Mathematical Biosciences and Engineering

MBE, 19(12): 13747−13781.
DOI: 10.3934/mbe.2022641
Received: 31 July 2022
Revised: 03 September 2022
Accepted: 07 September 2022
Published: 19 September 2022

http://www.aimspress.com/journal/MBE

Research article

Two-stage feature selection for classification of gene expression data based on an improved Salp Swarm Algorithm

Xiwen Qin, Shuang Zhang, Dongmei Yin, Dongxue Chen and Xiaogang Dong*

School of Mathematics and Statistics, Changchun University of Technology, Changchun 130012, China

* **Correspondence:** Email: dongxiaogang@ccut.edu.cn.

Abstract: Microarray technology has developed rapidly in recent years, producing a large number of ultra-high dimensional gene expression data. However, due to the huge sample size and dimension proportion of gene expression data, it is very challenging work to screen important genes from gene expression data. For small samples of high-dimensional biomedical data, this paper proposes a two-stage feature selection framework combining Wrapper, embedding and filtering to avoid the curse of dimensionality. The proposed framework uses weighted gene co-expression network (WGCNA), random forest and minimal redundancy maximal relevance (mRMR) for first stage feature selection. In the second stage, a new gene selection method based on the improved binary Salp Swarm Algorithm is proposed, which combines machine learning methods to adaptively select feature subsets suitable for classification algorithms. Finally, the classification accuracy is evaluated using six methods: lightGBM, RF, SVM, XGBoost, MLP and KNN. To verify the performance of the framework and the effectiveness of the proposed algorithm, the number of genes selected and the classification accuracy was compared with the other five intelligent optimization algorithms. The results show that the proposed framework achieves an accuracy equal to or higher than other advanced intelligent algorithms on 10 datasets, and achieves an accuracy of over 97.6% on all 10 datasets. This shows that the method proposed in this paper can solve the feature selection problem related to high-dimensional data, and the proposed framework has no data set limitation, and it can be applied to other fields involving feature selection.

Keywords: high-dimensional data; feature selection; swarm intelligence optimization algorithm; gene expression data; cancer classification

1. Introduction

Because the incidence rate and mortality rate of cancer are very high, it has been widely a concern all over the world, so diagnosing cancer has become a very difficult task [1]. Cancer is also a common research object in bioinformatics research. In cancer diseases, there are many features with different information, which can be used to distinguish the tissue or organ source of cancer distribution according to these features [2]. The rapid development of microarray technology in recent years has produced a large amount of ultra-high-dimensional gene expression data. Therefore, the use of gene expression data for cancer diagnosis has great advantages. However, due to the huge sample size and dimension ratio of gene expression data, small sample size and high gene dimension, it is a very challenging task to screen key genes from gene expression data. And due to the curse of dimensionality [3], irrelevant and redundant genes will increase the difficulty of model training and will also have adverse effects on the accuracy of the model.

The premise of cancer treatment is an accurate diagnosis. With the extensive development of machine learning and artificial intelligence, machine learning classification methods have occupied a certain position in the field of cancer diagnosis. In recent years, more and more researchers [4] use machine learning algorithms for cancer diagnosis. And for classification problems, feature reduction also plays a very important role, which is very effective in preventing overfitting, reducing computational complexity, and reducing model interpretability [5].

Feature selection can be roughly divided into three categories: filtering, Wrapper and embedding, which aims to solve high-dimensional problems. The wrapper [6] method is to simplify the data through the feature selection algorithm, and then construct a feature subset to train the classification algorithm, and the feature selection fitness value is the performance of the classification algorithm. The filter [7] method first calculates the correlation between the features in the dataset and the target variable, and filters the data by comparing the magnitude of the correlation. The embedding [8] introduces a regularization term in the loss function of the classification method to constrain the model, and selects features according to the performance of the classification method.

Researchers have proposed many feature selection methods based on gene expression data to achieve robust feature selection and accurate cancer diagnosis [9]. L. Sun et al. [10] proposed a neighborhood rough set-based feature selection method for cancer classification of gene expression data using an uncertainty measure based on neighborhood entropy. A. Kumar et al. [11] in their paper constructed an integrated active learning approach to achieve simplification of gene expression data using a fuzzy rough set approach. This method can improve the classification accuracy with limited samples in the training dataset. J. Lee et al. [12] proposed a new multivariate feature ranking method to improve the quality of gene selection and ultimately the accuracy of microarray data classification. They embedded the formal definition of relevance into the Markov blanket (MB) to create a new feature ranking method. X. Zheng and C. Zhang [13] proposed a model based on latent representation learning, which treats each gene as a feature and performs feature selection by computing the intra-association between samples of gene expression data and the relationship between features, i.e., in the latent representation space, rather than by comparing the importance of features in the dataset. L. Li et al. [14] Proposed a stable machine learning recursive feature elimination (StabML-RFE) strategy. They employed eight different machine learning methods and sequentially removed the least important features with recursive feature elimination (RFE). Then, each feature is sorted, and the top ranked features are selected to form the best feature subset, and a stability measure is established to evaluate

the robustness of different feature selection techniques. The selected biomarkers are also verified by different methods.

In recent years, swarm intelligence optimization algorithms have shown to be very powerful in feature selection because of their simplicity and global search capability. A. K. Shukla et al. [2] proposed a novel hybrid wrapper algorithm TLBO-GSA incorporating features of teaching-learning-based optimization (TLBO) and Gravitational Search Algorithm (GSA). This method first selects relevant genes from the gene expression dataset using mRMR and then selects informative genes from approximate data generated by mRMR using the proposed method. H. Wang et al. [15] proposed a new multidimensional population-based bacterial colony optimization method, referred to as BCO-MDP, for feature selection for classification. C. Shen and K. Zhang [16] constructed a two-stage feature selection framework based on the gray wolf optimization algorithm. In the first stage, an integer optimization problem was constructed by first training the parameters of a multilayer perceptron based on the group lasso regularization term using a modified gray wolf optimization algorithm for the initial screening of features and determination of the number of hidden layer layers; in the second stage, the multilayer perceptron was run again using the results of the stage to construct a discrete optimization problem for feature selection. C. Qu et al. [17] implemented a Harris Hawk optimization algorithm based on variable neighborhood learning for feature selection of gene expression data. It is also a two-stage framework that first performs one stage of feature selection using F-score to compress the feature space. Then the second phase of feature selection is performed using the Harris Hawk optimization algorithm based on variable neighborhood learning. A. Dabba et al. [18] used an improved Moth-flame optimization algorithm to combine the Moth-flame optimization algorithm with mutual information maximization to achieve feature selection of gene expression data. L. Sun et al. [19] constructed a feature selection algorithm combining an ant colony optimization algorithm with RelieF to achieve feature selection for tumor classification problems. Uzma et al. [20] constructed a two-stage gene selection method, aggregating three filtering methods in the first stage, and then using a genetic algorithm in the second stage in combination with an unsupervised autoencoder-based method to implement the gene selection problem for the subsequent classification task.

From the above introduction, we can see that some intelligent algorithms have been used to build cancer diagnosis frameworks, but these frameworks have some defects worthy of improvement, such as falling into local optimization. Since the minimization of the selected genes is not considered, the maximum fitness evaluation value and parameters are required to be adjusted, so the classification results are not ideal. And due to the traditional fitness function to select genes, the performance of the classifier cannot be maximized with a small subset of features. In this study, a cancer classification framework based on a small number of possible genes was established in order to accurately classify the gene expression data of cancer. The proposed framework used a two-stage feature selection method for optimal gene selection and machine learning classification. The accuracy of the algorithm and the number of selected features are combined as a fitness evaluation to accurately determine whether it is cancer.

The overall goal of this paper is to propose a feature selection framework for the feature selection problem of high-dimensional data. This framework can achieve high classification accuracy with fewer feature subsets. Specifically, this paper uses two-stage feature selection technology to achieve the classification of gene expression data. Since the dimension of the data is too large, the purpose of the first stage is to remove irrelevant and redundant features while retaining as many relevant features as possible. Because different feature selection algorithms have different advantages and disadvantages.

mRMR is a filtering feature selection algorithm. The filtering feature selection method measures the importance of features through relevant statistics, but because the process of feature selection is independent of the learner, the selected feature subset may not obtain good classification accuracy. RF feature selection is an embedded feature selection algorithm. In the embedded feature selection, the feature selection algorithm itself is embedded in the learning algorithm as a component, and some machine learning algorithms or models are used for training to obtain the weight coefficients of each feature (between 0 and 1). These weight coefficients often represent the contribution or importance of features to the model, but the adjustment of parameters has a great impact on the method. WGCNA is a system biology method, which is used to describe the correlation pattern between genes and can be used to find highly correlated gene sets. Therefore, in the first stage, we select the combination of these three methods to select the gene subset with rich information. The purpose of the second stage feature selection is to use as few features as possible to achieve higher classification accuracy, so this paper adopts the improved binary Salp Swarm Algorithm (SSA) [21] for feature selection in the second stage. Among the wrapper-based algorithms, the SSA has superior global search capability and faster convergence speed, and the main advantages of SSA are less computational effort and fewer control parameters compared to the existing optimization algorithms, namely, Particle Swarm Optimization (PSO) [22], Grey Wolf Optimizer (GWO) [23], Whale Optimization Algorithm (WOA) [24], and Sine Cosine Algorithm (SCA) [25].

