frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 20 October 2022
DOI 10.3389/fnins.2022.1034971

OPEN ACCESS

EDITED BY
Shuo Wang,
Beijing Tiantan Hospital, Capital Medical University, China

REVIEWED BY
Zhe Xu,
National Clinical Research Center for Neurological Diseases, China
Ming Xiao,
Sichuan University, China

*CORRESPONDENCE
Mingquan Ye
ymq@wnmc.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Translational Neuroscience, a section of the journal Frontiers in Neuroscience

RECEIVED 02 September 2022
ACCEPTED 30 September 2022
PUBLISHED 20 October 2022

CITATION
Li Q, Wang P, Yuan J, Zhou Y, Mei Y and Ye M (2022) A two-stage hybrid gene selection algorithm combined with machine learning models to predict the rupture status in intracranial aneurysms. *Front. Neurosci.* 16:1034971. doi: 10.3389/fnins.2022.1034971

A two-stage hybrid gene selection algorithm combined with machine learning models to predict the rupture status in intracranial aneurysms

Qingqing Li[1,2†], Peipei Wang[1,2†], Jinlong Yuan[3], Yunfeng Zhou[4], Yaxin Mei[1,2] and Mingquan Ye[1,2*]

[1]School of Medical Information, Wannan Medical College, Wuhu, Anhui, China, [2]Research Center of Health Big Data Mining and Applications, Wannan Medical College, Wuhu, Anhui, China, [3]Department of Neurosurgery, Yijishan Hospital of Wannan Medical College, Wannan Medical College, Wuhu, Anhui, China, [4]Department of Radiology, Yijishan Hospital of Wannan Medical College, Wannan Medical College, Wuhu, Anhui, China

An IA is an abnormal swelling of cerebral vessels, and a subset of these IAs can rupture causing aneurysmal subarachnoid hemorrhage (aSAH), often resulting in death or severe disability. Few studies have used an appropriate method of feature selection combined with machine learning by analyzing transcriptomic sequencing data to identify new molecular biomarkers. Following gene ontology (GO) and enrichment analysis, we found that the distinct status of IAs could lead to differential innate immune responses using all 913 differentially expressed genes, and considering that there are numerous irrelevant and redundant genes, we propose a mixed filter- and wrapper-based feature selection. First, we used the Fast Correlation-Based Filter (FCBF) algorithm to filter a large number of irrelevant and redundant genes in the raw dataset, and then used the wrapper feature selection method based on the he Multi-layer Perceptron (MLP) neural network and the Particle Swarm Optimization (PSO), accuracy (ACC) and mean square error (MSE) were then used as the evaluation criteria. Finally, we constructed a novel 10-gene signature (YIPF1, RAB32, WDR62, ANPEP, LRRCC1, AADAC, GZMK, WBP2NL, PBX1, and TOR1B) by the proposed two-stage hybrid algorithm FCBF-MLP-PSO and used different machine learning models to predict the rupture status in IAs. The highest ACC value increased from 0.817 to 0.919 (12.5% increase), the highest area under ROC curve (AUC) value increased from 0.87 to 0.94 (8.0% increase), and all evaluation metrics improved by approximately 10% after being processed by our proposed gene selection algorithm. Therefore,

these 10 informative genes used to predict rupture status of IAs can be used as complements to imaging examinations in the clinic, meanwhile, this selected gene signature also provides new targets and approaches for the treatment of ruptured IAs.

KEYWORDS

intracranial aneurysms, rupture status, gene selection, machine learning, FCBF-MLP-PSO, informative genes

Introduction

An intracranial aneurysm (IA) is an abnormal swelling of cerebral vessels, which can occur without causing any symptoms (Pontes et al., 2021). A subset of these IAs can rupture, causing aneurysmal subarachnoid hemorrhage (aSAH), often resulting in death or severe disability (Tawk et al., 2021). In general, assessing the number of subarachnoid hemorrhages requires imaging studies such as *trans*-cranial Doppler, computed tomography (CT), and magnetic resonance imaging (MRI) which are sometimes difficult to obtain especially in complicated patients and are technically demanding for physicians (Rose, 2011). With the rapid growth of RNA-sequencing (RNA-seq) technologies, massive sequencing data have been produced in the area of tumor research. In the field of IA research, evaluating the status of IAs by transcriptomic profiling has also become a hotspot, with several studies focusing on mining the molecular biomarkers from transcriptomic data to predict the status of an IA (Gao et al., 2020; Poppenberg et al., 2020). To date, few studies have used the method of feature selection combined with machine learning in this area; however, in the early stages of subarachnoid hemorrhage, using the appropriate method to distinguish quantity controlled molecular markers can help to precisely predict the status of an IA and may provide new therapeutic targets.

Intracranial aneurysm transcriptomic sequencing data, similar to gene expression data from other tumors, often have the properties of a small sample size and high dimensional features, large amounts of redundant or unnecessary features may not only result in misdiagnosis and failure to diagnose but can also be time-consuming and reduce the effectiveness of the categorization (Chen et al., 2016; Xiong et al., 2021). To better acquire effective information from gene expression data, currently, our objective is to successfully reduce the feature dimensionality and obtain an informative subset of genes with the best categorization performance. With traditional statistical methods, multiple genes from the same pathway are selected, as genes in the same pathway tend to have the same or similar expression pattern, which can lead to the introduction of a particular set of gene signatures involved in one particular biological process by overrepresentation and thus introduce considerable redundancy (Perscheid, 2021). Machine learning models have unique advantages in addressing issues such as clustering, classification, and regression of high-dimensional biological multi-omics data (Camacho et al., 2018), feature selection is a key step of machine learning in the preprocessing of gene expression sequencing data, which is beneficial to precision medicine, can help discover disease mechanisms and reduce the cost of clinical diagnosis by finding the optimal set of features based on the performance of classification models (Chen et al., 2021).

When managing RNA-seq data and gene microarray data, the feature selection process that before machine learning modeling was commonly referred to as informative gene selection. The purpose of gene selection is to eliminate completely unrelated and noise features, weak correlation and redundancy features, and to filter out strong correlation features related to modeling. The optimal subset of features obtained by feature selection should theoretically make the model run faster, with higher model performance, and unlike feature extraction, the value of the eigenvalues in the data does not change after feature selection. Depending on the way the subset of features is evaluated, gene feature selection methods can be classified as filter-based methods, wrapper-based methods and embedded-based methods, as well as hybrid-based methods and ensemble-based methods, which have been popular in recent years (Jain et al., 2018; Ye et al., 2019). Without taking into account any particular learning algorithm, filter methods eliminate genes with minimal information based on the statistical properties of variables, and mainly include correlation-based feature selection, the Markov blanket filter method, and the mutual information-based methods (Yu and Liu, 2003; Tang et al., 2018). Among them, the Fast Correlation-Based Filter (FCBF) is a filtering solution based on the two features through the Symmetrical Uncertainty (SU) method as a measure, which have been widely used in high-dimensional gene expression profiles (Lei and Liu, 2003). Wrapper methods can obtain a relatively small subset of genes with better classification through performance by evaluating the performance of the predefined learning algorithm (Huang et al., 2020). Informative gene selection and the training procedure are conducted

concurrently using embedded methods, which incorporate gene selection into the learning algorithm (Medjahed et al., 2017).

