuracy, AUC, and F1-Score measures, the performance of RFSPL model is higher than that of the competitors.

Table 2. Classification performance in various prediction models based on five lung cancer datasets

Metric	Model	GSE4115	GSE33356	GSE3141	GSE8894	GSE40419	Average
Accuracy	Random Forest	0.7193	0.8889	0.5882	0.5952	0.9388	0.7461
	RFSPL	**0.8261**	**0.9472**	**0.7059**	**0.6905**	**0.9796**	**0.8299**
	Adboost	0.7544	0.9167	0.5588	0.5476	0.9388	0.7433
	Logistics regression	0.7368	0.9167	0.6176	0.5962	0.9184	0.7569
	SVM	0.6667	0.8889	0.5882	0.6429	0.6122	0.6798
F1-Score	Random Forest	0.7037	0.8889	0.4615	0.5854	0.9412	0.7161
	RFSPL	**0.8148**	**0.9444**	**0.7222**	**0.6829**	**0.9811**	**0.8291**
	Adboost	0.7586	0.9189	0.4828	0.5778	0.9412	0.7359
	Logistics regression	0.7692	0.9189	0.6061	0.5405	0.9200	0.7510
	SVM	0.6780	0.8947	0.7083	0.6341	0.7324	0.7295

4.2 Statistical analysis

To compare the statistical significance of performance of our proposed with existing methods, we adopted the Friedman test used in the study [40]. The Friedman test can measure statistical differences in the various methods, according to the performance ranking of different approaches on the five datasets. The Friedman test estimator F_F is measured in Equation (7).