The rest of this paper is organized as follows. The second section is the method, which introduces the feature selection algorithm and classification method used in this paper; the third section is the proposed method, which introduces the proposed improved SSA and binary feature selection; the fourth section is the empirical study, which introduces the details and parameter settings of the empirical part of this paper; the fifth section is the results and discussion, which analyzes the results of the experiments in this paper and compares them with advanced methods. The sixth section is the conclusion, a summary of the work of this paper, and future research directions.

2. Methods

2.1. Weighted gene co-expression network

Weighted gene co-expression network analysis (WGCNA) [26,27] is a systems biology approach used to describe the association patterns of different genes, aiming to find co-expressed gene modules and to explore the association between gene networks and phenotypes of interest, as well as the core genes in the network. WGCNA uses the information of thousands or nearly 10,000 genes with the greatest variation or all genes to identify gene sets of interest, and perform significant association analysis with phenotypes. The first is to make full use of the information, and the second is to convert the association between thousands of genes and phenotypes into associations between several gene sets and phenotypes, eliminating the problem of multiple hypothesis testing and correction.

From the methodological point of view, WGCNA is divided into two parts: expression clustering analysis and phenotype association, which mainly include several steps of correlation coefficient calculation between genes, co-expression network construction, gene module identification, and module-trait association.

Step 1: Use Pearson's correlation coefficient to calculate the correlation coefficient between any two genes and construct the co-expression similarity matrix S_{ij},

$$S_{ij} = \left|cor(x_i, x_j)\right| \tag{1}$$

where x_i and x_j are the i-th and j-th genes.

Step 2: Construct the adjacency matrix a_{ij}, construct the scale-free network, and determine the power index β. Based on the adjacency matrix, construct the topological overlap matrix (TOM matrix), and the TOM matrix uses w_{ij} to represent the connectivity of the connected nodes

$$a_{ij} = S_{ij}^{\beta} \tag{2}$$

$$w_{ij} = \frac{l_{ij}+a_{ij}}{min\{k_i,k_j\}+1-a_{ij}} \tag{3}$$

where $l_{ij} = \sum_u a_{iu}\, a_{uj}$ and $k_i = \sum_u a_{iu}$, u is the total number of genes analyzed for co-expression.

Step 3: The topological overlap is transformed into a dissimilarity matrix using $1 - w_{ij}$. Hierarchical clustering trees are constructed, gene modules are generated using dynamic shearing, and genes with similar expression patterns are clustered within the same branch to determine gene modules.

Step 4: Modules are associated with external phenotypic information of interest to find modules with high phenotypic correlation, and the genes within the modules are the selected genes.

2.2. Random Forest

Random Forest is an algorithm proposed by Leo Breiman (2001) where the model uses a collection of decision trees to perform various tasks (training, classification, and prediction of samples). The random forest can also filter features by evaluating the importance of each feature in the model, which is an embedded feature selection algorithm.

1) k variables are randomly selected from the collected data set for a total of m variables (where k is less than or equal to m), and then a decision tree is built based on these k variables.

2) Repeating the above process n times to construct n different decision trees.

3) Then for each decision tree the outcome is predicted using random variables and all predicted outcomes are recorded, resulting in n outcomes from n decision trees.

4) The number of votes obtained for each prediction result is calculated, i.e., the prediction result with the highest number of votes is taken as the final prediction result of the random forest algorithm.

Random forest feature selection means that the feature variables in the random forest are sorted in descending order according to VI (Variable Importance), and then you find your own will to select the desired number of features.

2.3. Max-Relevance and Min-Redundancy

Maximum Correlation Minimum Redundancy (mRMR) is a filtered feature selection method proposed by H. Peng et al. [28], which can use mutual information, correlation or distance similarity scores to select features. The principle is very simple, which is to find the set of features in the original set of features that are most correlated with the final output target variable, but the features are least correlated with each other.

Max-Relevance:

$$max\ D\ (S,c),\ D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i;c) \tag{4}$$

Min-Redundancy:

$$min\ R\ (S),\ R = \frac{1}{|S|^2}\sum_{x_i,x_j \in S} I\left(x_i;x_j\right) \tag{5}$$

The mRMR score is:

$$max\ \Phi\ (D,R),\ \Phi = D - R \tag{6}$$

where $I(x_i;c)$ and $I\left(x_i;x_j\right)$ are the mutual information between features and categories, features and features, respectively, and S is the subset of features $\{x_i\}$.

2.4. Salp Swarm Algorithm

The Salps population is divided into two groups: leaders and followers. The leader is the Salps at the front of the food chain, while the rest of the Salps are considered followers. As these names imply, the leader leads the population and the followers follow each other (direct or indirect leaders).

Population initialization:

$$X_j^i = rand(N,D) \times (ub(j) - lb(j)) + lb(j) \tag{7}$$

Initialize the location of the salps $X_j^i (i = 1,2,\cdots,N, j = 1,2,\cdots,D)$.

The leader's position update formula is:

$$x_j^i = \begin{cases} F_j + c_1((ub_j - lb_j))c_2 + lb_j & c_3 \geq 0.5 \\ F_j - c_1((ub_j - lb_j))c_2 + lb_j & c_3 < 0.5 \end{cases} \tag{8}$$

Where x_j^i is the leader position in the j-th dimension. Choose the first salp as the leader. Where F_j is the position of the food source in the j-th dimension, ub_j is the upper bound of the j-th dimension, lb_j is the lower bound of the j-th dimension, and c_1, c_2, c_3 are random numbers.

c_1 is the most important parameter in SSA because it balances exploration and exploitation, and is defined as follows.

$$c_1 = ze^{-(\frac{4l}{L})^2} \tag{9}$$

where l is the current iteration and L is the maximum number of iterations.

The formula for updating the position of followers is:

$$x_j^i = \frac{1}{2}(x_j^i + x_j^{i-1}) \tag{10}$$

Where $i \geq 2$, x denotes the position of the i-th follower in the j-th dimension.

2.5. Classification methods

LightGBM (Light Gradient Boosting Machine) is a framework for implementing the GBDT algorithm, which uses weak classifiers to iteratively train to obtain the optimal model, supports efficient parallel training, has faster training speed, better accuracy, and is less prone to overfitting. LightGBM is efficient and fast in processing large-scale data sets.

SVM (Support Vector Machine) is a generalized linear classifier that minimizes the empirical error and maximizes the geometric edge area at the same time. SVM maps the vectors into a higher dimensional space with a maximum interval hyperplane and separates the hyperplanes to maximize the distance between the two parallel hyperplanes. SVM has many unique advantages in dealing with small sample, nonlinear and high-dimensional pattern recognition problems.

XGBoost (eXtreme Gradient Boosting), also called extreme gradient boosting tree, is an implementation of the boosting algorithm that focuses on reducing bias, i.e., reducing the error of the model. Therefore, it uses multiple base learners, each of which is relatively simple, to avoid overfitting. XGBoost is suitable for structured data, and it is fast and effective in processing large-scale data sets.

MLP (Multi-Layer perceptron) is the basic algorithm of Deep Neural Networks (DNN), which can have multiple hidden layers in the middle except for the input and output layers, and the simplest MLP contains only one hidden layer, i.e., a three-layer structure. MLP has good fault tolerance and strong self-adaptive and self-learning functions.

KNN uses the training data to partition the feature vector space and uses the result of the partition as the final algorithmic model. For any n dimensional input vector, corresponding to a point in the feature space, respectively, the output is the category label or a predicted value corresponding to that feature vector. The prediction of labels only depends on the labels of several samples closest to the unknown samples. KNN algorithm is suitable for classification of data sets with unbalanced samples.

3. Proposed methods

3.1. Improved Salp Swarm Algorithm

Instead of random number generation by the original algorithm, chaotic mapping can generate chaotic numbers between 0 and 1. Chaotic sequences can often achieve better results than randomly generated random numbers during operations such as population initialization, selection, crossover, and mutation. In this paper, the PWLCM chaotic mapping is used to initialize the position of the Salps population. The formula is:

$$X_j^i = lb_i + (ub_i - lb_i)y_j^i \tag{11}$$

$$X(t+1) = F_p(X(t)) = \begin{cases} X(t)/p\ , & 0 < X(t) < p \\ (X(t)-p)/(0.5-p)\ , & p < X(t) < 0.5 \\ (1-X(t)-p)/(0.5-p)\ , & 0.5 < X(t) < 1-p \\ (1-X(t))/p, & 1-p < X(t) < 1-p \end{cases} \tag{12}$$

Where $p \in (0, 0.5)$.