In this study, we propose a two-stage hybrid algorithm FCBF-Multi-layer Perceptron (MLP)-Particle Swarm Optimization (PSO), that is, a mixed filter- and wrapper-based feature selection. Firstly, using the FCBF algorithm to filter a large number of irrelevant and redundant genes in the raw dataset, and then using the wrapper feature selection method based on the MLP neural network and PSO, MLP as a classifier, the PSO algorithm as the search strategy, and using accuracy (accuracy) and mean square error (MSE) as the evaluation criteria, the feature set with fewer variable numbers and high classification ACC is finally obtained as the optimal feature subset. Our findings demonstrated that the proposed method can improve classification performance while also obtaining a smaller subset of informative genes, which will benefit the mining of IA rupture-related biomarkers.

Materials and methods

Data collection

The gene expression omnibus (GEO) datasets[1] are maintained by NCBI to store gene expression profiles by RT-PCR, high-throughput sequencing, microarray and so on. We downloaded four gene expression datasets from the GEO, GSE13353 series on the GPL570 platform (Affymetrix Human Genome U133 Plus 2.0 Array) (Kurki et al., 2011), GSE15629 on the GPL6244 platform (Affymetrix Human Gene 1.0 ST Array) (Pera et al., 2010), GSE54083 on the GPL4133 platform (Agilent-014850 Whole Human Genome Microarray 4x44K G4112F) (Nakaoka et al., 2014), and GSE122897 on the GPL16791 platform (Illumina HiSeq 2500) (Kleinloog et al., 2016), since when we searched the GEO database and found that only these four GSE gene expression datasets meet our requirements that contained clear information on whether IAs ruptured or not. Data from different platforms were normalized and centered before the "sva" R package was used to remove the batch effects. After removing negative controls that were not IA, samples with explicit status of ruptured and unruptured aneurysms were retained. Finally, 88 samples entered subsequent analysis, including 48 ruptured and 40 unruptured samples of IA. The ruptured and unruptured groups were labeled 1 and 0, respectively. 0 and 1 were also used as target labels for binary samples by the following feature selection and machine learning algorithms.

Recognition and enrichment analysis of differentially expressed genes

After data had been downloaded and integrated, R package "limma" (Ritchie et al., 2015) suitable for both RNA-seq and microarray studies was used to generate DEGs between the two groups based on linear models. P-value < 0.05 and the $|$ Fold-change$| > 1.5$ were the requirements for DEG significance. R package "pheatmap" was used to display the heatmap plot and visualize the results of differential expression analysis. Then, non-redundant gene biological terms in a functionally grouped network were clustered and visualized using the Cytoscape (version 3.8.0) desktop application and the "ClueGO" plug-in. The Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichments of up-regulated and down-regulated DEGs were separately performed by web-based Metascape (Zhou et al., 2019).

Evaluation of immune cell infiltration

TIMER 2.0[2] was used to provide a robust estimation of immune infiltration levels in each sample of IA (Li et al., 2020), algorithms including TIMER, CIBERSORT, quanTIseq, xCell, MCP-counter and EPIC were all implemented. The significance criterion was P-value < 0.05 by two-tailed t-test, the expression profiles of immune cells with significant differences between groups in each sample are shown by heatmaps, and the overall profiles of each group are shown by box plots.

The proposed two-stage hybrid feature selection algorithm

Related theory

(1) The FCBF is a filtering solution based on fast correlations, as proposed by Lei and Liu (2003). The core idea of the algorithm is to measure the correlation of the two features through the SU method as a measure, the SU value of each feature g_i vs. the category C is calculated as $SU(X, Y) = 2\left[\frac{H(X)-H(X|Y)}{H(X)+H(Y)}\right]$, namely (g_i, C), where $H(X)$ represents the information entropy, and $H(X|Y)$ represents conditional entropy. The filter-based feature selection method based on FCBF is suitable for solving the feature selection problem of large-scale data, and can effectively delete redundant and irrelevant features in high-dimensional data. However, because the filter-based feature selection method is separated

1 http://www.ncbi.nlm.nih.gov/geo

2 http://timer.cistrome.org/

from the evaluation strategy of the learning algorithm in the process of gene subset selection, it is difficult to determine whether the selected feature subset can make the classification algorithm achieve the best performance.

(2) The MLP is an artificial neural network that tends to structure. This algorithm maps a set of input vectors to a set of output vectors by including a network structure with an input layer, hidden layer and output layer (Yang et al., 2009). The output of the MLP neural network can be expressed as $f_k(d) = \frac{1}{1+exp[-g_k(d)]}$, where $g_k(d)$ is the weighted sum of the hidden layer nodes, $k = 1, 2, \cdots, m$; d is the eigenvalue vector of the MLP neural network, with $d = (d_1, d_2, \cdots, d_i)$. The MLP based wrapped feature selection method selects feature subsets that are more relevant to the classification algorithm than the filtered feature selection algorithm, which is not conducive to solving the difficulties of feature selection and classification caused by the complex sample and feature distribution characteristics of the bionomic data, and has low computational efficiency in high-dimensional data processing.

(3) The PSO is a group intelligent search algorithm that simulates birds feeding on food in nature (Rostami et al., 2020). The PSO algorithm places a population of particles in the D-dimensional search space and evaluates the fitness of each particle. In the PSO algorithm, the i-th particle can update its next-generation position and flight speed with the following two formulations:

$$x_i^{t+1} = x_i(t) + v_i^{t+1}$$

$$v_i^{t+1} = \omega^* v_i^t + c_1{}^* rand_1^* \left(pbest_i - x_i^t\right) + c_2{}^* rand_2^* \left(gbest_i - x_i^t\right)$$

where t represents the current number of iterations, the x_i^{t+1} and v_i^{t+1} indicate the position and flight speed of the particle under the iterations t+1, respectively;ω represents the inertia weight used to adjust the effect of the particle velocity of the previous generation on the current particle velocity; the factors c_1 and c_2 indicate the acceleration coefficient, representing the cognitive learning factors and the social learning factors, respectively, and they are used to adjust the contribution of the individual optimal position *pbest* and the global optimal position *gbest* to the particles, which is usually set to 2; the $rand_1$ and $rand_2$ are represented as random numbers in the [0,1] range.

