Hindawi
BioMed Research International
Volume 2023, Article ID 5333361, 13 pages
https://doi.org/10.1155/2023/5333361

Research Article

Identification of Smoking-Associated Transcriptome Aberration in Blood with Machine Learning Methods

FeiMing Huang,[1] QingLan Ma,[1] JingXin Ren,[1] JiaRui Li,[2] Fen Wang,[3] Tao Huang,[4,5] and Yu-Dong Cai[1]

[1]*School of Life Sciences, Shanghai University, Shanghai 200444, China*
[2]*Advanced Research Computing, University of British Columbia, Vancouver, Canada*
[3]*Guangdong AIB Polytechnic College, Guangzhou 510507, China*
[4]*Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[5]*CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China*

Correspondence should be addressed to Tao Huang; tohuangtao@126.com and Yu-Dong Cai; cai_yud@126.com

Received 3 September 2022; Revised 15 December 2022; Accepted 15 December 2022; Published 4 January 2023

Academic Editor: Bilal Alatas

Long-term cigarette smoking causes various human diseases, including respiratory disease, cancer, and gastrointestinal (GI) disorders. Alterations in gene expression and variable splicing processes induced by smoking are associated with the development of diseases. This study applied advanced machine learning methods to identify the isoforms with important roles in distinguishing smokers from former smokers based on the expression profile of isoforms from current and former smokers collected in one previous study. These isoforms were deemed as features, which were first analyzed by the Boruta to select features highly correlated with the target variables. Then, the selected features were evaluated by four feature ranking algorithms, resulting in four feature lists. The incremental feature selection method was applied to each list for obtaining the optimal feature subsets and building high-performance classification models. Furthermore, a series of classification rules were accessed by decision tree with the highest performance. Eventually, the rationality of the mined isoforms (features) and classification rules was verified by reviewing previous research. Features such as isoforms ENST00000464835 (expressed by LRRN3), ENST00000622663 (expressed by SASH1), and ENST00000284311 (expressed by GPR15), and pathways (cytotoxicity mediated by natural killer cell and cytokine–cytokine receptor interaction) revealed by the enrichment analysis, were highly relevant to smoking response, suggesting the robustness of our analysis pipeline.

1. Introduction

Tabaco smoking is among the leading causes of premature mortality in the world, and this condition can be avoided [1]. It has been demonstrated to be associated with human diseases such as respiratory disease, cardiovascular, cancer, and gastrointestinal (GI) disorders [2–5]. According to the World Health Organization (WHO), smoking causes over US$500 billion economic loss globally annually.

Although cigarette smoke is deemed as the main risk factor for chronic obstructive pulmonary disease, which increases oxidative stress in the airway epithelium, the pathogenesis of most smoking-induced diseases has not been fully determined [5]. A recent article systematically reviewed previous studies on smoking-associated DNA methylation and the alteration of gene expression in human blood samples, in which 1,758 genes with differentially methylated sites and differentially expressed genes between smokers and nonsmokers were reported [6]. Therefore, gene expression alterations are important for smoking response. Considering that alternative splicing is applied by up to 95% of human genes for producing proteins with different functions, a recent

study has focused on identifying smoking-associated isoforms and found that 3′ UTR lengthening was widely associated with cigarette smoking [7, 8]. A total of 945 differentially expressed isoforms were identified in this study by using the classic statistic method. Machine learning could be applied without preexisting knowledge to analyze RNA-seq data and deal with a large number of variables in a much smaller sample size [9].

In the present study, we applied the Boruta [10] and four feature ranking algorithms to identify the isoforms with important roles in distinguishing smokers from former smokers based on the expression data of 85,437 isoforms on current and former smokers collected in the previous study [8]. The incremental feature selection (IFS) method [11] was employed to further analyze the results yielded by above methods for extracting optimal features, building efficient classification models, and interesting classification rules. The literature review and further comparison of these features identified by four feature ranking methods demonstrated strong biological relevance of these features (isoforms) with smoking response.

2. Materials and Methods

2.1. Data. The RNA-seq data in whole-blood samples on 454 current and 767 former smokers were accessed from the Gene Expression Omnibus (GEO) database under the accession number GSE171730 [8]. We separated the samples into two classes based on their smoking history: current smokers and former smokers. Each sample contains 85,437 transcript features. Such data was deeply analyzed by investigating a binary classification problem containing two classes (current smokers and former smokers) and 85,437 features. The purpose of this study was to discover essential transcript features that can classify smokers and reveal different patterns for current and former smokers.

2.2. Boruta Feature Filtering. As lots of features were involved for each smoker sample and only a few of them have strong associations with smoke, the irrelevant features should be screened out first and excluded. Here, we selected the Boruta method [10, 12, 13] to complete this task.

The Boruta is a feature selection method using random forest (RF), which can be used to confirm whether original features are statistically more important than the random features in the prediction. Given a dataset, the Boruta first generates a random feature for each original one. Its values are produced by shuffling numbers under the original feature. RF is performed on such extended dataset to evaluate the importance of all original and random features. Original features with importance remarkably better than the highest importance on random features are labelled as confirmed, while those that perform worse are categorized as rejected. Confirmed features are excluded from the dataset, and the updated dataset is put into the next round. After a number of rounds, features in the rejection region are dropped, and confirmed features in each round are kept. Unlike wrapper approaches, which try to locate some powerfully relevant features, the Boruta chooses features that are strongly or weakly important to achieve the best classification accuracy.

The Boruta program downloaded from https://github.com/scikit-learn-contrib/boruta_py was used in the present study for analyzing the RNA-seq data. Default settings were adopted.

2.3. Feature Ranking Algorithms. Relevant features were selected by the Boruta method. However, their contributions for prediction were not the same. To clearly classify features with their importance, four feature ranking algorithms were employed, including max-relevance and min-redundancy (mRMR) [14], Monte Carlo feature selection (MCFS) [15], light gradient boosting machine (LightGBM) [16], and least absolute shrinkage and selection operator (LASSO) [17]. As each algorithm has its own defects and merits, the bias may be produced by only using one feature ranking algorithm. Each algorithm can give a part of contributions for discovering essential features. A full and systemic evaluation on features can be obtained by the usage of different algorithms. A brief description on these algorithms was as below.

2.3.1. Max-Relevance and Min-Redundancy. The mRMR algorithm aims at determining the feature subset that has the highest correlation with the target variable and the lowest correlation between the features in this set [14, 18–21]. mRMR uses mutual information to quantify feature-target and feature-feature correlations. However, it is difficult to obtain such feature subset. mRMR adopts a heuristic way to generate a feature list, which is constructed by repeatedly choosing one feature with trade-off on maximum correlation to the target variable and minimum redundancy to features that have been chosen. Such list was termed as mRMR feature list. We utilized the mRMR tool retrieved from Hanchuan Peng's web (http://home.penglab.com/proj/mRMR/ in this work), which was run under default parameters.

2.3.2. Monte Carlo Feature Selection. MCFS is a method for ranking features by randomly selecting features to build multiple decision trees [15]. It is commonly used to process biological data [22–24]. In the present research, m transcript features are chosen at random to build t classification trees for s times. Each tree is trained and tested using randomly selected training and test data from the entire dataset. As a result, $s \times t$ classification trees are built. The relative significance (RI) of a given feature g is assessed based on how many trees select this feature and how much it contributes to prediction:

$$\mathrm{RI}_g = \sum_{\tau=1}^{\mathrm{st}} (\mathrm{wAcc})^u \sum_{n_g(\tau)} \mathrm{IG}\left(n_g(\tau)\right) \left(\frac{\mathrm{no.in}\, n_g(\tau)}{\mathrm{no.in}\, \tau}\right)^v, \quad (1)$$

where wAcc is the weighted accuracy, $\mathrm{IG}(n_g(\tau))$ represents the information gain (IG) of node $n_g(\tau)$, $(\mathrm{no.in}\, n_g(\tau))$ denotes the number of samples in node $n_g(\tau)$, and $(\mathrm{no.in}\, \tau)$ stands for the number of samples in the tree root. u and v are two parameters. In this study, we used the MCFS program obtained from http://www.ipipan.eu/staff/m.draminski/mcfs

.html, which was performed with default parameters. The list generated by MCFS was called the MCFS feature list.