The leader's position update formula is:

$$x_j^i = \begin{cases} F_j + c_1((ub_j - lb_j))c_2 + lb_j & c_3 \geq 0.5 \\ F_j - c_1((ub_j - lb_j))c_2 + lb_j & c_3 < 0.5 \end{cases} \tag{13}$$

In the original method, the first Salps is selected as the leader, but only the first one is easy to fall into the local optimum, so this paper selects the first third of Salps as the leader.

The follower's position update formula is based on Newton's laws of motion:

$$V_t = V_0 + at \tag{14}$$

$$x = V_0 t + \frac{1}{2}at^2 \tag{15}$$

$$a = \frac{V_{final} - V_0}{t} \tag{16}$$

$$V_{final} = \frac{x_j^{i-1} - x_j^i}{t} \tag{17}$$

which also gets

$$x_j^i = \frac{1}{2}(x_j^i - x_j^{i-1}) \tag{18}$$

This paper updates the formula with Eq (18) as a follower. The specific steps are shown in Algorithm 1.

Algorithm 1: Improved Salp Swarm Algorithm

Inputs: training set, test set, population size, the maximum number of iterations
Initialization:
01: The PWLCM chaotic mapping according to Eqs (11) and (12) initializes the position matrix A of the Salps
Optimization Process:
02: When iter < Tmax_iter
03 The value of the Salps fitness function is calculated according to Eq (19)
04: Ranking of fitness function values
05: If the optimal fitness value < the location of the food
06: Assign the position of the optimal fitness function value to the food position
07: Generate c_1 according to Eq (9), randomly generate c_2, c_3
08: The first one-third of the Salps updates the leaders' position according to Eq (13)
09: The back of the Salps updates the followers' position according to Eq (18)
10: Iter = iter + 1
Output: Optimal Salps position

3.2. Binary feature selection

The basic principle of feature selection in the SSA is to use an improved binary SSA to find an

optimal binary encode, each bit in the encode corresponds to a feature, if the i-th position is "1", the corresponding feature is selected and the feature will appear in the classifier if it is "0", it means that the corresponding feature is not selected and the feature will not appear in the classifier. The basic steps are:

Step 1: Encoding. Using the binary encoding method, the value of each position of the binary code, "0" means the feature is not selected and "1" means the feature is selected.

Step 2: Initial population generation. N initial matrices are randomly generated to form the initial population, and the number of populations is generally set to 50 to 100.

Step 3: The fitness function. The fitness function indicates the superiority or inferiority of an individual or solution.

Step 4: The update strategy of the population is determined by the fitness function value, and the next iteration is performed to continue the search for the optimal fitness function value.

Step 5: If the set number of iterations is reached, the best subset of genes is returned and used as the basis for feature selection, and the algorithm ends. Otherwise, go back to Step 4 to continue the next generation of iterations. The specific steps are shown in Figure 1.

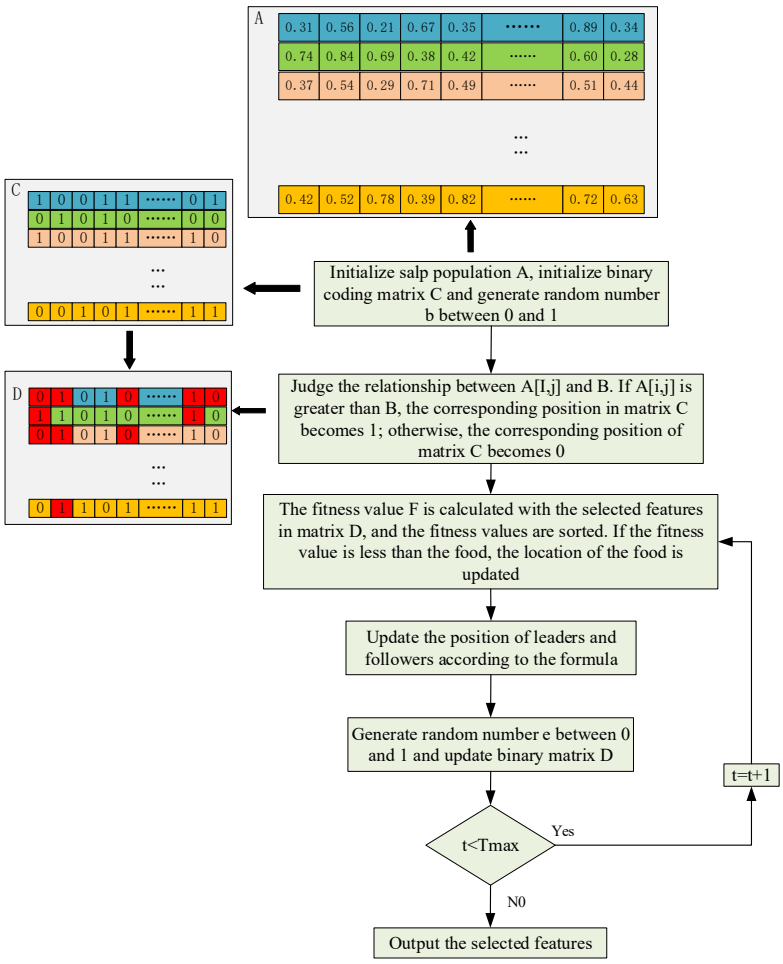

Figure 1. ISSA feature selection.

3.3. Function of fitness

For the intelligent algorithm feature selection problem, the construction of the fitness function is very important and mainly considers two aspects: first, the number of genes, i.e., the proportion of selected features to the total number of features. The fewer the selected features, the smaller the fitness value; and the second is the classification accuracy. So the fitness function is shown in Eq (19), which is mainly based on the classification ability and the number of features of the machine learning method.

$$fitness = \omega * (1 - accuracy) + (1 - \omega)(\frac{F}{N}) \tag{19}$$

Where ω is a constant between $0 \sim 1$, F is the number of features selected in each iteration, N is the total number of features, and $accuracy$ is the accuracy of the classification algorithm.

4. Empirical research

This section presents a two-stage feature selection framework to achieve more accurate cancer classification, and compares it with some feature selection algorithms based on intelligent optimization algorithms, which have been advanced in recent years. The overall framework of this paper is shown in Figure 2, which can be simply summarized as the following four steps:

Step 1: Preprocessing of gene expression datasets

In this paper, the 10 public datasets are pre-processed by first removing outlier points and duplicate values and then normalizing the datasets.

Step 2: First stage feature selection

In order to effectively filter out highly redundant and irrelevant genes, this paper adopts a more effective combination - the combination of three feature selection algorithms to identify key genes, including filtering and embedding, and the method to identify whether there is a common expression pattern between samples, which are mRMR, RF importance feature selection and WGCNA. The first N features of the three feature selection algorithms are merged to eliminate redundant and unimportant genes.

Step 3: Second stage feature selection

In this paper, an improved SSA is used to further compress the feature subsets. Firstly, the gene subset obtained in Step 2 was binary coded, and then the accuracy of the classification algorithm and the number of selected features were combined as the fitness function to enter the iteration and search for the optimal subset.

Step 4: Classifier

In this paper, six base classifiers, namely LightGBM, RF, SVM, XGBoost, MLP, and KNN, are used to classify by the feature subset selected in Step 3.