The main purpose of the two-stage hybrid feature selection method (FCBF-MLP-PSO) is to overcome the shortcomings of the existing filter-based or wrapper-based gene feature selection methods. First, the FCBF filtering feature selection method is used to quickly remove redundant features and generate candidate feature subsets, which can significantly reduce the computational complexity of feature selection for high-dimensional data. Then, the wrapped feature selection method based on MLP is adopted, and an improved particle swarm search strategy is introduced for secondary feature selection. The combined feature subset with strong discrimination ability is selected to overcome the problems of combined features being deleted by mistake and the deviation between the feature evaluation results and the final classification algorithm, thus significantly improving the classification ACC of gene expression in related diseases.

Steps of the proposed hybrid gene selection algorithm

This study proposed a mixed filter- and wrapper-based feature selection algorithm FCBF-MLP-PSO, the proposed hybrid algorithm includes the following steps:

Step 1: The FCBF algorithm was used to filter a large number of irrelevant and redundant genes in the original dataset through SU.

Step 2: The wrapper feature selection method based on the MLP neural network and PSO was used, MLP as a classifier, and PSO algorithm as the search strategy, using ACC and MSE as the evaluation criteria, to finally obtain the feature set with fewer variable numbers and high classification ACC as the optimal feature subset.

Step 3: Several classification algorithms were used to evaluate the effectiveness of gene subsets, including eXtreme Gradient Boosting (XGBoost), LightGBM, Random Forest (RF), Extra Tree (ET), Gaussian NB, K-nearest neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), and Linear Discriminant Analysis (LDA). Each algorithm used the comprehensive evaluation criteria of classification ACC, recall, F1 value, area under ROC curve (AUC) and confusion matrix by 10-fold cross-validation.

The WEKA 3.9.6 software platform and Python 3.8 were used to complete the aforementioned experiments. The experimental flow of the gene selection algorithm suggested in this research is shown in **Figure 1**. The method can effectively decrease the size of the raw gene sets, obtain fewer genes, and have higher classification ACC.

Results

Data description

After removing negative controls that were not IA and samples with vague status of ruptured and unruptured aneurysms, 88 samples subjected to debatching and normalization processing from four GEO datasets including 48 ruptured and 40 unruptured samples of IA were used in this study, and the details of each dataset are illustrated in **Table 1**. The sample size was not large, but the number of features was large; therefore, we need to solve the binary

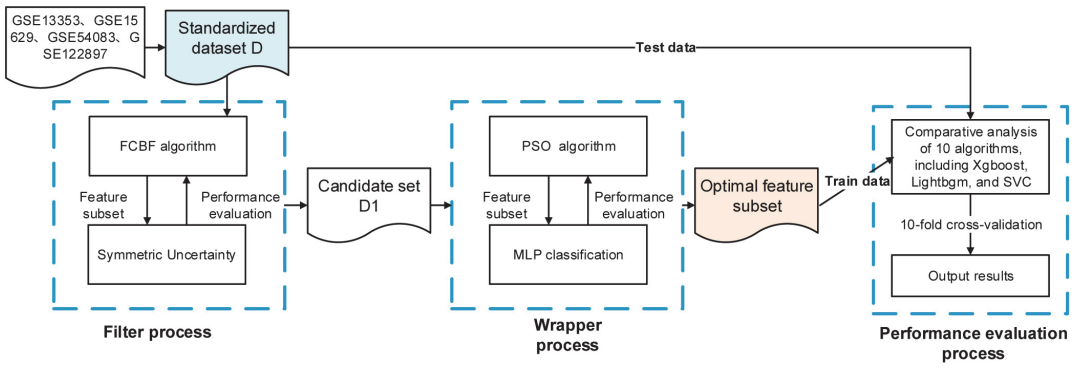

FIGURE 1
The experimental flow of the proposed hybrid gene selection algorithm FCBF-MLP-PSO.

classification problem of high-dimensional small samples. In order to control the quality of batch effect correction, we used two dimensionality reduction methods, including Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), to visualize the clustering of all samples from 4 different GSE datasets before and after adjustment of the batch effects (**Supplementary Figure 1**). **Supplementary Figures 1A,B** demonstrated that samples from different GSE datasets clearly clustered together before batch effect correction. While **Supplementary Figures 1C,D** showed that after de batch effect adjustment, the batch effect was eliminated and all samples from 4 different GSE datasets were almost evenly dispersed.

Differentially expressed mRNAs between ruptured and unruptured intracranial aneurysms

Gene expression data have many redundant features. To improve classification efficiency and to identify distinctive subgroup-specific patterns with different status of ruptured and unruptured aneurysms, we first conducted an analysis of DEGs. A total of 913 mRNAs passed the threshold screening, including 394 up-regulated genes and 519 down-regulated genes. The heatmap selects the first 25 genes exhibiting the most significant fold change for presentation (**Figure 2A**), and from the results of hierarchical clustering, the ruptured and unruptured aneurysms can still be separated into two distinct categories even if the data were derived from different batches and different platforms, indicating that the ruptured aneurysms can affect the gene expression patterns of the aneurysms and that our data preprocessing section is feasible. GO analyses using these 913 genes were further conducted and showed that these differentially expressed mRNAs were predominantly enriched in biological processes related to the immune system (**Figure 2B**). When enrichment analysis was performed separately for up-regulated and down-regulated genes (ruptured vs. unruptured IA), we found that the majority of the GO terms were contributed by down-regulated genes, the most significant of which were neutrophil degranulation and inflammatory response (**Figure 2C**). These results illustrate that ruptured and unruptured aneurysms are clearly distinguished at the transcriptional level, and these distinctions are closely related to the function of the immune system, with the ruptured aneurysm group exhibiting a marked down-regulation of immune-related genes. The deletion of two sets of genes with similar expression patterns, which reduces the number of features from tens of thousands to 913, can also effectively improve the efficiency of subsequent feature selection.

TABLE 1 Description of GEO datasets used in this research.

GEO number	Platform	No. of total	No. of ruptured	No. of unruptured	PMID
GSE13353	GPL570	19	11	8	21336216
GSE15629	GPL6244	14	8	6	20044533
GSE54083	GPL4133	13	8	5	24938844
GSE122897	GPL16791	42	21	21	27026628
	Total	**88**	**48**	**40**	

Bold values represent the average performance.

Differential immunological characteristics between ruptured and unruptured intracranial aneurysms

Considering that the GO terms enriched in neutrophil degranulation and inflammatory response were the most significant in the above results, we speculate that key factors in aneurysm rupture are associated with innate immunity. Multiple algorithms including TIMER, CIBERSORT, quanTIseq, xCell, MCP-counter and EPIC were used to estimate

the abundance of tissue-infiltrating immune subpopulations in ruptured aneurysms and unruptured aneurysms, and the differential immunological characteristics between the two groups are shown in Figure 3. We found that most innate immune cells (such as macrophages, monocytes, neutrophils) were highly infiltrated in ruptured IAs, and adaptive immune cells (such as $CD4^+$ T cells, $CD8^+$ T cells, and B cells) were highly infiltrated in unruptured IAs. These results showed that the distinct status of IAs could lead to differential immune cell composition, thereby influencing the immune responses.