2.3.3. Light Gradient Boosting Machine. LightGBM algorithm is an ensemble method using gradient boosting framework. It improves the gradient boosting decision tree and has the advantages, such as high efficiency, support for parallelism, low GPU memory consumption, and large-scale data processing [16]. LightGBM can estimate the feature importance based on the times appearing in the ensemble trees, and the high frequency indicates the importance of the feature. The study adopted the program of LightGBM (https://lightgbm.readthedocs.io/en/latest/) in Python, and we ran it with default parameters. For an easy description, the list produced by LightGBM was described as the LightGBM feature list.

2.3.4. Least Absolute Shrinkage and Selection Operator. In 1996, Robert Tibshirani proposed a new feature selection technique called the minimum absolute compression method or LASSO [17]. It employs the L1 paradigm to develop a penalty function, which can selectively eliminate features by assigning a higher penalty on features with higher coefficients and greater prediction errors, leading to a model with fewer features and that effectively reduces overfitting. Clearly, features with high coefficients do not contribute favorably to the prediction and should be scaled down. Consequently, features can be ranked according to their coefficients. In this study, the LASSO package collected in the scikit-learn was adopted, which was performed using its default parameters. Likewise, the list derived by LASSO was called the LASSO feature list.

2.4. Incremental Feature Selection. The feature ranking algorithms only sorted features in lists. It was still a problem for the selection of essential features. Therefore, the IFS method was employed to determine essential features from each list based on given classification algorithms [11], which can be further used to build the efficient classification models [25–28]. From each feature list, the IFS method first divides the feature list into n feature subsets whose feature numbers differ by 1 in turn. Subsequently, various feature subsets were then used to construct the models using a single classification algorithm, and the effectiveness of classification was assessed using tenfold cross-validation [29]. The optimal model can then be identified based on its performance. In addition, features employed in best model were referred to as the optimal features.

2.5. Synthetic Minority Oversampling Technique. In the investigated data, former smokers were 1.7 times as many as current smokers. It was not a balanced dataset, which may produce bias in a model that is built directly based on it. In view of this, we employed the powerful oversampling algorithm, synthetic minority oversampling technique (SMOTE) [30, 31], to tackle this problem. To equalize the distribution of data among various classes, this method synthesizes new samples of a minority class. It selects one sample from the minority class as a seed sample and then randomly chooses one of its k closest neighbors. The following is the synthesis equation:

$$s = x + \beta(x - y), \tag{2}$$

where x stands for the coordinates of the seed sample in the Euclidean space, y represents the coordinates of a randomly selected k-nearest neighbor of x, and β is an arbitrary number between 0 and 1.

Here, the SMOTE program reported at https://github.com/scikitlearn-contrib/imbalanced-learn was used. The default parameters were applied to run the program.

2.6. Classification Algorithm. The IFS technique required the use of at least one classification algorithm to construct the model on each feature subset. Four classification algorithms were used in this instance: the decision tree (DT), the random forest (RF), the k-nearest neighbor (KNN), and the support vector machine (SVM) [32–35]. These classification algorithms are theoretically sound and are widely used in machine learning.

2.6.1. k-Nearest Neighbor. KNN is a classic classification algorithm. For a test sample, k training samples that are nearest to the test sample are chosen based on a distance metric, and the class of the test sample is established based on the classes of these k training samples.

2.6.2. Random Forest. RF constructs many DTs to form a forest by using bootstrap aggregation. Besides the sample selection, features are also randomly selected when constructing each DT. When classifying a sample, each tree makes a prediction, and the class with the highest votes is considered the final decision of RF.

2.6.3. Support Vector Machine. SVM is a powerful classification algorithm. According to the distributions of training samples, it finds the optimal hyperplane for classifying samples in different classes. In many instances, it is difficult to build such a hyperplane in the original feature space. In order to map samples into a new space with a higher dimension, where the hyperplane is simple to construct, the kernel approach is used. We can establish its class based on which side of the hyperplane it is on.

2.6.4. Decision Tree. DT is a relatively simple classification algorithm that is utilized as a predictive model in medical diagnostics and biomarker screening [36]. Different from above three algorithms, the classification principle is much easier to be understood, which is the greatest merit of DT. By learning training samples, a tree is built. Such tree gives a completely open procedure for classifying test samples. This provides opportunities to figure out its classification principle. Thus, it is deemed as a white-box algorithm. A set of rules can also be used to represent DT in addition to its tree-like structure. Each rule shows a route from the root to a single leaf. For various classes, these rules can suggest multiple patterns.

In this study, the above algorithms are implemented by using the scikit-learn package written in Python [37].

2.7. Performance Evaluation. F1-measure is the primary metric utilized in this study to assess how well classification

models perform [38–40]. In fact, it is computed by combining the values of precision and recall. It indicates that the model is good when the F1-measure is high. These measurements can be calculated as follows:

$$\begin{aligned} \text{F1} - \text{measure} &= \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}, \\ \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \end{aligned} \tag{3}$$

where true positive (TP) is the number of positive samples that are correctly labelled as positive samples, false positive (FP) is the number of negative samples that are incorrectly labelled as positive samples, and false negative (FN) indicates the number of positive samples that are mistakenly labelled as negative samples.

Besides, the following measurements are also widely used for binary classification, including sensitivity (SN), specificity (SP), accuracy (ACC), and the Matthews correlation coefficient (MCC), where SN is same as recall, and others can be computed by

$$\begin{aligned} \text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}}, \\ \text{ACC} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \\ \text{MCC} &= \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \end{aligned} \tag{4}$$

where TP, FP, and FN are same as those in the above paragraph and TN indicates the number of negative samples that are correctly labelled as negative samples. They were also provided in this study to display a more complete evaluation on different models.

2.8. Functional Enrichment Analysis. Through the above machine learning algorithms, some essential features (transcripts) can be screened out. These transcripts were converted to the corresponding genes via "bitr" from the clusterProfiler package in R [41]. Then, the enrichment analysis of gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways was performed on these genes. The FDR, used to adjust the p value, was picked up as the key indicator for selecting enriched GO terms and KEGG pathways. Its cutoff was set to 0.05.

3. Results

In the current study, we developed a computational pipeline, which combined some feature analysis methods with the classification models, to investigate the RNA-seq data of two kinds of smokers. Figure 1 shows the entire procedures. The detailed results were displayed in this section.

3.1. Results of Feature Selection Methods. The 85,437 original transcript features were first filtered using the Boruta. 370 features were screened out, which were deemed to be highly correlated with the target variables. Subsequently, the 370 transcript features were ranked using four feature ranking algorithms, resulting in four feature lists (mRMR, MCFS, LightGBM, and LASSO feature lists), which are shown in Table S1. In Discussion, we performed a biological analysis of some top-ranked features yielded by four ranking algorithms.

3.2. Results of IFS Method and Feature Intersection. Four feature lists were obtained by using four feature ranking algorithms. Each of them was fed into the IFS computational framework and processed in the same manner. First, 370 feature subsets were generated, each containing a few of the most prominent features from the original feature list. One of the four classification algorithms listed in Classification Algorithm was used to build a model for each feature subset, and it was then tested using tenfold cross-validation. The model's performance was mainly measured by F1-measure. To display how the performance changed with different feature subsets, an IFS curve was created. All these curves under various feature lists and classification algorithms are shown in Figures 2–5, and the detailed performance is listed in Table S2.