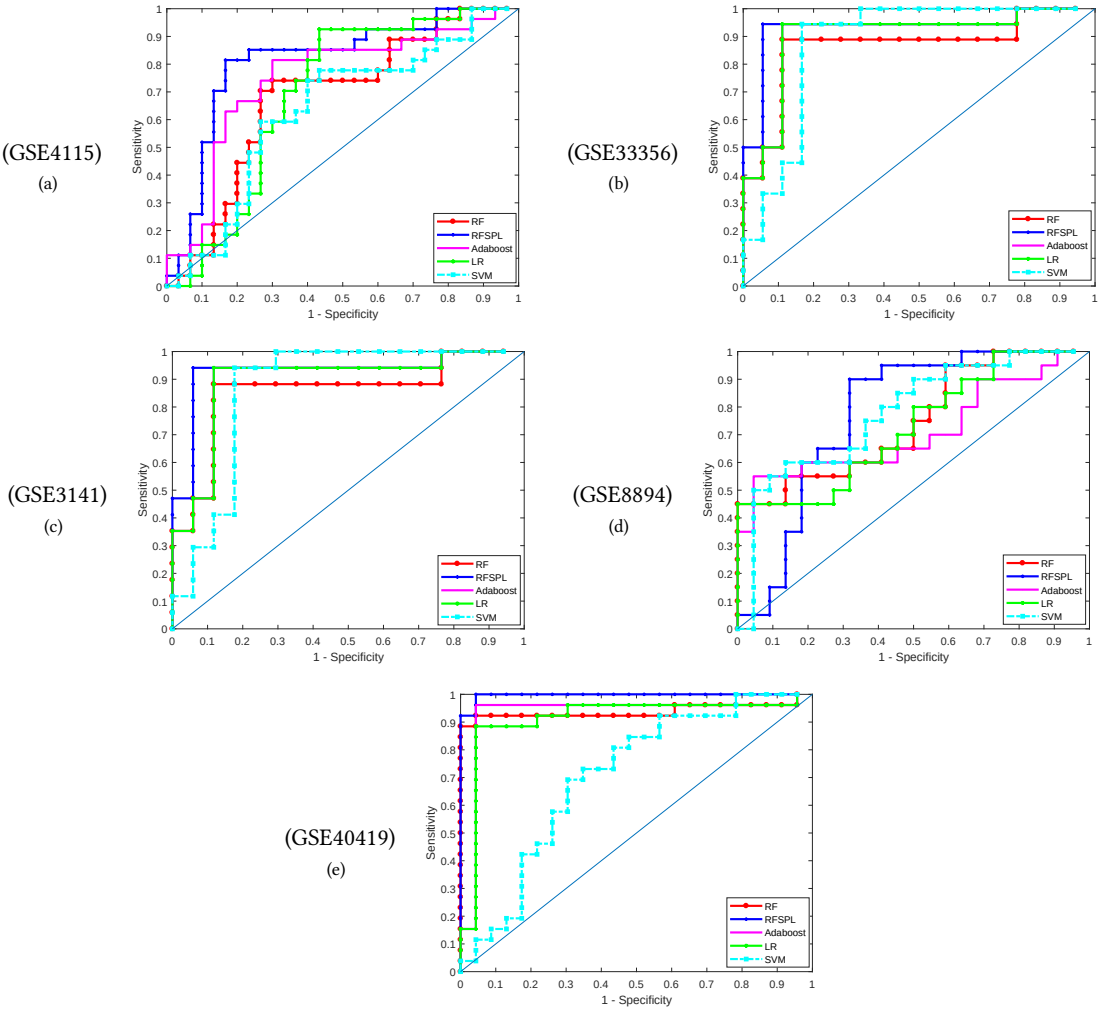

Fig. 1. AUC-ROC based on five public cancer datasets. The proposed RFSPL model always outperforms others in terms of AUC. We also know sensitivity and specificity from ROC, which indicates that the greater value of sensitivity, the greater the "tumor are judged to the tumor" (True Positive) and the smaller the "missed detection" (False Negative). Similarly, the higher the value of specificity, the higher the "health is judged to be healthy" (True Negative) and the smaller the "false alarm" (False Positive). For example, the value of sensitivity and specificity are 0.81 and 0.82 in the subfigure (a), respectively.

$$F_F = \frac{(N_D - 1)\chi_r^2}{N_D(N_M - 1) - \chi_r^2},$$
$$\chi_r^2 = \frac{12N_D}{N_M(N_M + 1)}\left(\sum_{i=1}^{N_M} R_i^2 - \frac{N_M(N_M + 1)^2}{4}\right),$$
$$R_i = \frac{\sum_{j=1}^{N_D} r_{ij}}{N_D}, \tag{7}$$

where N_M is the number of approaches, N_D is the number of datasets compared, R_i is the average ranking for the $i\text{-}th$ approach, and r^{ij} denotes the ranking of $i\text{-}th$ method on the $j\text{-}th$ dataset. For evaluated dataset, the model accuracy, AUC, and F1-Socre are ranked from one to the number of approaches, respectively. A comparison of five approaches is considered in this work. $r_{ij}^{\alpha} = 1$ represents the highest accuracy, AUC or F1-Score, and $r_{ij}^{\alpha} = 5$ represents the worst accuracy, AUC or F1-Score. For model diagnosis variance test, $r_{ij}^{\alpha} = 1$ is the lowest diagnosis accuracy, AUC or F1-Score variation, and $r_{ij}^{\alpha} = 5$ indicates the model with the highest accuracy, AUC or F1-Score variation as well. N_M-1 and (N_M-1)(N_D-1) freedom degrees of F_F is a Fisher distribution, and the confidence level is set as 0.05 in this case. N_M = 5 and N_D = 5, with degree of freedom N_M-1 = 4 and $(N_M - 1)(N_D - 1) = 16$ are applied, it can achieve a critical value of the Fisher distribution $F(4, 16) = 3.01$.

Table 3. Friedman test results on the five datasets

Test item	R_{RFSPL}	χ_r^2	$F_F(3.01)$	Decison
Accuracy	1	13.6	8.5	Positive
AUC	1.2	13.12	7.62	Positive
F1-Score	1	14.24	9.88	Positive

The accuracy, AUC, and F1-Score of Friedman test results are shown in Table 3. We know from Table 3 that average ranks of RFSPL (R_{RFSPL}) are the best value on the experimental models in terms of the multiple assessment performance measures. Additional, a significant result of the statistical difference of the ranking in the accuracy, AUC, and F1-Score are shown in Table 3.