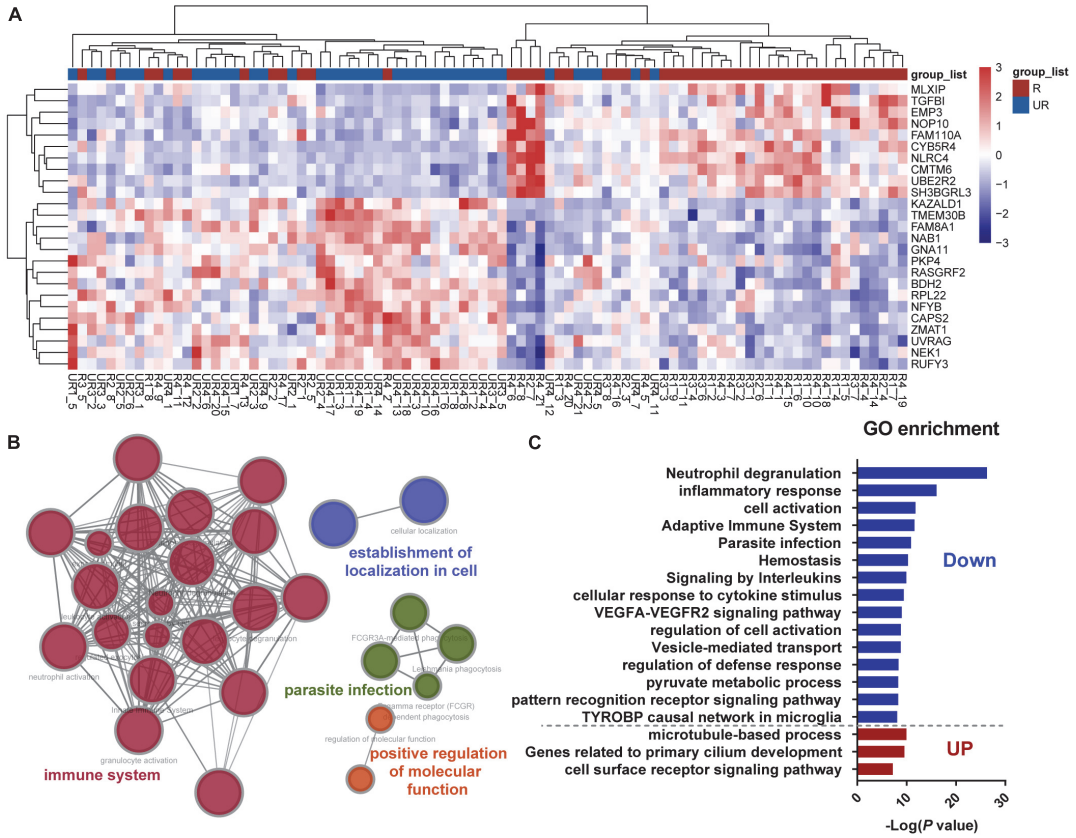

FIGURE 2
Differentially expressed mRNAs between ruptured and unruptured intracranial aneurysms. **(A)** Heatmap plot presents the first 25 genes exhibiting the most significant fold change across all samples, and the ruptured and unruptured aneurysm samples are hierarchically clustered. **(B)** Gene Ontology (GO) enrichment of the differentially expressed mRNAs between the two groups. **(C)** Pathway enrichments of up-regulated and down-regulated DEGs were analyzed separately.

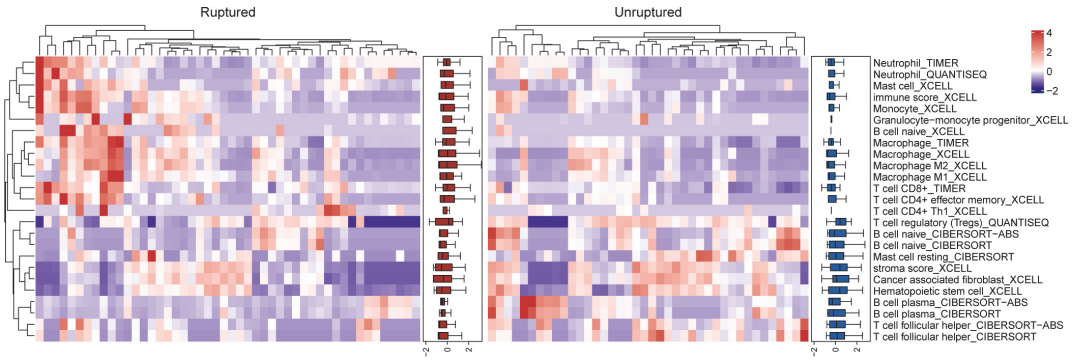

FIGURE 3
Differential immunological characteristics between ruptured and unruptured intracranial aneurysms.

Machine learning-based prediction of rupture status in intracranial aneurysms

In order to accurately predict whether an IA is ruptured using the integrated gene expression data, we attempted to develop an effective classification model that would distinguish ruptured cases from unruptured cases. Based on the 913 DEGs screened out above as features or modeling, 10 types of machine learning classification algorithms, including XGBoost, LightGBM, RF, ET, Gaussian NB, KNN, LR, DT, SVM, and LDA were used to establish the classification model. The evaluation indicators including ACC, recall, precision, F1 value, and mean AUC by 10-fold cross-validation were calculated and are listed in Table 2. The average ACC of all 10 models was 0.752, average recall was 0.732, average precision was 0.812, average F1 score was 0.753 and the average AUC of all models was 0.808. In addition, the ROC curve with 10-fold cross-validation (Figures 4A–J) and the confusion matrix (Figure 4K) for the LR model with the highest ACC and AUC (ACC = 0.817, AUC = 0.87) was plotted. As shown, the performance of these classifiers was reasonable at this point.

Key gene identification for better prediction of rupture status in intracranial aneurysms by the proposed hybrid gene selection algorithm

While the model prediction performance described above is acceptable, among the 913 genes, there are still many redundant and irrelevant features, which may lead to over-fitting of classification models, and still a large number of genes will be difficult to apply in clinical practice. In order to improve the generalization ability of classifiers and the feasibility of clinical application, we proposed a mixed filter- and wrapper-based feature selection algorithm FCBF-MLP-PSO. The FCBF algorithm was first used, and irrelevant and redundant genes were eliminated to obtain a set of 42 candidate features ranked by their importance to category (Supplementary Figure 2). Then genes were screened by MLP-PSO, and a valid subset of 10 features was selected with better ACC and MSE evaluation criteria from the set of candidate features. Finally, YIPF1, RAB32, WDR62, ANPEP, LRRCC1, AADAC, GZMK, WBP2NL, PBX1, and TOR1B containing a total of 10 effective feature subsets were determined.