For the IFS results on the mRMR feature list, Figure 2 shows the performance of four classification algorithms under various feature subsets. It can be shown that when the top 309, 49, 215, and 341 features in the mRMR feature list were adopted, DT, KNN, RF, and SVM produced the greatest F1-measure values of 0.766, 0.796, 0.840, and 0.852. These features made up the best feature set for the corresponding classification algorithm and can be used to create the best classification models. Table 1 lists the specific overall performance of these models, including ACC, MCC, and F1-measure, while Figure 6(a) illustrates other measurements (SN, SP, and precision). Clearly, the optimal SVM model provided the best performance among all optimal models. Thus, we set its optimal features (top 314 features in the mRMR feature list) as the optimal features extracted from the mRMR feature list.

For the IFS results on the MCFS feature list, the IFS curves are illustrated in Figure 3. With the same argument, the RF model using top 145 features yielded the best performance according to Figures 3 and 6 and Table 1. These 145 features constituted the optimal features extracted from the MCFS feature list.

Referring to the IFS results on the rest feature lists (LightGBM and LASSO feature lists), we can conclude that RF model with top 22 features in the LightGBM feature list and SVM model with top 369 features in the LASSO feature list provided the highest performance, refer to Figures 4–6 and Table 1. These features made up the optimal feature sets derived from the above two feature lists.

With the above analysis, four optimal feature subsets were obtained from four feature lists. The Venn diagram was plotted to show the relationships between them, as illustrated in Figure 7. Detailed intersection results of four

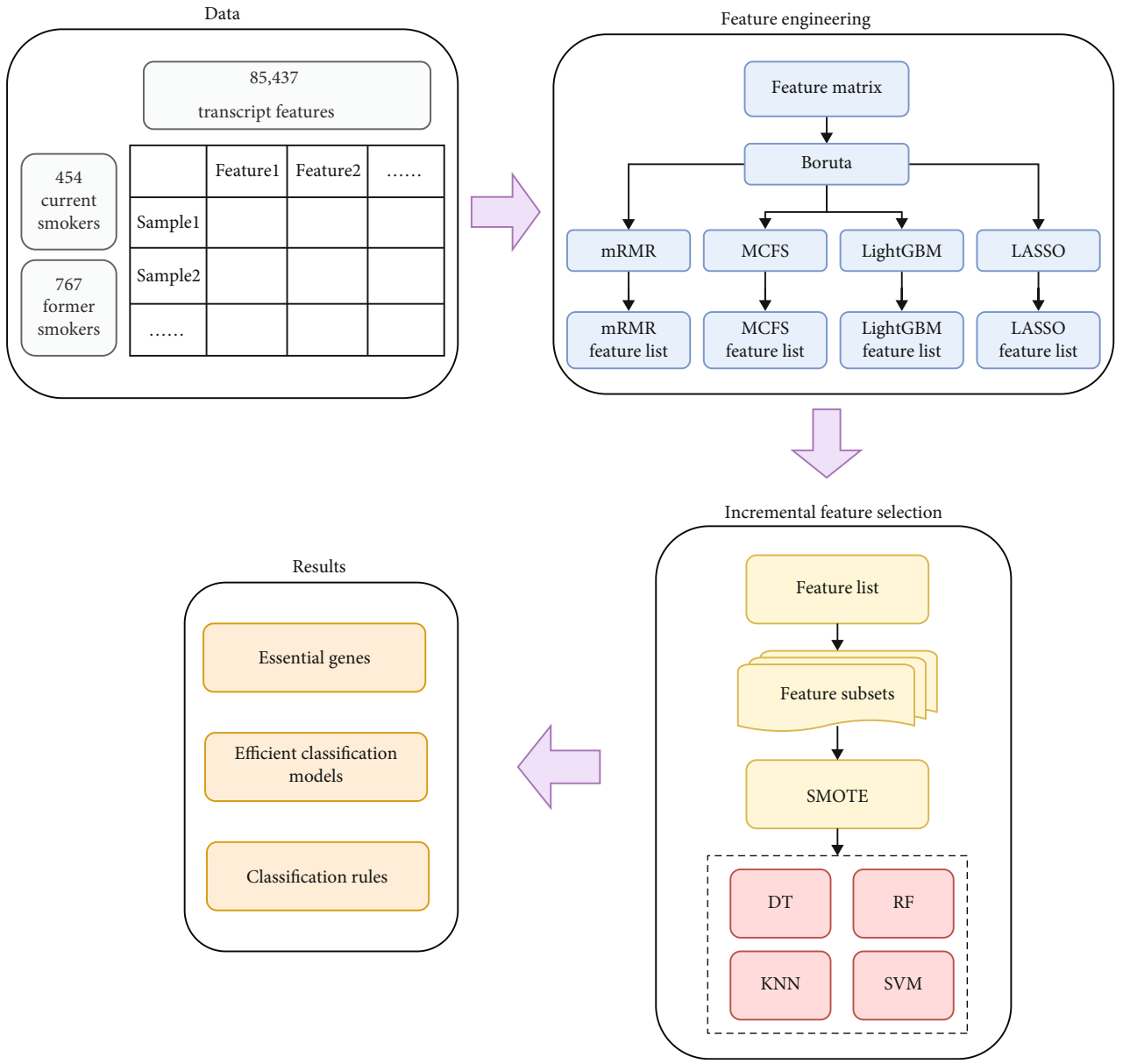

Figure 1: Process flow diagram for the full analysis. After being first filtered by the Boruta, the 85,437 transcript features from the 1221 smokers were then sorted by feature importance using four feature ranking algorithms: mRMR, MCFS, LightGBM, and LASSO. The IFS computational framework was then performed based on four sorted feature sets, and four effective classification algorithms were used in this process. Eventually, the essential genes (converted from important features) and the classification rules were extracted.

optimal feature subsets are shown in Table S3. We can see that 15 transcript features occurred in all four optimal feature subsets, which were deemed to be highly associated with smoking and would be analyzed biologically in the subsequent discussion section.

3.3. Classification Rules. According to the results listed in the above section, DT generally yielded the lowest performance. However, it can offer more insights than other three classification algorithms as it is a white-box algorithm. Thus, DT was picked up again in this section. The IFS findings show that the numbers of optimal features for DT on mRMR, MCFS, LightGBM, and LASSO feature lists were 309, 64, 32, and 370, respectively. We used these features to represent all smoker samples, and DT was applied to such data for generating four trees. From these trees, four sets of classification rules were accessed, as shown in Table S4. The numbers of rules used to distinguish two types of samples in each set of rules are shown in Figure 8. The detailed analysis of the rules that can distinguish the two classes with the largest number of samples would be provided in Discussion.

3.4. Results of Enrichment Analysis. Four optimal feature subsets extracted from four feature lists were merged into one set, resulting in 370 features. The corresponding genes to the transcripts in the merged set were picked up for

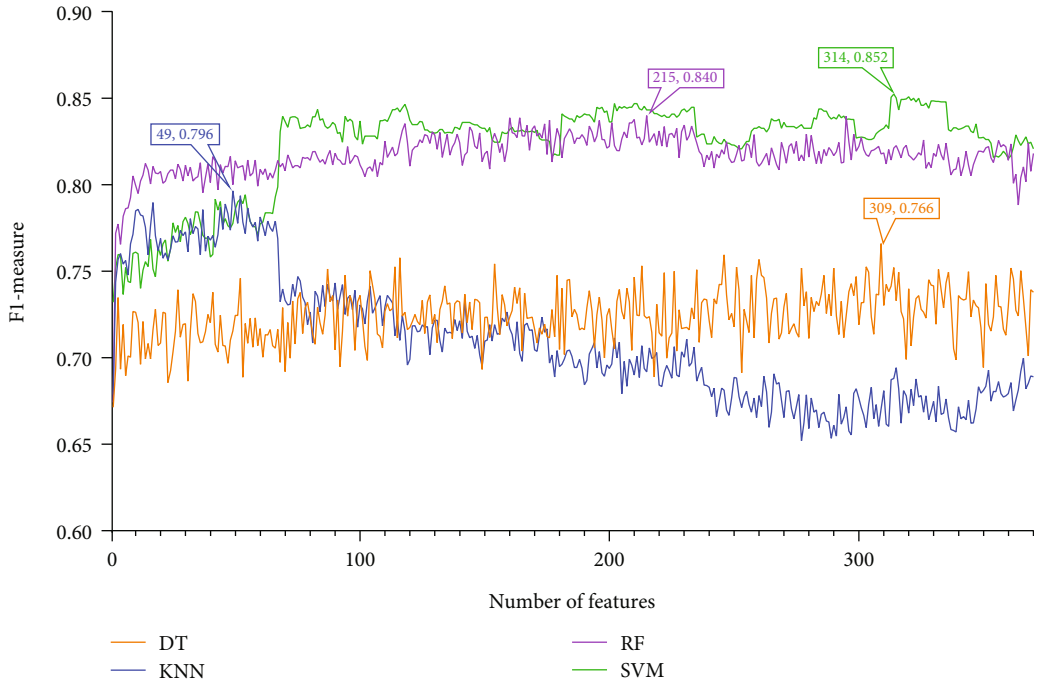

Figure 2: IFS curves for showing the performance of four classification algorithms according to F1-measure on the mRMR feature list.