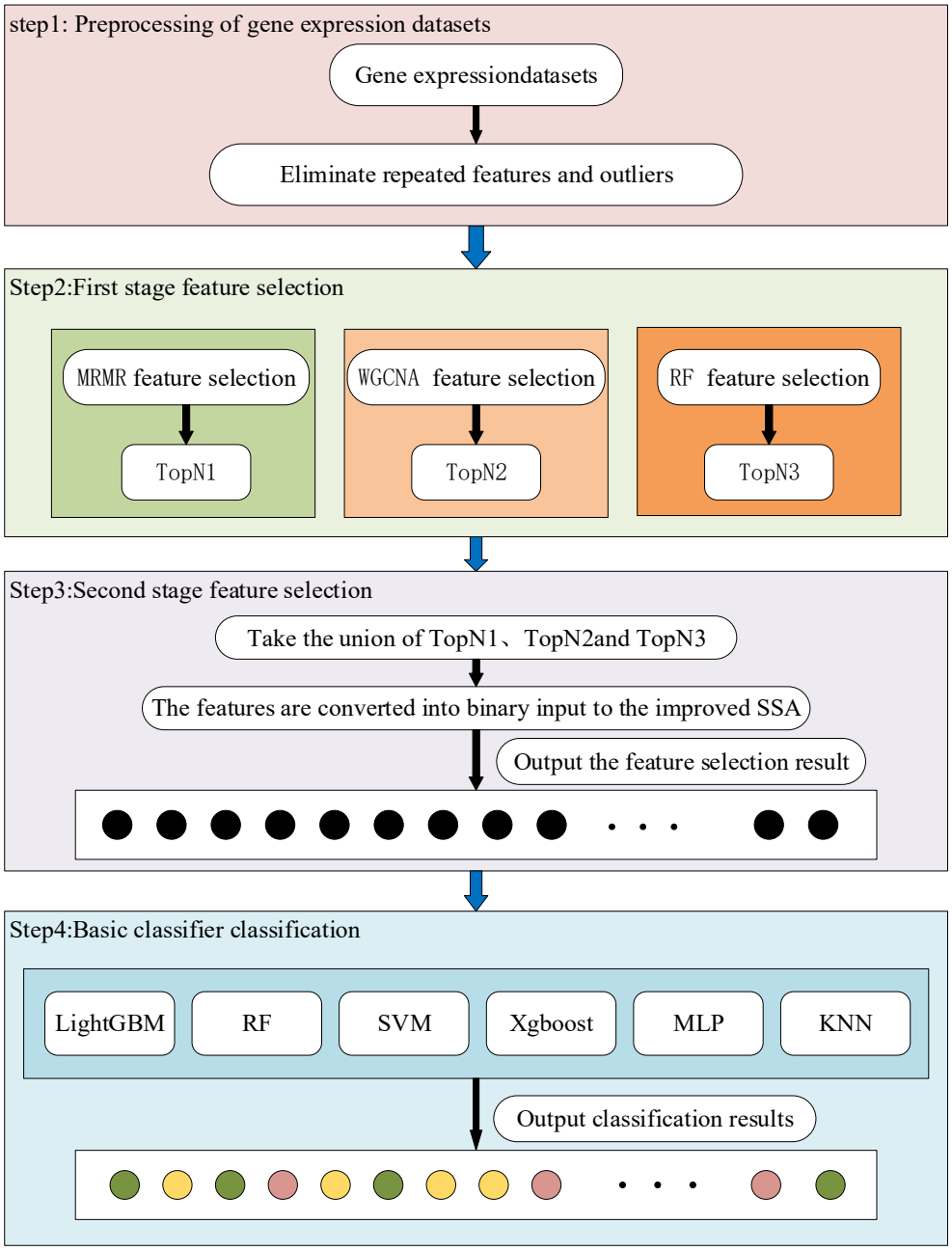

Figure 2. Overall flowchart of two-stage feature selection for cancer classification.

4.1. Datasets

To verify the effectiveness and extensiveness of the proposed framework, 10 cancer gene expression datasets [29] are used in this paper, among which 5 datasets are binary classification data and 5 datasets are multi-label classification data, as shown in Table 1, which shows some basic information of the datasets, including the number of genes, the number of samples and the number of categories. In the subsequent experiments, 80% of the samples in each dataset are selected for training in this paper, and the remaining samples are used as the test set.

Table 1. Datasets.

Dataset	Instances	Genes	Classes
Breast	97	24,481	2
CNS	60	7129	2
Colon	60	2000	2
Leukemia_3c	72	7129	3
Leukemia_4c	72	7129	4
Leukemia	72	7129	2
Lung	181	12,600	5
MLL	72	12,582	3
Ovarian	253	15,154	2
SRBCT	83	2308	4

4.2. Data preprocessing

In order to overcome the influence of dimension on the results and ensure the effectiveness of the analysis results, the data is preprocessed. First, by calculating the mean and variance of each feature, the genes that are all 0 are eliminated; Then detect outliers through sample clustering. As shown in Figure 3, take the red line as the dividing line, and the samples above the red line as outliers, and eliminate these samples; Finally, the max-min normalization processing is carried out on the data.

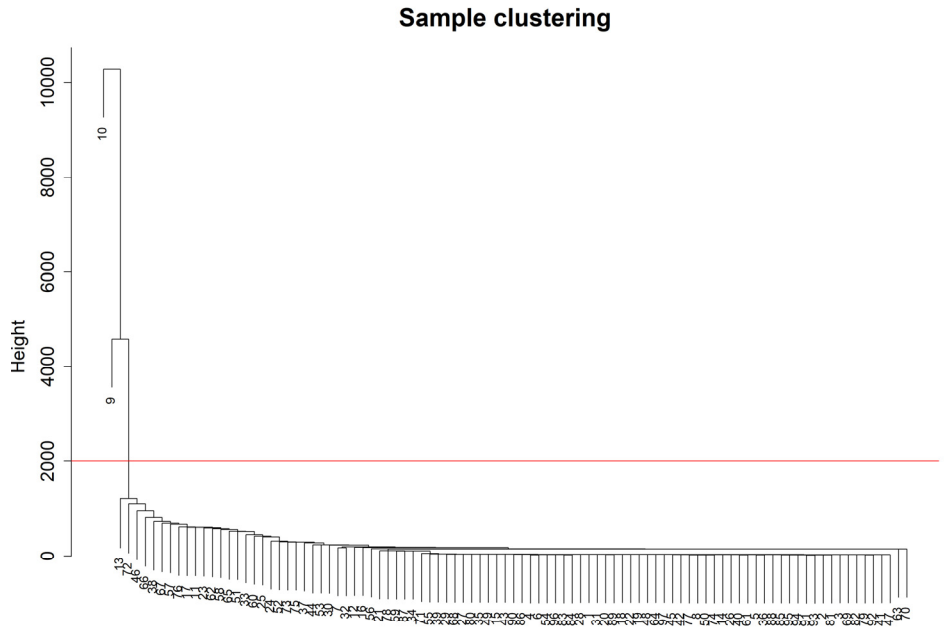

Figure 3. Sample Clustering tree of Breast dataset.

4.3. Experimental setup

Considering the efficiency and computational complexity, the population number of the proposed ISSA method is set to 100, the number of iterations is 200, and the weight in the fitness function is 0.99. To avoid unfair comparisons, the population size settings and iterations of other algorithms are kept consistent with the ISSA algorithm. The parameter settings used in this algorithm are shown in Table 2.

Table 2. Parameter setting.

S.No	Parameters	Value
1	Population Size	100
2	Number of generations	200
3	w in the fitness function	0.99
4	c1 and c2	(0, 1)
5	c3	(0, 2]
6	Inertia weight in PSO	0.6
7	The SCA of a	2

4.4. Comparison algorithm

The performance of the proposed algorithm is tested on 10 gene expression datasets to evaluate the effectiveness of the proposed algorithm. In the part of feature selection, it is compared with PSO, GWO, SCA, WOA and SSA, and six classification algorithms are used. In order to further evaluate the performance of the proposed method, it is also compared with the advanced methods in the literature, as shown in Table 3.

Table 3. Comparison method.

Key	Method name	Reference
TLBOGSA	Teaching learning-based algorithm and gravitational search algorithm	[2]
IGWO-MLP	A two-stage improved gray wolf optimization and multilayer perceptron	[16]
RFACO	ReliefF and Ant Colony Optimization Algorithm	[19]
EAK	Ensemble of three filter methods and Autoencoder-based k-means clustering	[20]
FADNE	Neighborhood entropy-based uncertainty measures	[10]
ATLBO	Adaptive inertia weight teaching-learning-based optimization algorithm	[30]
MOGT	Via multi-objective graph theoretic-based method	[31]
AC-MOFOA	Adaptive chaotic multi-objective forest optimization algorithm	[32]
GWO-TRIZ	Gray Wolf Optimizer enhanced with TRIZ-inspired operators	[33]
IG-MBKH	Information gain (IG) and an improved binary krill herd	[34]

4.5. Evaluation indicators

In this paper, accuracy, precision, recall, and F1-score are used as the evaluation metrics of the classifier. The formula is as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{20}$$

$$precision = \frac{TP}{TP+FP} \tag{21}$$

$$recall = \frac{TP}{TP+FN} \tag{22}$$

$$F1-score = \frac{2TP}{2TP+FP+FN} \tag{23}$$

where TP, TN, FP, FN are true positive, true negative, false positive, and false negative, respectively.