4.3 Gene selection

Some of the genes which emerged nearly in five experiments are displayed in Figure 2 via RFSPL, and the possible functions of the part of the selected genes are also provided by searching them on the NCBI database. For instance, the function of Binds with low affinity to interleukin-13 (IL13) in the GSE4115 with IL4RA together can form a functional receptor. It also serves as an alternate accessory protein to the typical cytokine receptor gamma chain for interleukin-4 (IL4) signaling, but it cannot replace the function of IL2RG in allowing enhanced interleukin-2 (IL2) binding activity.

5 DISCUSSION AND CONCLUSION

As a typical method of ensemble technique, the random forest can address the classification problem of cancer prognosis data. Based on five lung cancer datasets, RFSPL can achieve higher accuracy, AUC, and F1-Score values in each class, compared with other methods. We believe that, during the tumorigenic process, our proposed method can select some important genes and can be considered as a reference for bioinformatics experts. As a valid learning style, our proposed method has the potential to identify in other different cancer types, including breast cancer, colorectal cancer, pancreatic cancer, and other similar applications [41, 42].

In terms of the comprehensive comparisons, the effectiveness of RFSPL is better than other classification models based on all datasets. RFSPL remains the best performance concerning the average of Accuracy, AUC, Sensitivity, and Specificity, relative to other classification models. For the GSE4115 dataset, RFSPL generates the best-ranking output, and especially the AUC of RFSPL outperforms the second best model by 5.5-19.4%. Regarding all datasets, the accuracy of RFSPL achieves 9.5-21.9% improvement over that of other classifies. For metrics F1-Score, the variances of values from different replications significantly reduced all effectiveness metrics in the RFSPL model. In this study, we