enrichment analysis. The significant enrichment results after FDR estimation are shown in Table S5. Furthermore, the top 5 entries of each of the three parts of GO enrichment results and the top 5 pathways of KEGG enrichment results were selected for visual presentation, as shown in Figure 9. GO enrichment results show that immunoglobulin complex, complement activation, and humoral immune response mediated by circulating were significantly enriched by the identified genes. The cytotoxicity caused by natural killer cells, cytokine-cytokine receptor interaction, and viral protein-cytokine interaction was all considerably enriched according to the KEGG enrichment. Further analysis indicated the association of these results with different smoking populations in Discussion.

4. Discussion

4.1. Functional Enrichment Analyses. To illustrate the biological significance of the features and rules identified in this study, we clustered the 370 features with GO terms and KEGG pathway, respectively. As expected, all the top-ranked GO terms are involved in immune response such as complement activation, humoral immune response, and immunoglobulin complex. Cigarette smoke exposure can affect immune response significantly and may cause multiple diseases [42]. Similarly, most of the enriched KEGG pathways also represented immune responses, such as natural killer cell-mediated cytotoxicity, interaction of cytokine-cytokine receptor, and interaction between viral protein and cytokine receptor. These pathways have been reported to be related to cigarette smoke in previous studies. For instance, cigarette smoke inhibits NK cell ability to kill tumor cell lines and increases the amount of proinflammatory cytokines while downregulates the anti-inflammatory cytokines [43, 44]. Smoke was also found to increase the viral load and cell death, which leads to more severe viral myocarditis [45]. Interestingly, the features were also enriched in pathways of diseases, such as graft-versus-host disease, autoimmune thyroid disease, and type I diabetes mellitus. Smoking was found to be able to both increase and decrease the risks of autoimmune thyroid diseases [46]. Moreover, the animal models treated with nicotine were found to alter the expression of pancreatic cytokine, which leads to a lower incidence of type I [47].

4.2. Analysis of Highly Ranked Transcripts in Four Algorithms. As shown above, these 370 features were ranked by four algorithms, and each of which produced a set of optimal features. We compared the four sets and investigated their biological significance related to smoking. Interestingly, the optimal features by LASSO included 369 features, and the mRMR produced an optimal set of 314 features, indicating that these two methods tend to capture features comprehensively. By contrast, the optimal feature sets have 22 and 145 features by LightGBM and MCFS, respectively, and the optimal models provided higher F1-measure (Table 1) than the optimal models on the optimal feature sets of LASSO and mRMR, indicating that these two feature ranking algorithms can capture the features with higher effects.

Meanwhile, 15 features were shared by all four optimal feature sets, suggesting the significance of these features in responding to smoking. Based on the review of literature, we found that 12 features were proved to be relevant to

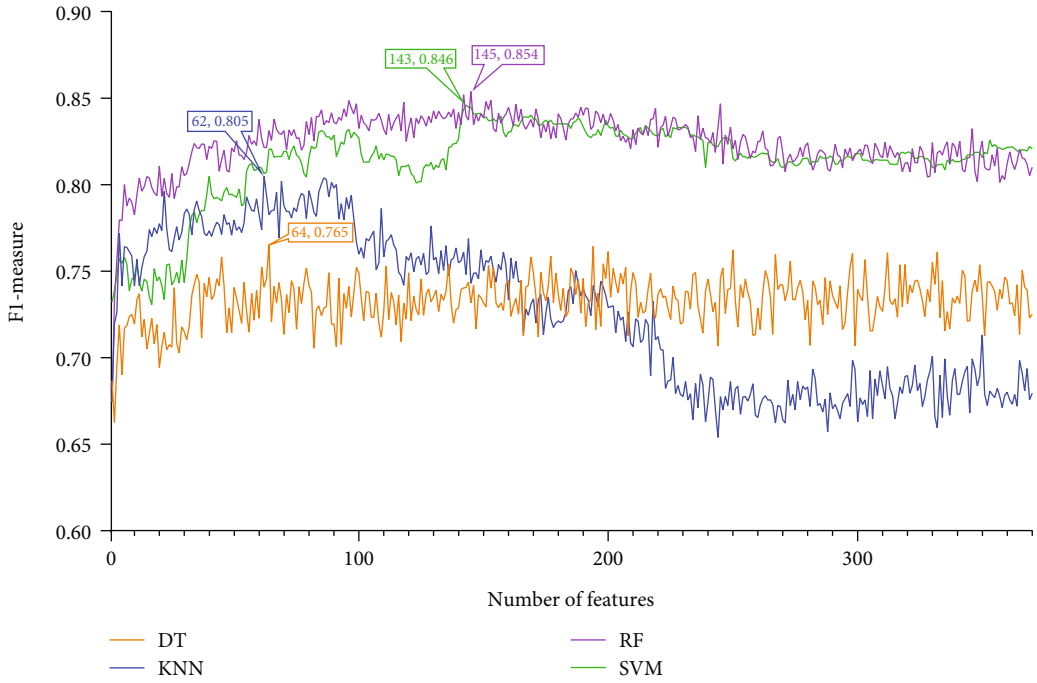

Figure 3: IFS curves for showing the performance of four classification algorithms according to F1-measure on the MCFS feature list.

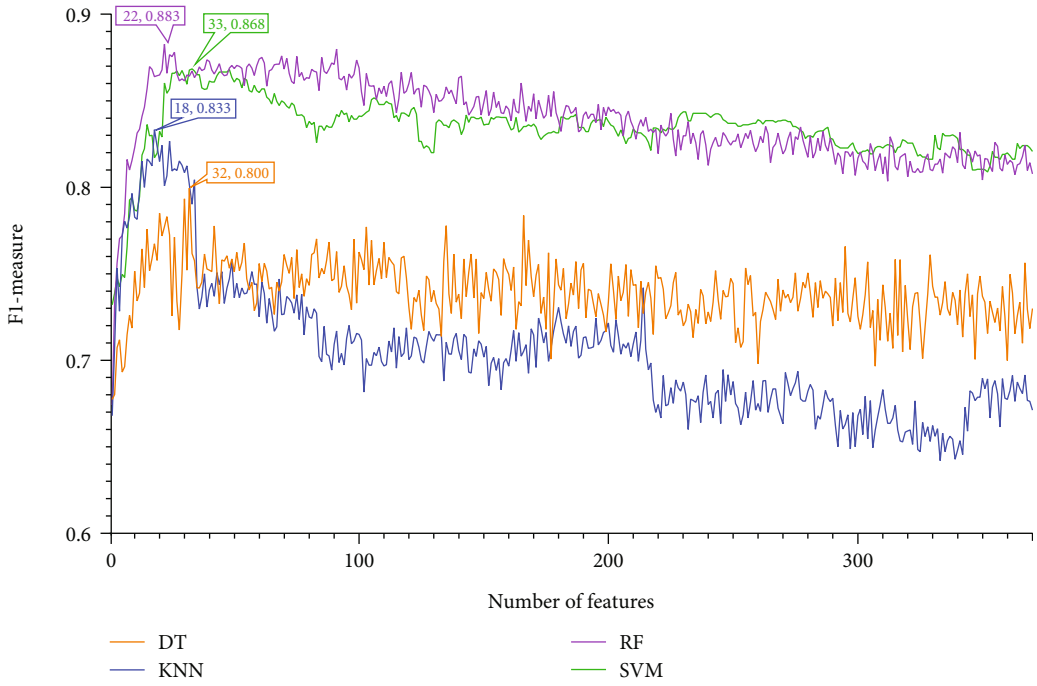

Figure 4: IFS curves for showing the performance of four classification algorithms according to F1-measure on the LightGBM feature list.