5. Results and discussion

5.1. The first stage feature selection

WGCNA obtained the dissimilarity matrix by TOM matrix and then obtained the final gene module based on hierarchical clustering with the dynamic cut method. Figure 4 shows the correlation between different modules and categories of gene expression datasets obtained in the WGCNA method. Only the Breast and CNS datasets are listed in this paper, and the results of the rest of the datasets are in the appendix. For each dataset, the modules with a high correlation with the category and small p-value are selected in this paper, and the genes inside these modules are analyzed for gene enrichment to extract the most important key genes for further feature selection. The upper values in each module are correlations, the values in parentheses represent p-values, the red modules indicate positive correlations and the blue modules indicate negative correlations. In general, genes within modules of the same color have a high degree of similarity, while genes in gray modules indicate genes that cannot be assigned to any module. The paper selects genes with greater relevance within the modules by performing gene enrichment analysis on the genes within each module, and finally, the Breast dataset selected 6 modules among all modules, and those modules were removed due to the weak relevance of other modules to the category, and a total of 76 genes were selected in all columns of the 6 modules; similarly, the CNS dataset selected 6 modules out of all modules and extracted a total of 87 features; the Colon dataset selected 2 modules out of all modules and extracted 96 features; the Leukemia_3c dataset selected 5 modules and extracted 86 features; the Leukemia_4c dataset selected 4 modules and extracted 119 features; the Leukemia dataset selected 5 modules and extracted 91 features; the Lung dataset selected 4 modules and extracted 137 features; the MLL dataset selected 6 modules and extracted 289 features; the Ovarian dataset selected 3 modules and extracted 89 features; and the SRBCT dataset selected 2 modules and extracted 28 features.

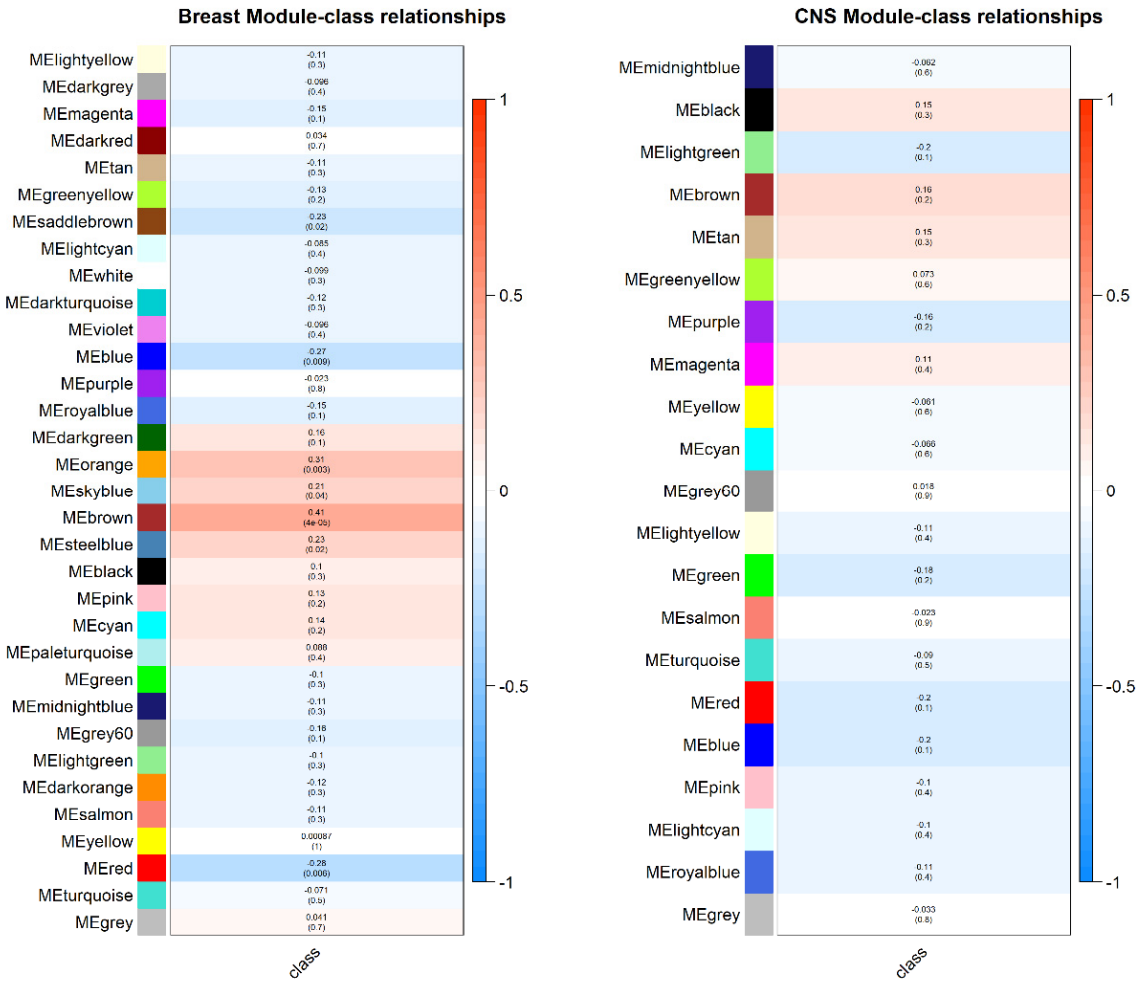

Figure 4. Correlation of modules with categories in Breast and CNS datasets WGCNA.

In the feature selection based on the random forest method, there is a certain randomness in the importance ranking of the features obtained from the random forest. Therefore, in this paper, we first run the random forest 10 times repeatedly and use the Gini index of the model with the best accuracy as the criterion for importance ranking. Then repeat the ten-fold cross-validation five times and draw the error rate curve, and take the number of features corresponding to the position with a lower error rate as the number of features to be extracted; then extract the corresponding number of features according to the feature importance ranking result. As shown in Figure 5, only CNS and Leukemia_3c datasets are listed, the rest are in the appendix, and the red line is the corresponding characteristic number when the lowest point occurs for the first time. As can be seen from Figure 5, Breast selected the top 170 features in terms of feature importance; CNS selected 55 features; Colon selected 30 features; Leukemia_3c, Leukemia_4c, and Leukemia selected 70, 170, and 140 features, respectively; Lung, MLL, Ovarian, and SRBCT 80, 150, 180 and 60 features were selected for Lung, MLL, Ovarian, and SRBCT, respectively.

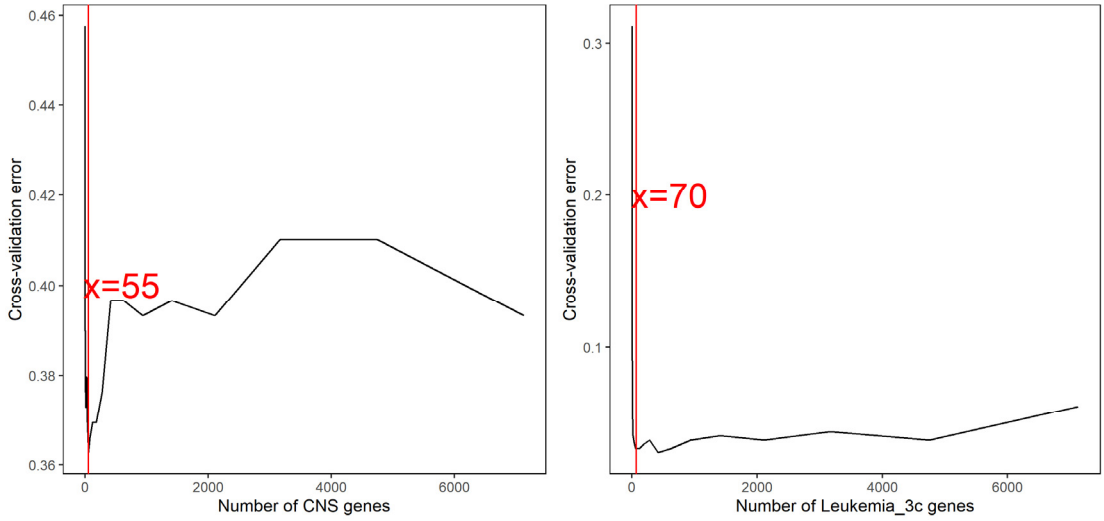

Figure 5. Error rate curve of RF.

In mRMR score-based feature selection, the maximum correlation between features is calculated first, then the minimum redundancy between features is calculated, and the mRMR score of each feature can be obtained by the maximum correlation minus the minimum redundancy, and finally, the mRMR score is ranked, and the corresponding number of features can be selected according to the ranking result. In this paper, when the number of features is less than 5000, 50 features are selected; when the number of features is less than 10,000, 80 features are selected; when the number of features is more than 10,000, 100 features are selected, and this is the criterion for the number of features selected by mRMR. In Table 4, this paper summarizes the number of features selected by the three feature selection methods and the final number of features after taking the concatenation of the three methods, which are the initial number of features for the next step of feature selection. Tables 5 and 6 are the classification results of all features with all features and separate feature selection methods.This paper lists two classification methods, LightGBM and XGBoost, and the results of the other four methods are in the appendix.