Next, the selected 10 features were fed into the 10 types of different machine learning classification predictive models. The results showed that when using the hybrid gene selection algorithm proposed above, all evaluation indicators including ACC, recall, precision, F1 value, and mean AUC by 10-fold cross-validation, were improved (Table 3). For example, the average ACC value increased from 0.752 to 0.840 (11.7% increase), and the average AUC value increased from 0.808 to 0.901 (11.5% increase). The highest ACC value increased from 0.817 to 0.919 (12.5% increase), the highest AUC value increased from 0.87 to 0.94 (8.0% increase). In addition, the ROC curve with 10-fold cross-validation (Figures 5A–J) and the confusion matrix (Figure 5K) for SVM classifier with the highest ACC was plotted. The confusion matrix also showed that the predicted label matched the true label better, as shown by the larger number of upper left and lower right diagonals, and cumulatively there were only seven cases of sample prediction error. These results indicated that the deletion of redundant and irrelevant features could improve the model's ACC, whereas reduced features have greater implications for clinical diagnosis.

Distribution of the selected 10 informative genes across samples between the two groups

To explore the relationship between the 10 informative genes selected by our proposed hybrid gene selection algorithm

TABLE 2 Classification performance tests before gene selection.

Model name	Accuracy	Recall	Precision	F1	AUC
XGBClassifier	0.772	0.730	0.856	0.763	0.82 ± 0.14
LGBMClassifier	0.714	0.750	0.763	0.732	0.81 ± 0.11
RandomForestClassifier	0.763	0.750	0.802	0.749	0.81 ± 0.10
ExtraTreesClassifier	0.758	0.765	0.787	0.758	0.83 ± 0.08
GaussianNB	0.817	0.745	0.885	0.804	0.84 ± 0.11
KNeighborsClassifier	0.713	0.680	0.767	0.713	0.76 ± 0.12
LogisticRegression	0.817	0.815	0.858	0.826	0.87 ± 0.11
DecisionTreeClassifier	0.639	0.675	0.743	0.715	0.66 ± 0.20
SupportVectorMachineClassifier	0.794	0.725	0.895	0.777	0.84 ± 0.07
LinearDiscriminantAnalysis	0.736	0.680	0.752	0.691	0.84 ± 0.11
Mean	**0.752**	**0.732**	**0.812**	**0.753**	**0.808**

Bold values represent the average performance.

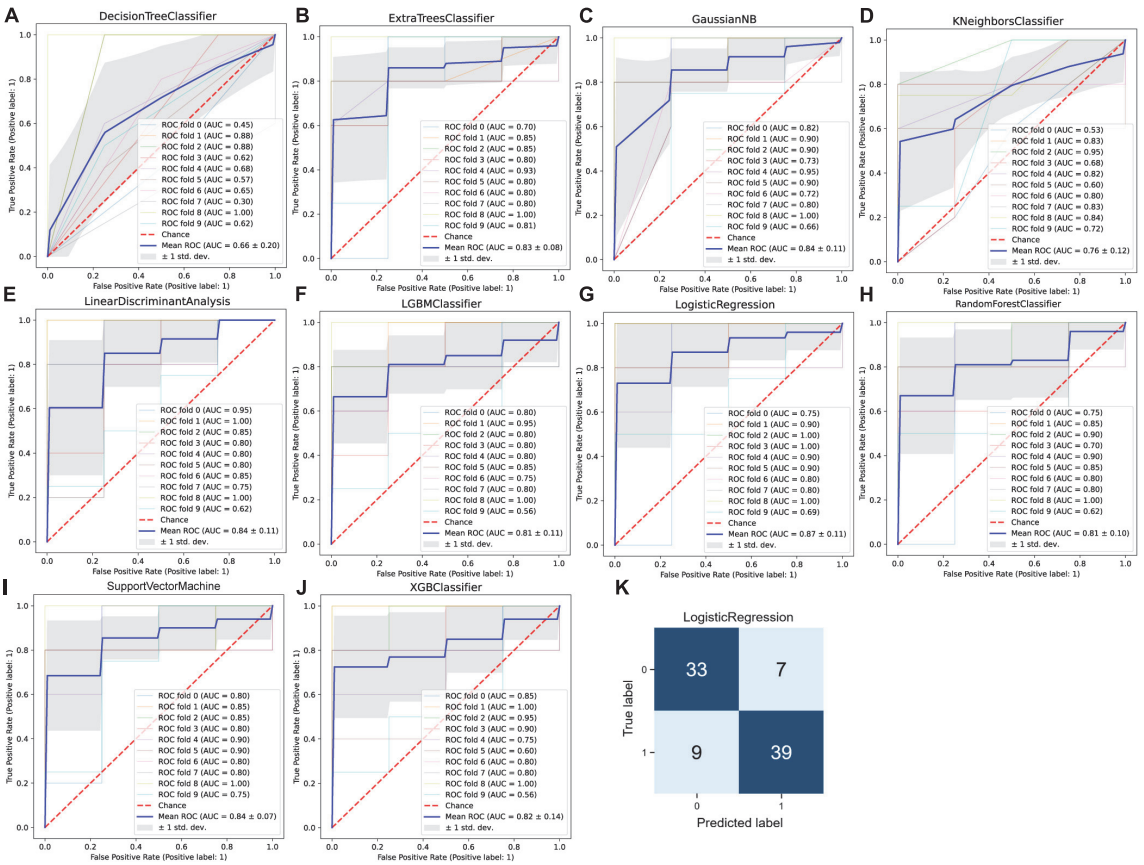

FIGURE 4

Machine learning-based prediction of rupture status in intracranial aneurysms. **(A)** Decision tree, **(B)** extra tree, **(C)** gaussian NB, **(D)** K-nearest neighbor, **(E)** linear discriminant analysis, **(F)** LightGBM, **(G)** logistic regression, **(H)** random forest, **(I)** support vector machine, and **(J)** XGBoost classifiers were used to establish the classification model. The receiver operating characteristic (ROC) curve was drawn after 10-fold cross-validation for each model, and the mean area under the ROC curve (AUC) was calculated, the mean AUC was combined with the standard deviation (SD). **(K)** The confusion matrix for the logistic regression model with the highest accuracy.

TABLE 3 Classification performance tests after gene selection by the proposed hybrid algorithm.