smoking. Features ENST00000464835 and ENST00000308478 are two transcripts of gene *LRRN3*, which are most significantly associated with smoking [48]. ENST00000622663 and ENST00000367467 were two transcripts of the gene *SASH1*, whose expression changes were associated with smoking in human monocytes [49]. Two GWAS studies found that gene *GPR15* is strongly associated with smoking, and its expressed transcript ENST00000284311 is among the optimal features by all four methods in this study [50, 51]. Other features including ENST00000392054 (expressed by PID1), ENST00000586582 (expressed by SEMA6B), ENST00000393590 (expressed by P2RY6), ENST00000422622

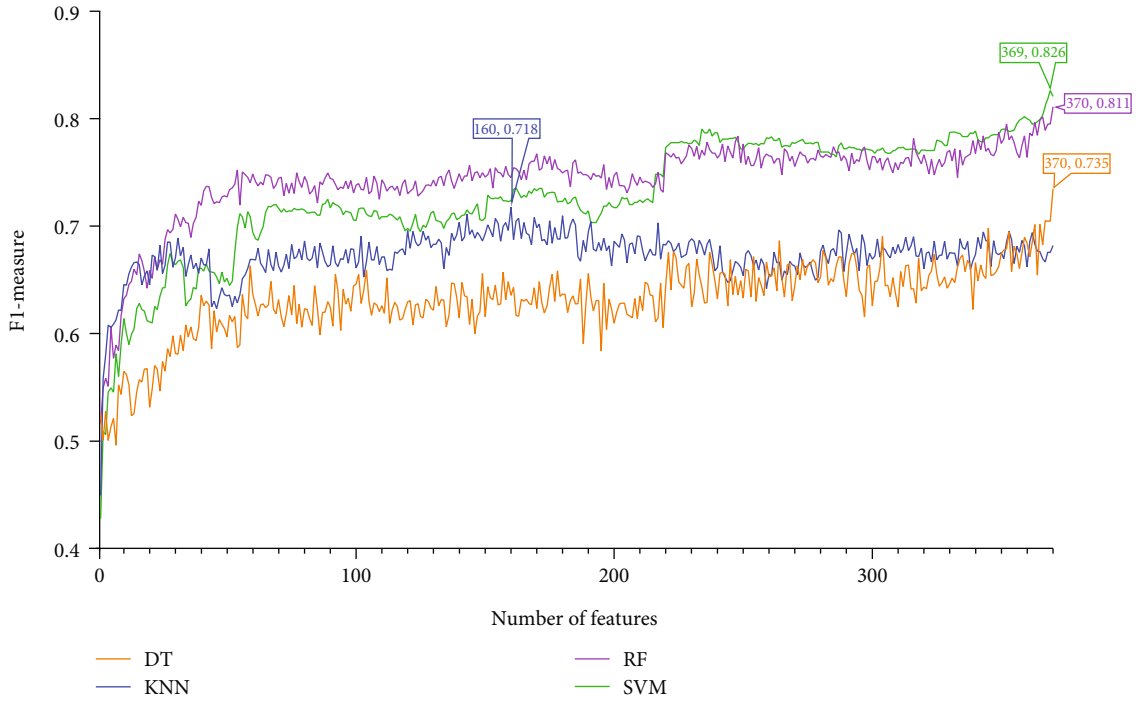

Figure 5: IFS curves for showing the performance of four classification algorithms according to F1-measure on the LASSO feature list.

Table 1: Performance of the optimal models based on various classification algorithms and lists yielded by various feature ranking algorithms.

Feature ranking algorithms	Classification algorithms	Number of features	F1-measure	MCC	ACC
mRMR	DT	309	0.766	0.623	0.822
	KNN	49	0.796	0.665	0.831
	SVM	314	0.852	0.761	0.886
	RF	215	0.840	0.742	0.878
MCFS	DT	64	0.765	0.620	0.820
	KNN	62	0.805	0.679	0.838
	SVM	143	0.846	0.750	0.880
	RF	145	0.854	0.764	0.888
LightGBM	DT	32	0.800	0.676	0.846
	KNN	18	0.833	0.728	0.867
	SVM	33	0.868	0.788	0.899
	RF	22	0.883	0.811	0.911
LASSO	DT	370	0.735	0.568	0.794
	KNN	160	0.718	0.525	0.753
	SVM	369	0.826	0.718	0.866
	RF	370	0.811	0.695	0.856

(expressed by SSPN), ENST00000339223 (expressed by FPR3), ENST00000341184 (expressed by MGAT3), and ENST00000316418 (expressed by AHRR) are supported by previous studies [52–58]. Therefore, the features identified using all four methods have strong associations with smoking.

In comparison with LASSO or mRMR, the two other algorithms LightGBM and MCFS produced relative smaller number of optimal features. However, these two methods shared only half or less optimal features with each other. A further investigation unveiled that those optimal features by MCFS but not LightGBM were either transcripts of immunoglobulins or unclear in the relevance to smoking. This result indicated the bias or capability of MCFS methods in capturing the features involved in immune response. By

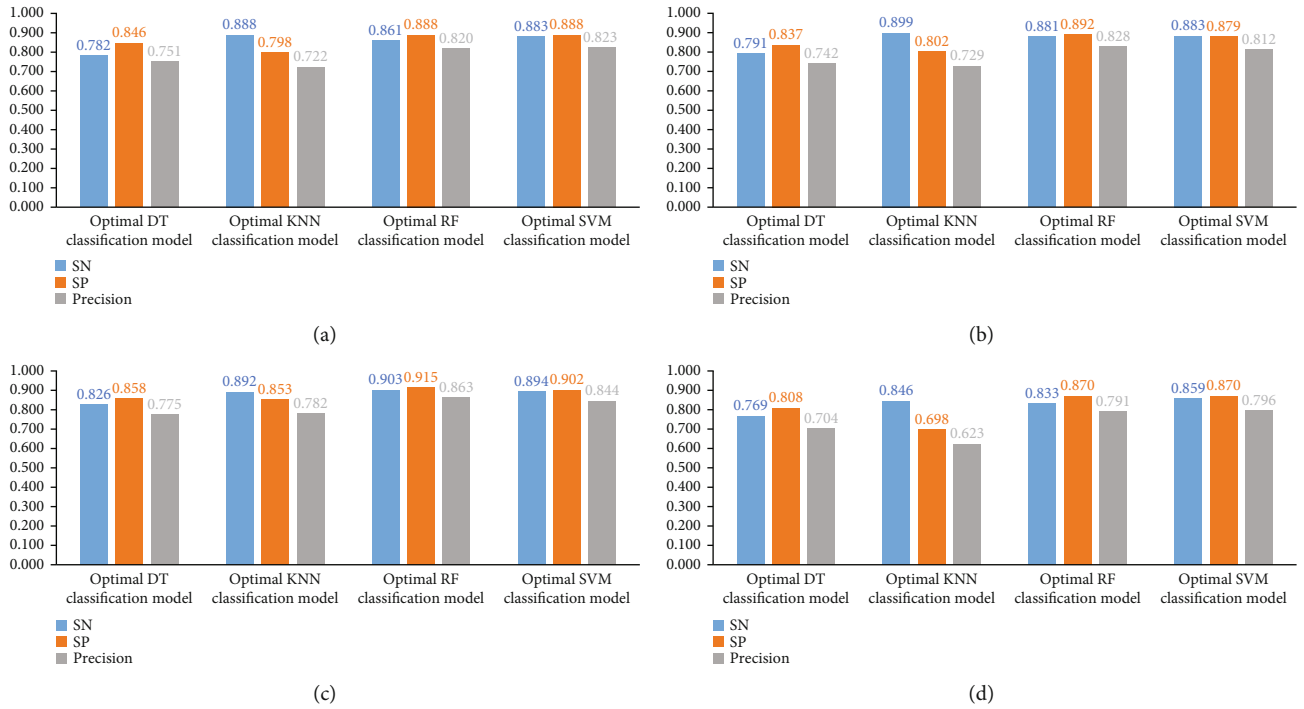

Figure 6: Performance of the optimal models based on different feature lists. (a) mRMR feature list, (b) MCFS feature list, (c) LightGBM feature list, and (d) LASSO feature list.