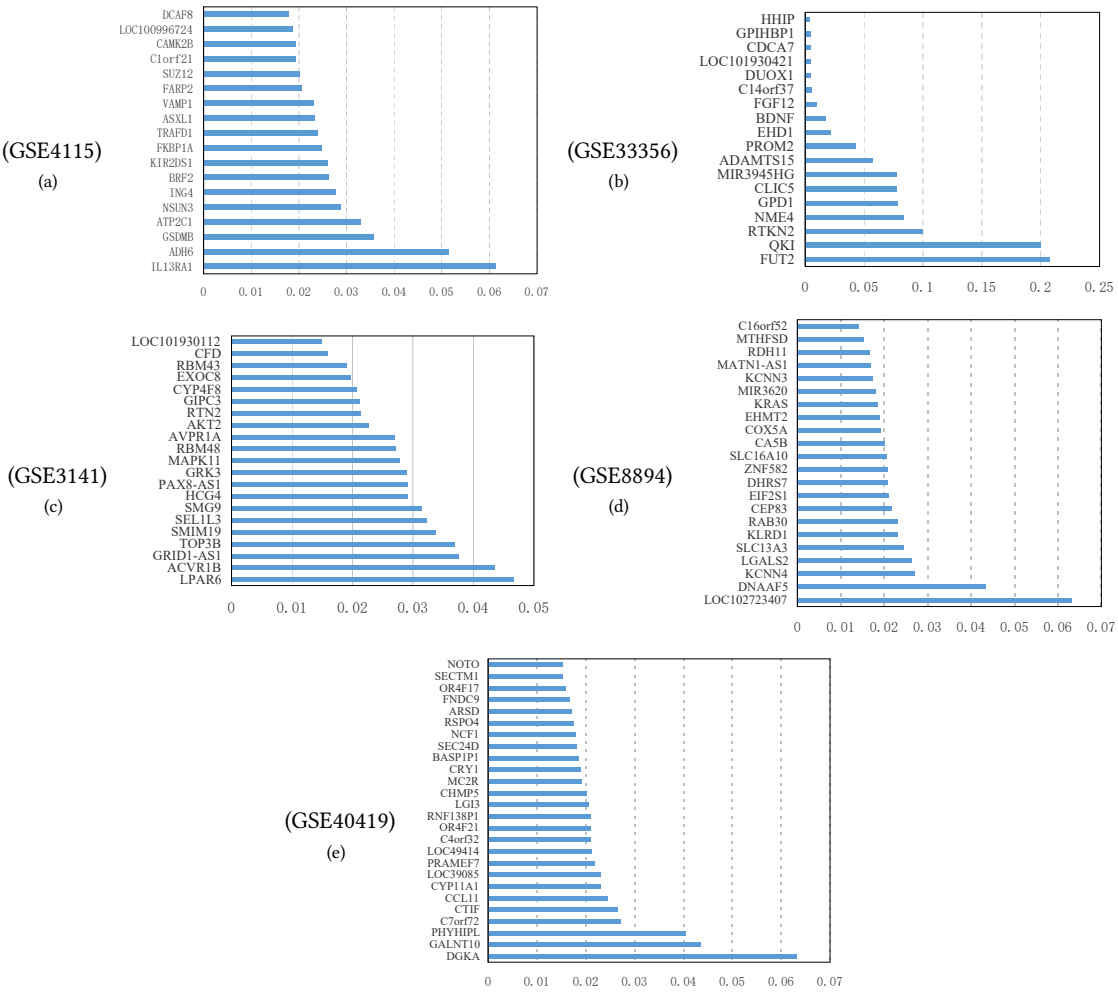

Fig. 2. Selected important genes in the five datasets. Transverse axes represent feature value in the model of RFSPL, and longitudinal axes represent the selected important genes. For p-value compared, the value of appendix A is less than 0.01 in all, which indicates whether hypothesis tests are statistically significant.

proposed a random forest with self-paced learning (RFSPL) analysis of cancer, which is important progress in applied soft computing and bioinformatics. The biological data commonly is high-dimensional and small-sample size, it causes a big challenge on mining and learning algorithms. Our proposed RFSPL method can reduce the noise of datasets and improve the classification performance to solve the dimension disaster problems.

Several aspects can be extended in future work. The SPL can learn from high- to low-quality samples, yet the limitation of learning from the diversity, which indicates that we can investigate to incorporate diversity information into the proposed method. More specific, the random forest model structures also can be widely devoted to other types' cancer datasets, and the model feature preparation process also works for some relevant feature selection and extraction techniques. Then, the other types of bagging methods can also be found in other disease diagnoses, such as thyroid

cancer, oral cancer, and diabetes. For the computation time, parallel computation techniques can accelerate the training process for the RFSPL model efficiently. What is more, it cost time and human resources if we need to achieve some useful monitoring information (category labels). Since it is easy to obtain lots of unlabeled samples, the semi-supervised classification may count for the future research direction. Additionally, deep learning has been successfully applied in various tasks, such as images. However, deep learning has some limitations; for example, deep learning is suitable in a large amount of data instead of small samples. The most sophisticated models take days to train using many expensive GPUs. There is not a strong theoretical foundation to support the outcomes, as the determining of the training method or hyperparameters is a black art.

6 ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (Grant No. 61703416), the Natural Science Foundation of Hunan Province, China (Grant No. 2018JJ3614). The authors would like to thank reviewers for their constructive comments.

REFERENCES

[1] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2018. *CA: a cancer journal for clinicians*, 68(1):7–30, 2018.

[2] Lindsey A. Torre, Rebecca L. Siegel, and Ahmedin Jemal. *Lung Cancer Statistics*. Springer International Publishing, 2016.

[3] Howard Lee and Yi Ping Phoebe Chen. Image based computer aided diagnosis system for cancer detection. *Expert Systems with Applications*, 42(12):5356–5365, 2015.

[4] Azian Azamimi Abdullah and Syamimi Mardiah Shaharum. Lung cancer cell classification method using artificial neural network. *information engineering letters*, 2(1), 2012.