Model name	Accuracy	Recall	Precision	F1	AUC
XGBClassifier	0.815	0.850	0.815	0.827	0.91 ± 0.09
LGBMClassifier	0.830	0.815	0.868	0.837	0.89 ± 0.07
RandomForestClassifier	0.831	0.825	0.826	0.836	0.93 ± 0.08
ExtraTreesClassifier	0.874	0.855	0.927	0.868	0.94 ± 0.07
GaussianNB	0.850	0.810	0.902	0.845	0.93 ± 0.08
KNeighborsClassifier	0.840	0.79	0.915	0.835	0.92 ± 0.10
LogisticRegression	0.872	0.895	0.887	0.884	0.93 ± 0.07
DecisionTreeClassifier	0.690	0.685	0.798	0.757	0.71 ± 0.23
SupportVectorMachineClassifier	0.919	0.940	0.930	0.929	0.93 ± 0.07
LinearDiscriminantAnalysis	0.874	0.895	0.882	0.883	0.92 ± 0.07
Mean	**0.840**	**0.836**	**0.875**	**0.850**	**0.901**

Bold values represent the average performance.

and the grouping of ruptured and unruptured IAs, we analyzed the distribution of the differential expression of the selected 10 informative genes across samples. From the distribution plot, it can be seen that the expression pattern between the ruptured group and unruptured group was significantly different (Figure 6), as the peaks of the two groups in

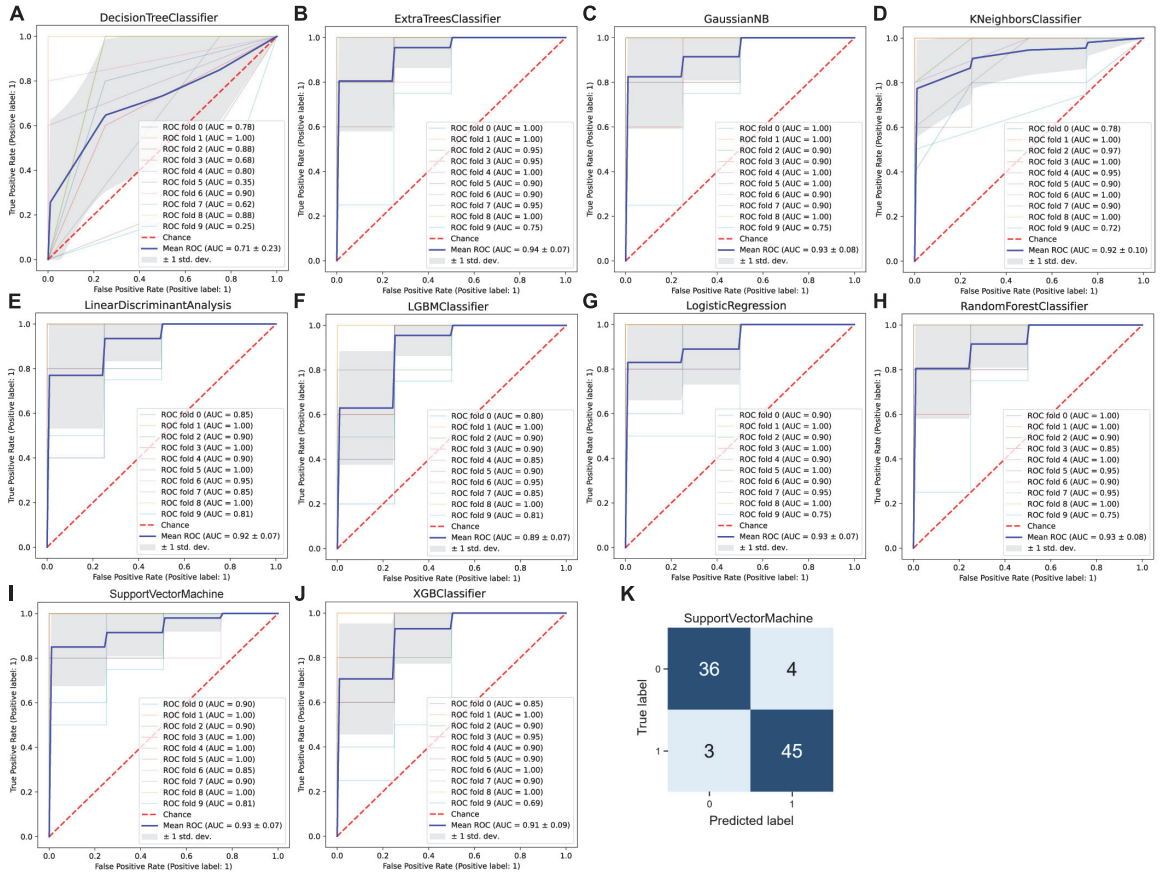

FIGURE 5
Machine learning-based prediction of rupture status in intracranial aneurysms after gene selection by the proposed hybrid gene selection algorithm. **(A)** Decision tree, **(B)** extra tree, **(C)** gaussian NB, **(D)** K-nearest neighbor, **(E)** linear discriminant analysis, **(F)** LightGBM, **(G)** logistic regression, **(H)** random forest, **(I)** support vector machine, and **(J)** XGBoost classifiers were used to establish the classification model. The receiver operating characteristic (ROC) curve was drawn after 10-fold cross-validation for each model, and the mean area under the ROC curve (AUC) was calculated, the mean AUC was combined with the standard deviation (SD). **(K)** The confusion matrix for the support vector machine model with the highest ACC.

the distribution plots were completely inconsistent. We used the normalized and the batch effect processed data when performing feature selection as well as classifier prediction, and for a better view of the roles of these 10 informative genes we additionally confirmed each gene expression per sample in the raw four GSE datasets. Violin plots were used to present the total dataset and boxplots were used to present individual raw GSE datasets. **Supplementary Figure 3** shows that YIPF1, RAB32, WDR62, ANPEP, AADAC, and TOR1B were up-regulated in ruptured IAs, while GZMK, LRRCC1, PBX1, and WBP2NL were down-regulated to some extent, indicating that YIPF1, RAB32, WDR62, ANPEP, AADAC, and TOR1B are risk factors, while the others are protective factors. Although the expression of each informative gene in the total dataset was significantly different between the ruptured and unruptured groups, some genes did not reach statistical significance in a certain GSE due to the small sample size. However, the expression pattern differences of these two groups can still be observed in each GSE dataset.

Discussion

Several studies have investigated the differences in gene expression in IAs vs. normal artery tissues to identify key genes and pathways implicated in the formation of aneurysms. For example, an 18-gene signature distinguished the presence of unruptured IAs with an area under the ROC curve of 0.91 by the SVM model using whole blood transcriptome (Poppenberg et al., 2020). Using the method of taking an intersection after statistical analysis of several GEO databases, an 11-gene signature involving the leukocyte *trans*-endothelial migration pathway has been determined (Gao et al., 2020). However, studies comparing ruptured vs. unruptured aneurysms are relatively few. Using imaging techniques and meta-analysis can accurately assess whether IAs will rupture (Etminan and Rinkel, 2016). Despite this, the exact molecular mechanisms that ultimately cause aneurysm rupture are still uncertain, and the