contrast, those optimal features identified by LightGBM but not MCFS were involved in diverse functions or bioprocesses. For example, ENST00000297785 was a transcript expressed by the gene *ALDH1*, which was upregulated by smoking, while another feature ENST00000423064, a transcript of *HGF*, was found to be upregulated in male smokers [59]. And many risk factors including smoking were found to be associated with serum *HGF* [60]. Therefore, different algorithms identify features with different focus, and a combination of multiple algorithms is recommended for a comprehensive analysis to complement with each other.

4.3. Analysis of Classification Rules. Further analysis on the rules was also conducted to investigate the relevance to smoking. The top-ranked rule in each of the four algorithms received much more passed courts than the second, indicating that the top rule outperformed all the others in terms of precision. Therefore, we focused on the gene expression pattern in the top rule of each algorithm. Among the 38 features in the four top rules, only two features were shared by all four top rules from four algorithms, ENST00000284311 and ENST00000586582, which are the transcripts of *GPR15* and *SEMA6B*. Both features were strongly associated with smoking responses by previous studies [50–52]. The only feature shared by the three top rules was ENST0000 0390539, which is a transcript of immunoglobulin. Immunoglobulin has four other features, indicating that these immunoglobulin isoforms might play a more important role in smoking response than others. Other than the immune response, the epigenetic change is also a key

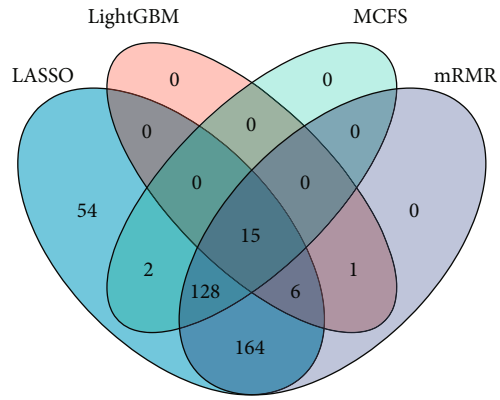

Figure 7: Venn diagram of the optimal feature subsets obtained from the mRMR, MCFS, LightGBM, and LASSO feature lists. The circles indicate transcripts that were identified as optimal features by different ranking algorithms.

player in smoking response. Our study revealed four features from these 38 top-rank rules, namely, ENST0000 0395002 from FAM13A, ENST00000394718 from AKAP5, ENST00000341184 from MGAT3, and ENST00000316418 from AHRR. These genes were associated with smoking response [57, 61–63]. Moreover, 17 features have not been studied, indicating their roles in smoking response. These genes or features could be good candidates and further investigated in future studies to unveil the mechanisms of smoking response.

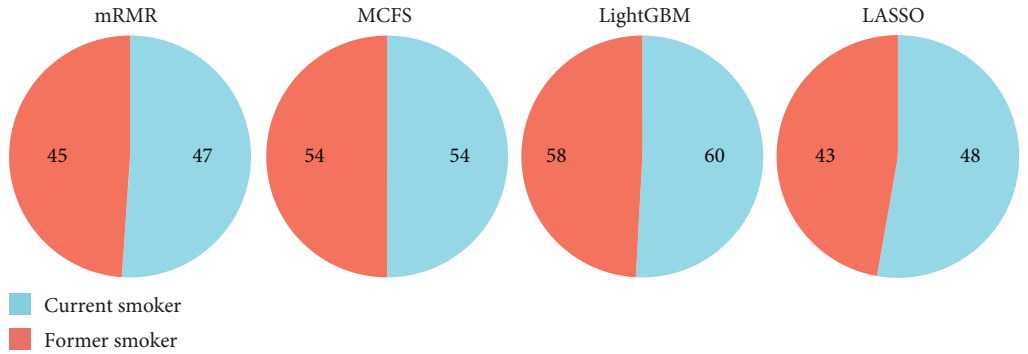

Figure 8: Number of rules utilized to distinguish each class in each classification rule set extracted based on the mRMR, MCFS, LightGBM, and LASSO feature lists. A large number of classification rules were built to describe two kinds of smokers. The number of classification rules for each kind of smokers was approximately the same, indicating that our method was unbiased.