[5] Z. Cai, D. Xu, Q. Zhang, J. Zhang, S. M. Ngai, and J. Shao. Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Molecular Biosystems*, 11(3):791–800, 2015.

[6] Maciej Zięba, Jakub M Tomczak, Marek Lubicz, and Jerzy Świątek. Boosted svm for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied soft computing*, 14:99–108, 2014.

[7] Golrokh Mirzaei, Anahita Adeli, and Hojjat Adeli. Imaging and machine learning techniques for diagnosis of alzheimerâĂŹs disease. *Reviews in the Neurosciences*, 27(8):857–870, 2016.

[8] Aboul Ella Hassanien, Hossam M Moftah, Ahmad Taher Azar, and Mahmoud Shoman. Mri breast cancer diagnosis hybrid approach using adaptive ant-based segmentation and multilayer perceptron neural networks classifier. *Applied Soft Computing Journal*, 14(1):62–71, 2014.

[9] Qingyong Wang, Liang-Yong Xia, Hua Chai, and Yun Zhou. Semi-supervised learning with ensemble self-training for cancer classification. In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 796–803. IEEE, 2018.

[10] M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197, 2010.

[11] Changsheng Li, Junchi Yan, Fan Wei, Weishan Dong, Qingshan Liu, and Hongyuan Zha. Self-paced multi-task learning. In *AAAI*, pages 2175–2181, 2017.

[12] Ye Tang, Yu Bin Yang, and Yang Gao. Self-paced dictionary learning for image classification. In *ACM International Conference on Multimedia*, pages 833–836, 2012.

[13] Liang-Yong Xia, Qing-Yong Wang, Zehong Cao, and Yong Liang. Descriptor selection improvements for quantitative structure-activity relationships. *International Journal of Neural Systems*, pages 1–16, 2019.

[14] Min Wei Huang, Chih Wen Chen, Wei Chao Lin, Shih Wen Ke, and Chih Fong Tsai. Svm and svm ensembles in breast cancer prediction. *Plos One*, 12(1):e0161501, 2017.

[15] D. L. Langer, Van Der Kwast Th, A. J. Evans, J Trachtenberg, B. C. Wilson, and M. A. Haider. Prostate cancer detection with multi-parametric mri: logistic regression analysis of quantitative t2, diffusion-weighted imaging, and dynamic contrast-enhanced mri. *Journal of Magnetic Resonance Imaging*, 30(2):327–334, 2010.

[16] Zehong Cao, Weiping Ding, Yu-Kai Wang, Farookh Khadeer Hussain, Adel Al-Jumaily, and Chin-Teng Lin. Effects of repetitive ssveps on eeg complexity using multiscale inherent fuzzy entropy. *Neurocomputing*, 2019.

[17] Dursun Delen. Analysis of cancer data: a data mining approach. *Expert Systems*, 26(1):100–112, 2010.

[18] Dursun Delen, Glenn Walker, and Amit Kadam. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine*, 34(2):113–127, 2005.

[19] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[20] Dustin Baumgartner and Gursel Serpen. Performance of global–local hybrid ensemble versus boosting and bagging ensembles. *International Journal of Machine Learning and Cybernetics*, 4(4):301–317, 2013.

[21] Dieu Tien Bui, Tien-Chung Ho, Biswajeet Pradhan, Binh-Thai Pham, Viet-Ha Nhu, and Inge Revhaug. Gis-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with adaboost, bagging, and multiboost ensemble frameworks. *Environmental Earth Sciences*, 75(14):1101, 2016.

[22] VF Rodriguez-Galiano, M Chica-Olmo, F Abarca-Hernandez, Peter M Atkinson, and C Jeganathan. Random forest classification of mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Remote Sensing of Environment*, 121:93–107, 2012.

[23] Juan José Rodriguez, Ludmila I Kuncheva, and Carlos J Alonso. Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1619–1630, 2006.

[24] RamÃşn DÃŋazuriarte and Sara Alvarez De AndrÃĺs. Gene selection and classification of microarray data using random forest. *Bmc Bioinformatics*, 7(1):1–13, 2006.