Table 4. Number of features selected by each feature selection method.

Dataset	all	WGCNA	RF	mRMR	Union
Breast	24,481	76	170	100	325
CNS	7129	87	55	80	216
Colon	2000	96	30	50	148
Leukemia_3c	7129	86	70	80	182
Leukemia_4c	7129	119	170	80	336
Leukemia	7129	91	140	80	241
Lung	12,600	137	80	100	283
MLL	12,582	289	170	100	489
Ovarian	15,154	89	170	100	300
SRBCT	2308	28	60	50	104

Table 5. Classification accuracy of LightGBM method.

Dataset	All	WGCNA	RF	mRMR	Union
Breast	0.58	0.63	0.84	0.84	0.79
CNS	0.58	0.42	0.92	0.67	0.58
Colon	0.77	0.62	0.92	0.92	0.77
Leukemia_3c	0.80	0.80	0.93	0.87	0.93
Leukemia_4c	0.8	0.73	0.87	0.93	0.87
Leukemia	0.8	0.87	0.87	0.93	0.93
Lung	0.83	0.90	0.93	0.93	0.98
MLL	0.79	0.93	0.86	0.93	0.86
Ovarian	0.98	0.76	0.96	0.98	0.96
SRBCT	0.94	0.94	0.94	0.94	0.94

Table 6. Classification accuracy of XGBoost method.

Dataset	All	WGCNA	RF	mRMR	Union
Breast	0.63	0.69	0.84	0.68	0.74
CNS	0.58	0.50	0.83	0.75	0.58
Colon	0.85	0.85	0.92	0.85	0.62
Leukemia_3c	0.87	0.73	0.93	0.87	0.87
Leukemia_4c	0.73	0.73	0.87	0.93	0.87
Leukemia	0.8	0.8	0.80	0.93	0.93
Lung	0.85	0.95	0.95	0.90	0.98
MLL	0.71	0.86	0.93	0.86	0.86
Ovarian	0.9	0.73	0.92	0.98	0.96
SRBCT	0.88	0.81	0.88	0.88	0.94

5.2. The second stage feature selection

To reduce the number of selected genes, achieve further improvement in classification accuracy and a further reduction in computational effort, this paper performs the next step of feature selection based on an improved binary SSA, where the input dimension of the method is the concatenation of the final number of features selected by the three feature selection algorithms in the previous section. The features are first binary coded, with 0 indicating that this feature is not selected and 1 indicating that this feature is selected. Since the features selected by the intelligent algorithm will vary from one classifier to another, to verify the effectiveness of the method proposed in this paper, each classifier is repeatedly run 10 times, and each evaluation index at the end is the average of the 10 times. As shown in Table 7, six classification methods, namely LightGBM, RF, SVM, XGBoost, MLP, and KNN, are used in this paper to classify the features selected by ISSA, and the classification results are evaluated using four evaluation metrics. Compared with Tables 5 and 6, the method proposed in this paper achieves high classification accuracy with fewer features on 10 public datasets, and achieves good performance on datasets with far fewer samples than the number of features.

Table 7. Classification results after ISSA.

Dataset	Performance	LightGBM	RF	SVM	XGBoost	MLP	KNN
Breast	Acc	0.990	0.938	0.897	0.975	0.980	1.000
	precision	0.990	0.943	0.910	0.975	0.980	1.000
	recall	0.990	0.938	0.897	0.975	0.980	1.000
	f1-score	0.990	0.940	0.895	0.975	0.980	1.000
CNS	Acc	0.992	0.936	0.944	0.992	0.992	0.985
	precision	0.993	0.942	0.949	0.993	0.992	0.986
	recall	0.992	0.936	0.944	0.992	0.992	0.985
	f1-score	0.992	0.928	0.937	0.991	0.990	0.985
Colon	Acc	0.976	0.952	0.968	0.920	0.968	0.968
	precision	0.980	0.961	0.973	0.994	0.974	0.972
	recall	0.976	0.952	0.968	0.992	0.968	0.968
	f1-score	0.977	0.954	0.968	0.993	0.969	0.968
Leukemia_3c	Acc	1.000	0.993	1.000	0.965	0.986	1.000
	precision	1.000	0.995	1.000	0.979	0.988	1.000
	recall	1.000	0.993	1.000	0.965	0.986	1.000
	f1-score	1.000	0.993	1.000	0.969	0.986	1.000
Leukemia_4c	Acc	1.000	0.967	0.972	0.993	0.958	0.979
	precision	1.000	0.962	0.978	0.994	0.969	0.970
	recall	1.000	0.967	0.972	0.993	0.958	0.979
	f1-score	1.000	0.961	0.971	0.993	0.958	0.973
Leukemia	Acc	1.000	0.993	1.000	1.000	1.000	1.000
	precision	1.000	0.994	1.000	1.000	1.000	1.000
	recall	1.000	0.993	1.000	1.000	1.000	1.000
	f1-score	1.000	0.993	1.000	1.000	1.000	1.000
Lung	Acc	0.992	0.959	0.982	0.994	0.981	0.986
	precision	0.992	0.962	0.983	0.995	0.983	0.987
	recall	0.992	0.960	0.981	0.994	0.981	0.986
	f1-score	0.990	0.955	0.976	0.993	0.979	0.985
MLL	Acc	1.000	1.000	1.000	1.000	1.000	1.000
	precision	1.000	1.000	1.000	1.000	1.000	1.000
	recall	1.000	1.000	1.000	1.000	1.000	1.000
	f1-score	1.000	1.000	1.000	1.000	1.000	1.000
Ovarian	Acc	1.000	1.000	0.998	1.000	1.000	0.990
	precision	1.000	1.000	0.998	1.000	1.000	0.990
	recall	1.000	1.000	0.998	1.000	1.000	0.990
	f1-score	1.000	1.000	0.998	1.000	1.000	0.990
SRBCT	Acc	1.000	1.000	1.000	1.000	1.000	1.000
	precision	1.000	1.000	1.000	1.000	1.000	1.000
	recall	1.000	1.000	1.000	1.000	1.000	1.000
	f1-score	1.000	1.000	1.000	1.000	1.000	1.000

In order to verify the validity of the features we selected, we selected features that appeared more than five times in ten results, divided the samples in these features into the Colon Cancer patient group and the normal group, compared their differences, and performed Mann-Whitney test. The results are shown in Figure 6, in which only 10 features are listed, which "*": $p < 0.05$; "**": $p < 0.01$; "***": $p < 0.001$; "****": $p < 0.0001$.

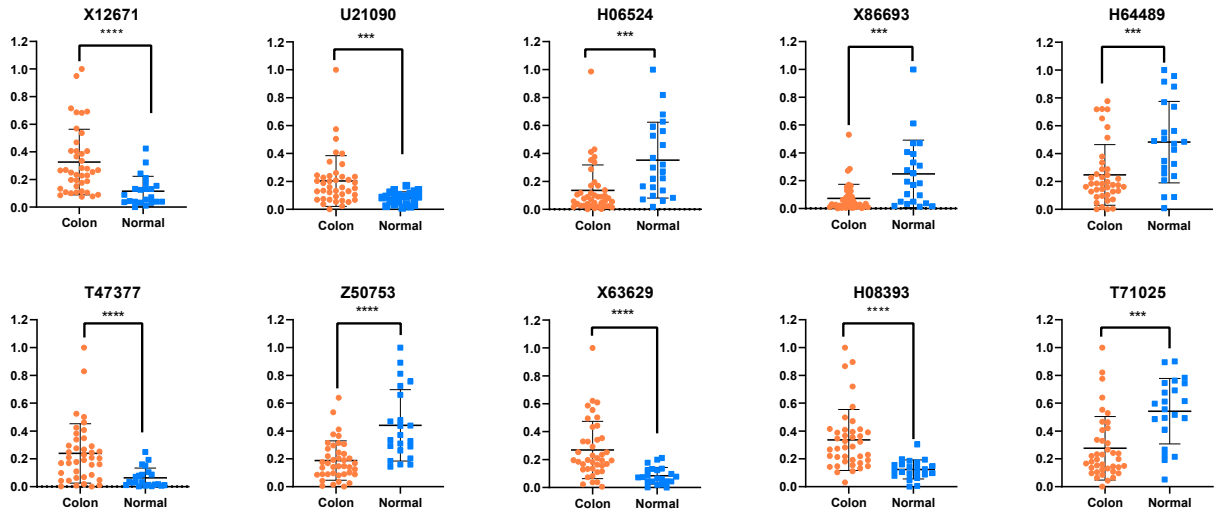

Figure 6. Gene expression profiles of a subset of selected features of colon cancer.