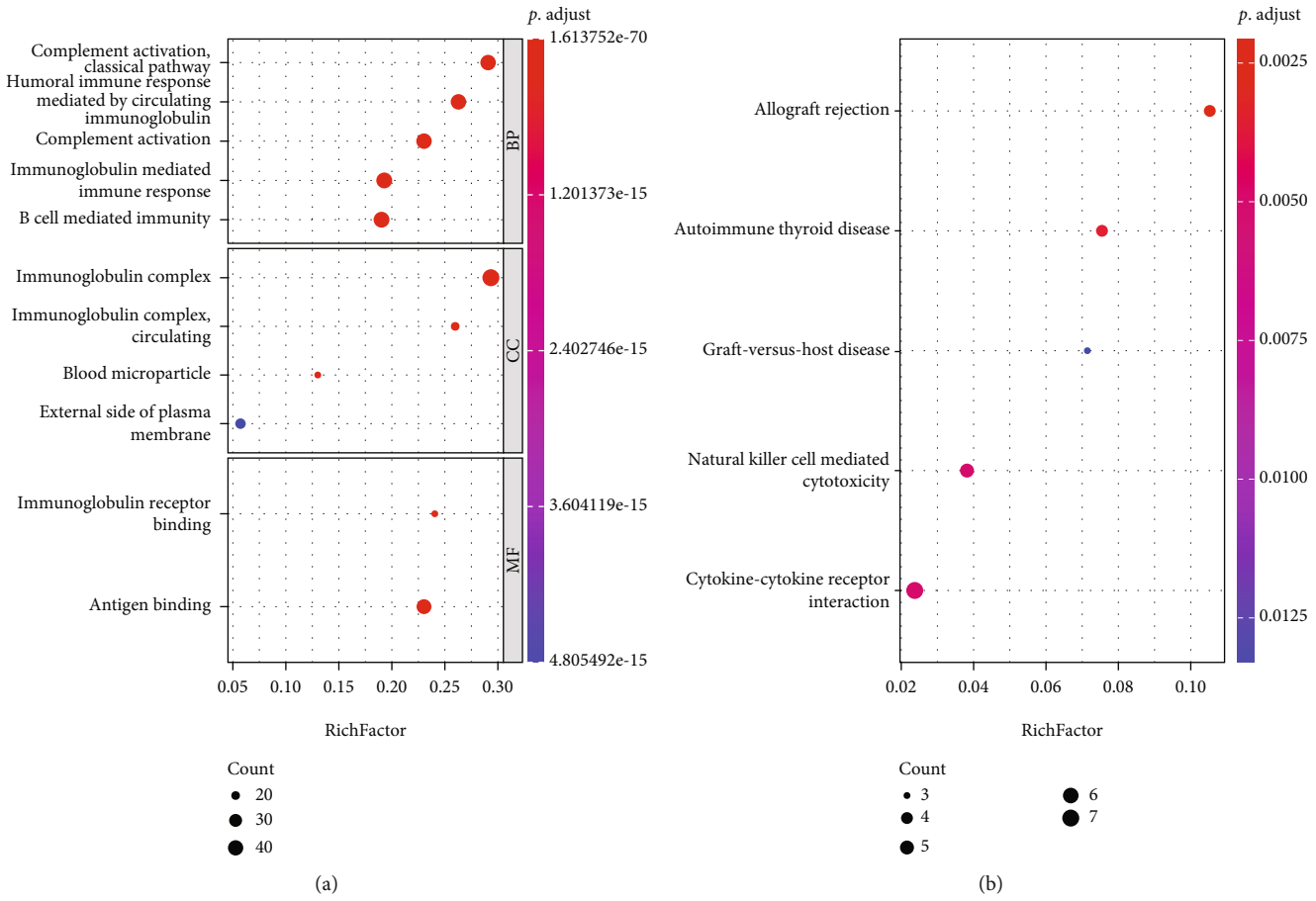

Figure 9: Analysis of KEGG pathway and gene ontology enrichment using the union of four best feature subsets determined by four feature ranking algorithms. The GO terms and KEGG pathways were filtered by the criterion of FDR < 0.05. The top 5 significant GO enrichment results and KEGG pathway enrichment results are shown.

5. Conclusions

In this study, some widely used machine learning methods were applied on transcript expression data to reveal the essential features of different populations with different smoking history. Three aspects of the results were obtained. First, a list of features that could be used to determine the difference between current and former smokers were extracted. These features provided a more detailed description of the alteration of biological processes in the human body by smoking at the transcript level. Second, efficient classification models were built to identify current and former smokers. Finally, specific classification rules for distinguishing current smokers from former smokers were built.

These rules quantitatively described the role of transcript expression in differentiating smoking populations, thus providing a theoretical basis for the treatment of smoking-related diseases.

Data Availability

The original data used to support the findings of this study are available at the GEO database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE171730).

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the National Key R&D Program of China (2022YFF1203202), the Strategic Priority Research Program of Chinese Academy of Sciences (XDA26040304 and XDB38050200), the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002).

Supplementary Materials

Table S1: feature ranking results obtained by mRMR, MCFS, LightGBM, and LASSO methods. Table S2: IFS results on different feature lists. Table S3: intersection of the optimal feature subsets extracted from mRMR, MCFS, LightGBM, and LASSO feature lists. The features that appear in 4, 3, 2, and 1 optimal feature subsets are shown. Table S4: classification rules generated by the optimal DT model. Table S5: GO and KEGG enrichment results after merging the optimal feature subsets of the four feature ranking algorithms. *(Supplementary Materials)*

References

[1] J. M. Samet, "Tobacco smoking: the leading cause of preventable disease worldwide," *Thoracic Surgery Clinics*, vol. 23, no. 2, pp. 103–112, 2013.

[2] A. Chatziioannou, The EnviroGenomarkers project consortium, P. Georgiadis et al., "Blood-based omic profiling supports female susceptibility to tobacco smoke- induced cardiovascular diseases," *Scientific Reports*, vol. 7, no. 1, p. 42870, 2017.

[3] B. A. Forey, A. J. Thornton, and P. N. Lee, "Systematic review with meta-analysis of the epidemiological evidence relating smoking to COPD, chronic bronchitis and emphysema," *BMC Pulmonary Medicine*, vol. 11, no. 1, p. 36, 2011.

[4] K. W. Lee and Z. Pausova, "Cigarette smoking and DNA methylation," *Frontiers in Genetics*, vol. 4, p. 132, 2013.

[5] L. F. Li, R. L. Y. Chan, L. Lu et al., "Cigarette smoking and gastrointestinal diseases: the causal relationship and underlying molecular mechanisms (review)," *International Journal of Molecular Medicine*, vol. 34, no. 2, pp. 372–380, 2014.

[6] C. P. Silva and H. M. Kamens, "Cigarette smoke-induced alterations in blood: a review of research on DNA methylation and gene expression," *Experimental and Clinical Psychopharmacology*, vol. 29, no. 1, pp. 116–135, 2021.

[7] W. Jiang and L. Chen, "Alternative splicing: human disease and quantitative analysis from high- throughput sequencing," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 183–195, 2021.

[8] Z. Xu, J. Platig, S. Lee et al., "Cigarette smoking-associated isoform switching and 3′ UTR lengthening via alternative polyadenylation," *Genomics*, vol. 113, no. 6, pp. 4184–4195, 2021.

[9] D. Bzdok, N. Altman, and M. Krzywinski, "Statistics versus machine learning," *Nature Methods*, vol. 15, no. 4, pp. 233-234, 2018.

[10] M. B. Kursa and W. R. Rudnicki, "Feature selection with the-Borutapackage," *Journal of Statistical Software*, vol. 36, no. 11, pp. 1–13, 2010.

[11] H. A. Liu and R. Setiono, "Incremental feature selection," *Applied Intelligence*, vol. 9, no. 3, pp. 217–230, 1998.

[12] L. Chen, Z. Li, T. Zeng et al., "Predicting gene phenotype by multi-label multi-class model based on essential functional features," *Molecular Genetics and Genomics*, vol. 296, no. 4, pp. 905–918, 2021.

[13] Y.-H. Zhang, H. Li, T. Zeng et al., "Identifying transcriptomic signatures and rules for SARS-CoV-2 infection," *Frontiers in Cell and Development Biology*, vol. 8, article 627302, 2020.

[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[15] M. Draminski, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, and J. Komorowski, "Monte Carlo feature selection for supervised classification," *Bioinformatics*, vol. 24, no. 1, pp. 110–117, 2008.

[16] G. Ke, Q. Meng, T. Finley et al., "LightGBM: a highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems30 (NIP 2017)*, Curran Associates Inc., Red Hook, NY, United States, 2017.

[17] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 73, no. 1, pp. 273–282, 1996.

[18] F. Yuan, L. Lu, Y. H. Zhang, S. P. Wang, and Y. D. Cai, "Data mining of the cancer-related lncRNAs GO terms and KEGG pathways by using mRMR method," *Mathematical Biosciences*, vol. 304, pp. 1–8, 2018.

[19] X. Yu, X. Y. Pan, S. Q. Zhang et al., "Identification of gene signatures and expression patterns during epithelial-to-mesenchymal transition from single-cell expression atlas," *Frontiers in Genetics*, vol. 11, article 605012, 2020.

[20] X. Zhao, L. Chen, and J. Lu, "A similarity-based method for prediction of drug side effects with heterogeneous information," *Mathematical Biosciences*, vol. 306, pp. 136–144, 2018.

[21] L. Chen, Z. D. Li, S. Q. Zhang, Y.-H. Zhang, T. Huang, and Y.-D. Cai, "Predicting RNA 5-methylcytosine sites by using essential sequence features and distributions," *BioMed Research International*, vol. 2022, Article ID 4035462, 11 pages, 2022.

[22] L. Chen, J. R. Li, Y. H. Zhang et al., "Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method," *Journal of Cellular Biochemistry*, vol. 119, no. 4, pp. 3394–3403, 2018.

[23] X. Chen, Y. Jin, and Y. Feng, "Evaluation of plasma extracellular vesicle microRNA signatures for lung adenocarcinoma and granuloma with Monte-Carlo feature selection method," *Frontiers in Genetics*, vol. 10, p. 367, 2019.

[24] J. Li, L. Lu, Y. H. Zhang et al., "Identification of leukemia stem cell expression signatures through Monte Carlo feature selection strategy and support vector machine," *Cancer Gene Therapy*, vol. 27, no. 1-2, pp. 56–69, 2020.

[25] Y. H. Zhang, W. Guo, T. Zeng et al., "Identification of microbiota biomarkers with orthologous gene annotation for type 2 diabetes," *Frontiers in Microbiology*, vol. 12, article 711244, 2021.