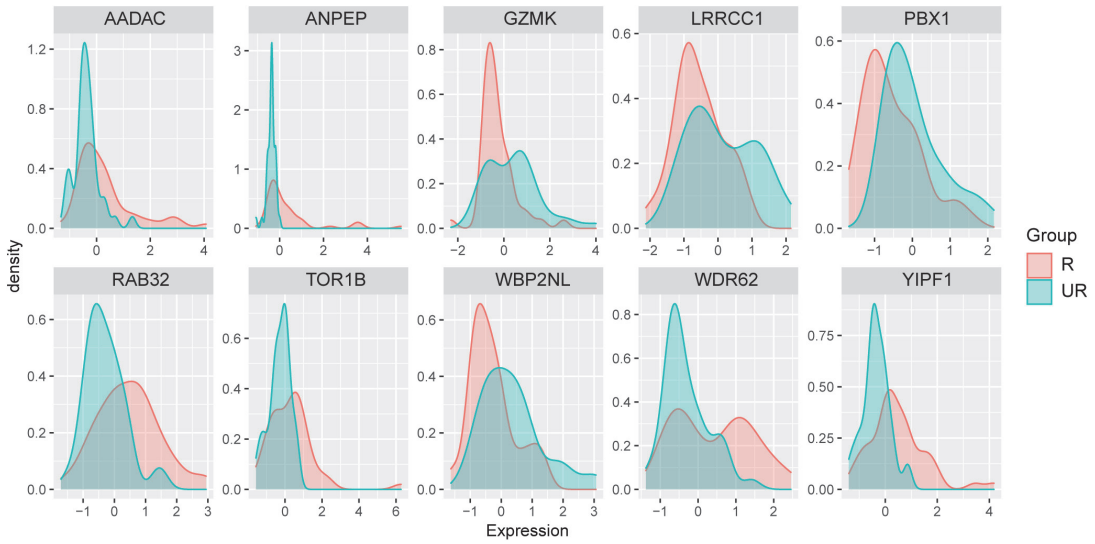

FIGURE 6
Distribution of the differential expression of the selected 10 informative genes across samples between ruptured and unruptured intracranial aneurysms.

molecular biomarkers involved in the process of IA rupture have yet to be identified.

The use of differential gene expression may identify genes and pathways involved in the process of aneurysm rupture. For example, the lysosome pathway is a new pathway for the rupture of IAs and evidence for the role of the immune response in aneurysmal rupture has been found (Kleinloog et al., 2016). Another study reported that DEGs were mainly enriched in pathways of the major histocompatibility complex class II protein complex and antigen processing and presentation (Wang et al., 2018). We also found that the distinct status of IAs could lead to differential innate immune responses, as neutrophils account for 50–70% of circulating leukocytes and are therefore the most common cells involved in innate immune responses, and they can act as signaling mediators performing various antimicrobial functions and inflammatory responses through activation and degranulation (Klopf et al., 2021). Although some pathways that play a role in the process of IA rupture have been discovered, the key genes that distinguish whether or not rupture occurs have not been elucidated. Using weighted gene co-expression network analysis, Wang et al. (2020) identified six hub genes associated with IA rupture, which tended to enrich in one pathway. In addition, the gene with the highest ranking for classification importance was not necessarily the one with the greatest fold change and vice versa (Li et al., 2022). The feature selection algorithm had the benefit over conventional statistical methods in this regard, favoring the selection of biomarkers that could be used to distinguish between IAs that had ruptured and those that had not. As we were unable to find larger studies investigating gene expression differences in IAs, in the present study, we combined four previous GSE datasets. We propose a new two-stage hybrid feature selection algorithm FCBF-MLP-PSO, which is a mixed filter- and wrapper-based gene selection algorithm. Thereafter, YIPF1, RAB32, WDR62, ANPEP, LRRCC1, AADAC, GZMK, WBP2NL, PBX1, and TOR1B containing a total of 10 effective feature subsets were determined. The highest ACC value increased from 0.817 to 0.919 (12.5% increase), the highest AUC value increased from 0.87 to 0.94 (8.0% increase), and all evaluation metrics improved by approximately 10% after being processed by our proposed gene selection algorithm. By analyzing the relationship of IAs in previous studies with each informative gene, only GZMK was reported to be associated with IA rupture (Wang et al., 2020).

In conclusion, we constructed a novel 10-gene signature (YIPF1, RAB32, WDR62, ANPEP, LRRCC1, AADAC, GZMK, WBP2NL, PBX1, and TOR1B) using the proposed two-stage hybrid algorithm FCBF-MLP-PSO and used different machine learning models to predict the rupture status in IAs (AUC = 0.94, ACC = 0.92, greater than the previous researches, as far as we know). Therefore, these 10 informative genes used to predict the rupture status of IAs can be used as complements to imaging examinations in the clinic, and provide new targets and ideas for the treatment of ruptured IAs.

Data availability statement

The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Author contributions

QL and MY conceived and designed the experiments. QL and PW analyzed the data and wrote the first draft of this manuscript. YZ, JY, and YM discussed and contributed to the data analysis. All authors reviewed and approved the final version of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (61672386), the Key Research and Development Plan of Anhui Province, China (2022a05020011), the Academic Support Project for Top-notch Talents in Disciplines (Majors) of Universities in Anhui Province, China (gxbjZD2022042), the Foreign Visiting Scholar Project for Outstanding Young Backbone Teachers of Universities in Anhui Province, China (gxgwfx2022026), and the Key University Science Research Project of Anhui Province (KJ2021A0848).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2022.1034971/full#supplementary-material

SUPPLEMENTARY FIGURE 1
The Principal Component Analysis (PCA) plots and the t-Distributed Stochastic Neighbor Embedding (t-SNE) plots before and after adjustment of the batch effects. **(A)** The PCA plots before adjustment of the batch effects, **(B)** the t-SNE plots before adjustment of the batch effects, **(C)** the PCA plots after adjustment of the batch effects, **(D)** the t-SNE plots after adjustment of the batch effects.

SUPPLEMENTARY FIGURE 2
Importance ranking of the 42 candidate features obtained by the FCBF algorithm.

SUPPLEMENTARY FIGURE 3
Expression of the selected 10 informative genes across samples in the total dataset and each of the raw datasets between ruptured and unruptured intracranial aneurysms. Violin plots are used to present the total dataset and box plots are used to present individual raw GSE datasets.