5.3. Comparative analysis

In recent years, using the wrapper method to improve the quality of feature subsets has become a research hotspot. In this paper, on 10 gene expression public data sets, the method proposed in this paper is compared with the current advanced methods, namely PSO, GWO, WOA, SCA and the original SSA. As shown in Tables 8 and 9, Table 8 is the classification result after constructing the fitness function feature selection based on the accuracy of LightGBM, and shows the mean and variance of the results of ten runs. It can be seen that better or similar results have been achieved on ISSA. Similar results. Similarly, Table 9 is the MLP (the results of the other four methods are in the appendix). To be fair, experiments are performed based on WGCNA, mRMR and RF feature selection.

Table 8. Classification results of different intelligent algorithms on LightGBM.

Dataset	Performance	PSO	GWO	WOA	SCA	SSA	ISSA
Breast	Mean	0.936	0.93	0.947	0.984	0.958	0.99
	Var	± 0.0049	± 0.0046	± 0.0020	± 0.0031	± 0.0031	± 0.0004
CNS	Mean	0.926	0.968	0.934	0.984	0.975	0.992
	Var	± 0.0039	± 0.0017	± 0.0044	± 0.0011	± 0.0032	± 0.0006
Colon	Mean	0.939	0.915	0.953	0.946	0.954	0.976
	Var	± 0.0048	± 0.0030	± 0.0029	± 0.0039	± 0.0041	± 0.0015
Leukemia_3c	Mean	1	1	0.993	1	0.986	1

Continued on next page

Dataset	Performance	PSO	GWO	WOA	SCA	SSA	ISSA
Leukemia_3c	Var	0	0	± 0.0005	0	± 0.0009	0
Leukemia_4c	Mean	0.993	1	0.98	0.986	0.993	1
	Var	± 0.0005	0	± 0.0020	± 0.0009	± 0.0005	0
Leukemia	Mean	1	1	1	0.9993	1	1
	Var	0	0	0	± 0.0005	0	0
Lung	Mean	0.992	0.967	0.983	0.988	0.976	0.992
	Var	± 0.0001	± 0.0009	± 0.0002	± 0.0004	± 0.0004	± 0.0001
MLL	Mean	1	0.972	1	1	0.986	1
	Var	0	± 0.0013	0	0	± 0.0009	0
Ovarian	Mean	1	0.974	1	0.998	1	1
	Var	0	± 0.0010	0	± 0.00004	0	0
SRBCT	Mean	1	0.964	1	1	1	1
	Var	0	± 0.0018	0	0	0	0

Table 9. Classification results of different intelligent algorithms on MLP.

Dataset	Performance	PSO	GWO	WOA	SCA	SSA	ISSA
Breast	Mean	0.92	0.92	0.957	0.903	0.925	0.98
	Var	± 0.0076	± 0.0059	± 0.0025	± 0.0055	± 0.0040	± 0.0007
CNS	Mean	0.901	0.934	0.917	0.935	0.959	0.992
	Var	± 0.0060	± 0.0044	± 0.0109	± 0.0058	± 0.0035	± 0.0006
Colon	Mean	0.916	0.961	0.937	0.922	0.906	0.968
	Var	± 0.0030	± 0.0029	± 0.0050	± 0.0052	± 0.0009	± 0.0017
Leukemia_3c	Mean	0.979	0.979	0.986	0.993	0.986	0.986
	Var	± 0.0011	± 0.0011	± 0.0009	± 0.0005	± 0.0009	± 0.0009
Leukemia_4c	Mean	0.952	0.926	0.935	0.953	0.934	0.958
	Var	± 0.0041	± 0.0023	± 0.0058	± 0.0029	± 0.0058	± 0.0013
Leukemia	Mean	1	0.993	1	1	0.986	1
	Var	0	0.0005	0	0	± 0.0009	0
Lung	Mean	0.961	0.953	0.97	0.975	0.973	0.981
	Var	± 0.0007	± 0.0010	± 0.0011	± 0.0008	± 0.0005	± 0.0005
MLL	Mean	1	0.951	1	1	1	1
	Var	0	± 0.0005	0	0	0	0
Ovarian	Mean	0.996	0.994	0.99	0.998	0.992	1
	Var	± 0.00007	± 0.0002	± 0.0004	± 0.00005	± 0.0002	0
SRBCT	Mean	0.994	1	1	1	1	1
	Var	± 0.0004	0	0	0	0	0

Figure 7 is the grouping box graph of Lung data set (the graphs of the other nine data sets are in the appendix). The abscissa is six classification methods, each classification method is a group, and each group is the box graph of the results of ten runs of six intelligent optimization algorithms. From the graph, it can be seen that the median and quartiles of ISSA are above other methods.

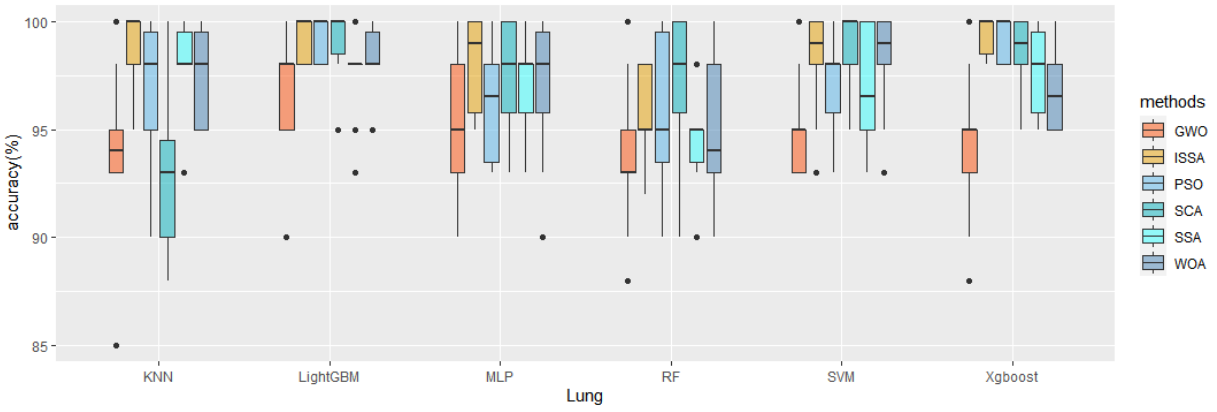

Figure 7. Boxplot comparison of different classification methods and different intelligent optimization algorithms for Lung dataset.

Figure 8 shows the number of features selected by the six intelligent optimization algorithms used in this paper for comparison on 10 datasets. This article uses six classification methods, each of which is run ten times, so the number of features in the figure is the average of 60 results. As can be seen from the figure, the number of features selected by ISSA is the least.

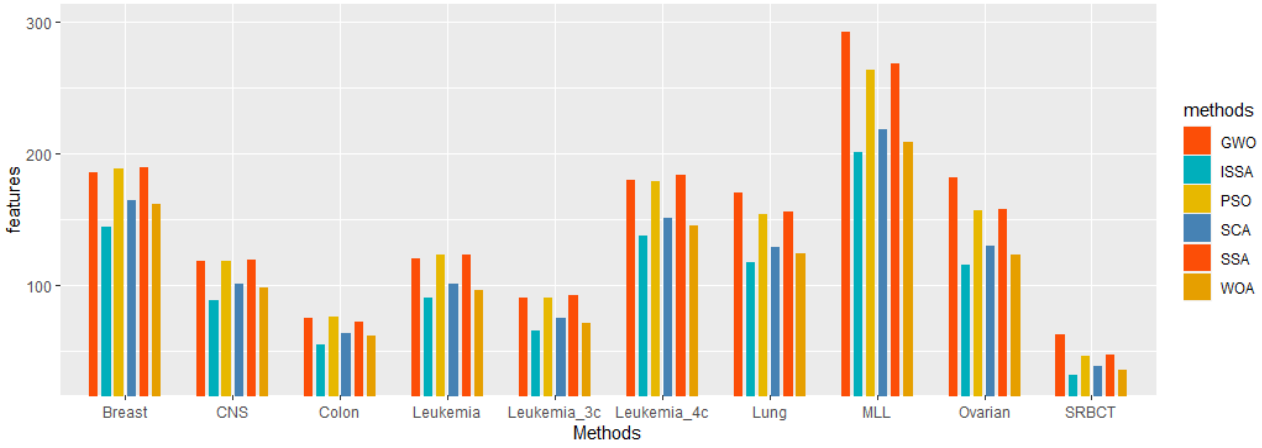

Figure 8. The average number of features selected by different intelligent optimization algorithms on ten datasets.