[26] Y. H. Zhang, Z. Li, T. Zeng et al., "Distinguishing glioblastoma subtypes by methylation signatures," *Frontiers in Genetics*, vol. 11, article 604336, 2020.

[27] S. Ding, D. Wang, X. Zhou et al., "Predicting heart cell types by using transcriptome profiles and a machine learning method," *Life*, vol. 12, no. 2, p. 228, 2022.

[28] X. Zhou, S. Ding, D. Wang et al., "Identification of cell markers and their expression patterns in skin based on single-cell RNA-sequencing profiles," *Life*, vol. 12, no. 4, p. 550, 2022.

[29] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Joint Conference on Artificial Intelligence*, Lawrence Erlbaum Associates Ltd., 1995.

[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[31] X. Pan, L. Chen, M. Liu, Z. Niu, T. Huang, and Y. D. Cai, "Identifying protein subcellular locations with embeddings-based node2loc," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 2, pp. 666–675, 2022.

[32] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[33] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[34] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[35] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[36] L. Chen, Z. Li, T. Zeng et al., "Identifying COVID-19-specific transcriptomic biomarkers with machine learning methods," *BioMed Research International*, vol. 2021, Article ID 9939134, 11 pages, 2021.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "scikit-learn: machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[38] D. Powers, "Evaluation: from precision, recall and f-measure to ROC., informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[39] Y. Yang and L. Chen, "Identification of drug–disease associations by using multiple drug anddisease networks," *Current Bioinformatics*, vol. 17, no. 1, pp. 48–59, 2022.

[40] B. Ran, L. Chen, M. Li, Y. Han, and Q. Dai, "Drug-drug interactions prediction using fingerprint only," *Computational and Mathematical Methods in Medicine*, vol. 2022, Article ID 7818480, 14 pages, 2022.

[41] T. Wu, E. Hu, S. Xu et al., "clusterProfiler 4.0: a universal enrichment tool for interpreting omics data," *The Innovations*, vol. 2, no. 3, article 100141, 2021.

[42] S. T. Lugg, A. Scott, D. Parekh, B. Naidu, and D. R. Thickett, "Cigarette smoke exposure and alveolar macrophages: mechanisms for lung disease," *Thorax*, vol. 77, no. 1, pp. 94–101, 2022.

[43] D. O'Shea, T. J. Cawood, C. O'Farrelly, and L. Lynch, "Natural killer cells in obesity: impaired function and increased susceptibility to the effects of cigarette smoke," *PLoS One*, vol. 5, no. 1, article e8660, 2010.

[44] Y. Arnson, Y. Shoenfeld, and H. Amital, "Effects of tobacco smoke on immunity, inflammation and autoimmunity," *Journal of Autoimmunity*, vol. 34, no. 3, pp. J258–J265, 2010.

[45] S. Bae, Q. Ke, Y. Y. Koh et al., "Exacerbation of acute viral myocarditis by tobacco smoke is associated with increased viral load and cardiac apoptosis," *Canadian Journal of Physiology and Pharmacology*, vol. 88, no. 5, pp. 568–575, 2010.

[46] W. M. Wiersinga, "Clinical relevance of environmental factors in the pathogenesis of autoimmune thyroid disease," *Endocrinology and Metabolism*, vol. 31, no. 2, pp. 213–222, 2016.

[47] J. G. Mabley, P. Pacher, G. J. Southan, A. L. Salzman, and C. Szabó, "Nicotine reduces the incidence of type I diabetes in mice," *The Journal of Pharmacology and Experimental Therapeutics*, vol. 300, no. 3, pp. 876–881, 2002.

[48] P. Beineke, K. Fitch, H. Tao et al., "A whole blood gene expression-based signature for smoking status," *BMC Medical Genomics*, vol. 5, no. 1, p. 58, 2012.

[49] D. R. Lorenz, V. Misra, and D. Gabuzda, "Transcriptomic analysis of monocytes from HIV-positive men on antiretroviral therapy reveals effects of tobacco smoking on interferon and stress response systems associated with depressive symptoms," *Human Genomics*, vol. 13, no. 1, p. 59, 2019.

[50] A. M. Andersen, M. K. Lei, S. R. H. Beach, and R. A. Philibert, "Inflammatory biomarker relationships with helper T cell GPR15 expression and cannabis and tobacco smoking," *Journal of Psychosomatic Research*, vol. 141, article 110326, 2021.

[51] J. M. Vink, R. Jansen, A. Brooks et al., "Differential gene expression patterns between smokers and non-smokers: cause or consequence?," *Addiction Biology*, vol. 22, no. 2, pp. 550–560, 2017.

[52] B. V. J. Cuppen, M. Rossato, R. D. E. Fritsch-Stork et al., "RNA sequencing to predict response to TNF-α inhibitors reveals possible mechanism for nonresponse in smokers," *Expert Review of Clinical Immunology*, vol. 14, no. 7, pp. 623–633, 2018.

[53] F. Martin, M. Talikka, J. Hoeng, and M. C. Peitsch, "Identification of gene expression signature for cigarette smoke exposure response—from man to mouse," *Human & Experimental Toxicology*, vol. 34, no. 12, pp. 1200–1211, 2015.

[54] L. M. Reynolds, K. Lohman, G. S. Pittman et al., "Tobacco exposure-related alterations in DNA methylation and gene expression in human monocytes: the multi-ethnic study of atherosclerosis (MESA)," *Epigenetics*, vol. 12, no. 12, pp. 1092–1100, 2017.

[55] M. Graff, L. Fernández-Rhodes, S. Liu et al., "Generalization of adiposity genetic loci to US Hispanic women," *Nutrition & Diabetes*, vol. 3, no. 8, article e85, 2013.

[56] S. D. Pouwels, V. R. Wiersma, I. E. Fokkema et al., "Acute cigarette smoke-induced eQTL affects formyl peptide receptor

expression and lung function," *Respirology*, vol. 26, no. 3, pp. 233–240, 2021.

[57] V. Barcelona, Y. Huang, K. Brown et al., "Novel DNA methylation sites associated with cigarette smoking among African Americans," *Epigenetics*, vol. 14, no. 4, pp. 383–391, 2019.

[58] H. Ohmomo, S. Harada, S. Komaki et al., "DNA methylation abnormalities and altered whole transcriptome profiles after switching from combustible tobacco smoking to heated tobacco products," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 31, no. 1, pp. 269–279, 2022.

[59] M. Patel, L. Lu, D. S. Zander, L. Sreerama, D. Coco, and J. S. Moreb, "ALDH1A1 and ALDH3A1 expression in lung cancers: correlation with histologic type and potential precursors," *Lung Cancer*, vol. 59, no. 3, pp. 340–349, 2008.

[60] L. Torres, E. Klingberg, M. Nurkkala, H. Carlsten, and H. Forsblad-d'Elia, "Hepatocyte growth factor is a potential biomarker for osteoproliferation and osteoporosis in ankylosing spondylitis," *Osteoporosis International*, vol. 30, no. 2, pp. 441–449, 2019.

[61] Q. Chen, M. de Vries, K. O. Nwozor et al., "A protective role of FAM13A in human airway epithelial cells upon exposure to cigarette smoke extract," *Frontiers in Physiology*, vol. 12, article 690936, 2021.

[62] A. Oldenburger, W. J. Poppinga, F. Kos et al., "A-kinase anchoring proteins contribute to loss of E-cadherin and bronchial epithelial barrier by cigarette smoke," *American Journal of Physiology. Cell Physiology*, vol. 306, no. 6, pp. C585–C597, 2014.

[63] F. Takeuchi, K. Takano, M. Yamamoto et al., "Clinical implication of smoking-related aryl-hydrocarbon receptor repressor (AHRR) hypomethylation in Japanese adults," *Circulation Journal*, vol. 86, no. 6, pp. 986–992, 2022.