References

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell* 173, 1581–1592. doi: 10.1016/j.cell.2018.05.015

Chen, K., Xue, B., Zhang, M., and Zhou, F. (2021). "An evolutionary multitasking-based feature selection method for high-dimensional classification," in *IEEE Transactions on Cybernetics*, Piscataway, NJ: IEEE, 99. doi: 10.1109/TCYB.2020.3042243

Chen, Y., Wang, L., Li, L., Zhang, H., and Yuan, Z. (2016). Informative gene selection and the direct classification of tumors based on relative simplicity. *BMC Bioinform.* 17:44. doi: 10.1186/s12859-016-0893-0

Etminan, N., and Rinkel, G. J. (2016). Unruptured intracranial aneurysms: development, rupture and preventive management. *Nat. Rev. Neurol,* 12, 699–713. doi: 10.1038/nrneurol.2016.150

Gao, Y., Zhao, C., Wang, J., Li, H., and Yang, B. (2020). The potential biomarkers for the formation and development of intracranial aneurysm. *J. Clin. Neurosci.* 81, 270–278. doi: 10.1016/j.jocn.2020.09.072

Huang, C., Huang, X., Fang, Y., Xu, J., Qu, Y., Zhai, P., et al. (2020). Sample imbalance disease classification model based on association rule feature selection. *Pattern. Recogniti. Lett.* 133, 280–286. doi: 10.1016/j.patrec.2020.03.016

Jain, I., Jain, V. K., and Jain, R. (2018). Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. *Appl. Soft Comput.* 62, 203–215. doi: 10.1016/j.asoc.2017.09.038

Kleinloog, R., Verweij, B. H., van der Vlies, P., Deelen, P., Swertz, M. A., de Muynck, L., et al. (2016). RNA sequencing analysis of intracranial aneurysm walls reveals involvement of lysosomes and immunoglobulins in rupture. *Stroke* 47, 1286–1293. doi: 10.1161/STROKEAHA.116.012541

Klopf, J., Brostjan, C., Eilenberg, W., and Neumayer, C. (2021). Neutrophil extracellular traps and their implications in cardiovascular and inflammatory disease. *Int. J. Mol. Sci.* 22:559. doi: 10.3390/ijms22020559

Kurki, M. I, Hakkinen, S. K., Frosen, J., Tulamo, R., von und zu Fraunberg, M., Wong, G., et al. (2011). Upregulated signaling pathways in ruptured human saccular intracranial aneurysm wall: an emerging regulative role of Toll-like receptor signaling and nuclear factor-kappaB, hypoxia-inducible factor-1A, and ETS transcription factors. *Neurosurgery* 68, 1667–1675. discussion 1675-1666., doi: 10.1227/NEU.0b013e318210f001

Lei, Y., and Liu, H. (2003). "Feature selection for high-dimensional data: a fast correlation-based filter solution," in *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003)*, Washington, DC.

Li, Q., Yang, H., Wang, P., Liu, X., Lv, K., and Ye, M. (2022). XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer. *J. Transl. Med.* 20:177. doi: 10.1186/s12967-022-03369-9

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2020). TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res.* 48, W509–W514. doi: 10.1093/nar/gkaa407

Medjahed, S. A., Saadi, T. A., Benyettou, A., and Ouali, M. (2017). Kernel-based learning and feature selection analysis for cancer diagnosis. *Appl. Soft Comput.* 51, 39–48. doi: 10.1016/j.asoc.2016.12.010

Nakaoka, H., Tajima, A., Yoneyama, T., Hosomichi, K., Kasuya, H., Mizutani, T., et al. (2014). Gene expression profiling reveals distinct molecular signatures associated with the rupture of intracranial aneurysm. *Stroke* 45, 2239–2245. doi: 10.1161/STROKEAHA.114.005851

Pera, J., Korostynski, M., Krzyszkowski, T., Czopek, J., Slowik, A., Dziedzic, T., et al. (2010). Gene expression profiles in human ruptured and unruptured intracranial aneurysms: what is the role of inflammation? *Stroke* 41, 224–231. doi: 10.1161/STROKEAHA.109.562009

Perscheid, C. (2021). Integrative biomarker detection on high-dimensional gene expression data sets: a survey on prior knowledge approaches. *Brief Bioinform.* 22:bbaa151. doi: 10.1093/bib/bbaa151

Pontes, F. G. B., da Silva, E. M., Baptista-Silva, J. C., and Vasconcelos, V. (2021). Treatments for unruptured intracranial aneurysms. *Cochrane Database Syst. Rev.* 5:CD013312. doi: 10.1002/14651858.CD013312.pub2

Poppenberg, K. E., Li, L., Waqas, M., Paliwal, N., Jiang, K., Jarvis, J. N., et al. (2020). Whole blood transcriptome biomarkers of unruptured intracranial aneurysm. *PLoS One* 15:e0241838. doi: 10.1371/journal.pone.0241838

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Rose, M. J. (2011). Aneurysmal subarachnoid hemorrhage: an update on the medical complications and treatments strategies seen in these patients. *Curr. Opin. Anaesthesiol.* 24, 500–507. doi: 10.1097/ACO.0b013e32834ad45b

Rostami, M., Forouzandeh, S., Berahmand, K., and Soltani, M. (2020). Integration of multi-objective PSO based feature selection and node centrality for medical datasets. *Genomics* 112, 4370–4384. doi: 10.1016/j.ygeno.2020.07.027

Tang, C., Cao, L., Zheng, X., and Wang, M. (2018). Gene selection for microarray data classification via subspace learning and manifold regularization. *Med. Biol. Eng. Comput,*. 56, 1271–1284. doi: 10.1007/s11517-017-1751-6

Tawk, R. G., Hasan, T. F., D'Souza, C. E., Peel, J. B., and Freeman, W. D. (2021). Diagnosis and treatment of unruptured intracranial aneurysms and aneurysmal subarachnoid hemorrhage. *Mayo Clin, Proc.* 96, 1970–2000. doi: 10.1016/j.mayocp.2021.01.005

Wang, Q., Chen, X., Yi, D., Song, Y., Zhao, Y. H., and Luo, Q. (2018). Expression profile analysis of differentially expressed genes in ruptured intracranial aneurysms: in search of biomarkers. *Biochem. Biophys. Res. Commun.* 506, 548–556. doi: 10.1016/j.bbrc.2018.10.117

Wang, Q., Luo, Q., Yang, Z., Zhao, Y. H., Li, J., Wang, J., et al. (2020). Weighted gene co-expression network analysis identified six hub genes associated with rupture of intracranial aneurysms. *PLoS One* 15:e0229308. doi: 10.1371/journal.pone.0229308

Xiong, Y., Li, Q., Wang, P., and Ye, M. (2021). Informative gene selection based on cost-sensitive fast correlation- based filter feature selection. *Curr. Bioinform.* 16, 1060–1068. doi: 10.2174/1574893616666210601111850

Yang, J. B., Shen, K. Q., Ong, C. J., and Li, X. P. (2009). Feature selection for MLP neural network: the use of random permutation of probabilistic outputs. *IEEE Trans. Neural Network.* 20, 1911–1922. doi: 10.1109/TNN.2009.2032543

Ye, M., Wang, W., Yao, C., Fan, R., and Wang, P. (2019). Gene selection method for microarray data classification using particle swarm optimization and neighborhood rough set. *Curr. Bioinform.* 14, 422–431.

Yu, L., and Liu, H. (2003). "Feature Selection for high-dimensional data: a fast correlation-based filter solution," in *In Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003*, Washington, DC.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat, Commun.* 10:1523. doi: 10.1038/s41467-019-09234-6