[25] Mohammed Khalilia, Sounak Chakraborty, and Mihail Popescu. Predicting disease risks from highly imbalanced data using random forest. *Bmc Medical Informatics and Decision Making*, 11(1):51–51, 2011.

[26] Fan Ma, Deyu Meng, Qi Xie, Zina Li, and Xuanyi Dong. Self-paced co-training. In *International Conference on Machine Learning*, pages 2275–2284, 2017.

[27] Dingwen Zhang, Junwei Han, Long Zhao, and Deyu Meng. Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework. *International Journal of Computer Vision*, pages 1–18, 2018.

[28] Deyu Meng, Qian Zhao, and Lu Jiang. A theoretical understanding of self-paced learning. *Information Sciences*, 414:319–328, 2017.

[29] Jesper Holst Pedersen, Witold Rzyman, Giulia Veronesi, Thomas A DâĂŹAmico, Paul Van Schil, Laureano Molins, Gilbert Massard, and Gaetano Rocco. Recommendations from the European Society of Thoracic Surgeons (ESTS) regarding computed tomography screening for lung cancer in Europe. *European Journal of Cardio-Thoracic Surgery*, 51(3):411–420, 02 2017.

[30] Rahul Paul, Samuel H Hawkins, Matthew B Schabath, Robert J Gillies, Lawrence O Hall, and Dmitry B Goldgof. Predicting malignant nodules by fusing deep features with classical radiomics features. *Journal of Medical Imaging*, 5(1):011021, 2018.

[31] Ahmed Hosny, Chintan Parmar, Thibaud P Coroller, Patrick Grossmann, Roman Zeleznik, Avnish Kumar, Johan Bussink, Robert J Gillies, Raymond H Mak, and Hugo JWL Aerts. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS medicine*, 15(11):e1002711, 2018.

[32] Rahul Paul, Samuel H Hawkins, Yoganand Balagurunathan, Matthew B Schabath, Robert J Gillies, Lawrence O Hall, and Dmitry B Goldgof. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. *Tomography*, 2(4):388, 2016.

[33] Ravi K Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A Helvie, Caleb Richter, and Kenny Cha. Cross-domain and multi-task transfer learning of deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 105750Q. International Society for Optics and Photonics, 2018.

[34] Te Pi, Xi Li, Zhongfei Zhang, Deyu Meng, Fei Wu, Jun Xiao, and Yueting Zhuang. Self-paced boost learning for classification. In *IJCAI*, pages 1932–1938, 2016.

[35] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. Self-paced curriculum learning. In *AAAI*, volume 2, page 6, 2015.

[36] Bin Liu, Ren Long, and Kuo-Chen Chou. idhs-el: identifying dnase i hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. *Bioinformatics*, 32(16):2411–2418, 2016.

[37] Jean Yves Audibert. Bagging predictors. *Annales De Linstitut Henri Poincare Probability and Statistics*, 40(6):685–736, 2004.

[38] L. Breiman. Random forests, machine learning 45. *Journal of Clinical Microbiology*, 2:199–228, 2001.

[39] Zehong Cao, Chin-Teng Lin, Kuan-Lin Lai, Li-Wei Ko, Jung-Tai King, Kwong-Kum Liao, Jong-Ling Fuh, and Shuu-Jiun Wang. Extraction of ssveps-based inherent fuzzy entropy using a wearable headband eeg in migraine patients. *IEEE Transactions on Fuzzy Systems*, 2019.

[40] M Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Publications of the American Statistical Association*, 32(200):675–701, 1939.

[41] Zehong Cao and Chin-Teng Lin. Inherent fuzzy entropy for the improvement of eeg complexity evaluation. *IEEE Transactions on Fuzzy Systems*, 26(2):1032–1035, 2017.

[42] Zehong Cao, Chun-Hsiang Chuang, Jung-Kai King, and Chin-Teng Lin. Multi-channel eeg recordings during a sustained-attention driving task. *Scientific data*, 6, 2019.