Table 10 shows the comparison between the framework proposed in this paper and the advanced research in recent years. The last line in the table is the work done in this paper, and the one with the highest accuracy among the six methods is selected. The top 10 rows are the performance of other papers on the dataset, and the first column is the abbreviation of the methods proposed by other papers. For specific methods, please refer to Table 3. "-" indicates that the paper did not use this dataset. In the 10 datasets, this paper has achieved equal or better accuracy, and the classification accuracy has reached more than 97.6% in all datasets, so the framework proposed in this paper has very important research significance.

Table 10. Comparison with other papers.

Methods	Breast	CNS	Colon	Leukemia_3c	Leukemia_4c	Leukemia	Lung	MLL	Ovarian	SRBCT
TLBOGSA	-	-	94.78	-	-	-	90.72	-	-	98.57
IGWO-MLP	-	-	-	-	-	-	95.64	-	-	99.14
RFACO	-	-	94	-	-	95.8	99.5	-	-	-
EAK	-	84.62	84.62	-	-	-	99.46	-	-	-
FADNE	-		83.8	-	-	92.9	98.8	-	-	93.6
ATLBO	92.53		91.57	-	-	92.71	90.78	-	-	96.04
MOGT	-		88.73	-	-	90.18	91.76	-	-	82.82
AC-MOFOA	86.53			97.66	-	-	93.97	-	-	90.72
GWO-TRIZ	-	97.38	94.13	99.86	98.84	100	97.52	99.9	100	100
IG-MBKH	-	90.34	96.47	99.44	-	100	96.12	99.72	100	100
Proposed	**100**	**99.2**	**97.6**	**100**	**100**	**100**	**99.4**	**100**	**100**	**100**

6. Conclusions

The internal relationship of cancer gene expression data sets makes cancer diagnosis full of challenges, so feature selection technology plays an important role in reducing the dimension of data and deleting irrelevant and redundant features. And because different feature selection algorithms have different advantages and disadvantages, combining different types of feature selection algorithms is a promising technique for solving feature selection problems. The two-stage framework for gene selection proposed in this paper combines embedding, filtering and wrapper to identify the optimal feature subset. This algorithm can significantly reduce the size of features while maintaining high-performance indicators. ISSA considers three factors: one is to use PWLCM chaotic mapping to increase the diversity of the initial population; second, it changes the number of leaders' choices to avoid falling into local optimization; the third is to improve the follower's update formula, which can search the optimal location faster, that is, to find the optimal subset. In the experiment, 10 high-dimensional benchmark datasets were used to test the performance of the method. These benchmark datasets are different in the number of genes, samples and categories, which is good enough to evaluate the generalization ability of the method.

At present, feature selection is a hot issue. New algorithms and new theories are all trying to solve the problem of feature selection. Feature selection by swarm intelligence optimization algorithm can help machine learning technology use the most important features, which improves the performance of learning algorithm, that is, learning speed or classification accuracy. The swarm intelligence algorithm has become more and more perfect in theory and has been proved to be a good method for solving practical optimization problems. The randomness of the algorithm can promote the diversity of solutions, avoid falling into local optimal solutions, and make it converge to a global optimal solution more quickly. However, since the algorithm is a random search algorithm, the solution of the problem and the analysis of the performance of the algorithm can only be proved by numerical experiment analysis, and the theoretical derivation is still insufficient. Therefore, swarm intelligence algorithm is an important direction of computer research and development, and will have a broad prospect in most science and engineering. In our further work, it is mixed with other intelligent algorithms to enhance its search ability, and other advanced machine learning algorithms are used in the classification part.

Acknowledgments

This work was supported by the Department Project of Jilin Province (Grant No. 20210101149JC), National Natural Science Foundation of China (Grant No. 12026430), Education Department Project of Jilin Province (Grant No. JJKH20210716KJ), and Science and Technology Department Project of Jilin Province (Grant No. 20200403182SF). The authors thank the editors for their comments and suggestions.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. A. Bashiri, M. Ghazisaeedi, R. Safdari, L. Shahmoradi, H. Ehtesham, Improving the prediction of survival in cancer patients by using machine learning techniques: experience of gene expression data: a narrative review, *Iran. J. Public Health*, **46** (2017), 165−172.
2. A. K. Shukla, P. Singh, M. Vardhan, Gene selection for cancer types classification using novel hybrid metaheuristics approach, *Swarm Evol. Comput.*, **54** (2020), 100661. https://doi.org/10.1016/j.swevo.2020.100661
3. A. Saha, S. Das, Clustering of fuzzy data and simultaneous feature selection: a model selection approach, *Fuzzy Set Syst.*, **340** (2018), 1−37. https://doi.org/10.1016/j.fss.2017.11.015
4. J. A. Cruz, D. S. Wishart, Applications of machine learning in cancer prediction and prognosis, *Cancer Inf.*, **2** (2006), 59−77. https://doi.org/10.1177/117693510600200030
5. A. K. Shukla, P. Singh, M. Vardhan, A hybrid framework for optimal feature subset selection, *J. Intell. Fuzzy Syst.*, **36** (2019), 2247−2259. https://doi.org/10.3233/JIFS-169936
6. I. Guyon, A. Elisseef, An introduction to variable and fea ture selection, *J. Mach. Learn. Res.*, **3** (2003), 1157–1182. https://doi.org/10.5555/944919.944968
7. L. C. Molina, L. Belanche, A. Nebot, Feature selection algo rithms: a survey and experimental evaluation, in *2002 IEEE International Conference on Data Mining*, (2002), 306–313. https://doi.org/10.1109/ICDM.2002.1183917
8. H. Liu, L. Yu, Toward integrating feature selection algorithms for classifcation and clustering, *IEEE Trans. Knowl. Data Eng.*, **17** (2005), 491–502. https://doi.org/10.1109/TKDE.2005.66
9. H. M. Zawbaa, E. Emary, C. Grosan, V. Snasel, Large-dimensionality small-instance set feature selection: a hybrid bio-inspired heuristic approach, *Swarm Evol. Comput.*, **42** (2018), 29–42. https://doi.org/10.1016/j.swevo.2018.02.021
10. L. Sun, X. Zhang, Y. Qian, J. Xu, S. Zhang, Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification, *Inf. Sci.*, **502** (2019), 18−41. https://doi.org/10.1016/j.ins.2019.05.072
11. A. Kumar, A. Halder, Ensemble-based active learning using fuzzy-rough approach for cancer sample classification, *Eng. Appl. Artif. Intell.*, **91** (2020), 103591. https://doi.org/10.1016/j.engappai.2020.103591

12. J. Lee, I. Choi, C. Jun, An efficient multivariate feature ranking method for gene selection in high-dimensional microarray data, *Expert Syst. Appl.*, **166** (2020), 113971. https://doi.org/10.1016/j.eswa.2020.113971
13. X. Zheng, C. Zhang, Gene selection for microarray data classification via dual latent representation learning, *Neurocomputing*, **461** (2021), 266−280. https://doi.org/10.1016/j.neucom.2021.07.047
14. L. Li, W. Ching, Z. Liu, Robust biomarker screening from gene expression data by stable machine learning-recursive feature elimination methods, *Comput. Biol. Chem.*, **100** (2022), 107747. https://doi.org/10.1016/j.compbiolchem.2022.107747
15. H. Wang, L. Tan, B. Niu, Feature selection for classification of microarray gene expression cancers using Bacterial Colony Optimization with multi-dimensional population, *Swarm Evol. Comput.*, **48** (2019), 172−181. https://doi.org/10.1016/j.swevo.2019.04.004
16. C. Shen, K. Zhang, Two-stage improved Grey Wolf optimization algorithm for feature selection on high-dimensional classification, *Complex Intell. Syst.*, **8** (2022), 1−21. https://doi.org/10.1007/s40747-021-00452-4
17. C. Qu, L. Zhang, J. Li, F. Deng, Y. Tang, X. Zeng, et al., Improving feature selection performance for classification of gene expression data using Harris Hawks optimizer with variable neighborhood learning, *Briefings Bioinf.*, **22** (2021). https://doi.org/10.1093/bib/bbab097
18. A. Dabba, A. Tari, S. Meftali, R. Mokhtari, Gene selection and classification of microarray data method based on mutual information and moth flame algorithm, *Expert Syst. Appl.*, **166** (2020), 114012. https://doi.org/10.1016/j.eswa.2020.114012
19. L. Sun, X. Kong, J. Xu, Z. Xue, R. Zhai, S. Zhang, A hybrid gene selection method based on reliefF and ant colony optimization algorithm for tumor classification, *Sci. Rep.*, **9** (2019), 8978. https://doi.org/10.1038/s41598-019-45223-x
20. Uzma, F. Al-Obeidat, A. Tubaishat, B. Shah, Z. Halim, Gene encoder: a feature selection technique through unsupervised deep learning-based clustering for large gene expression data, *Neural Comput. Appl.*, **34** (2020), 8309−8331. https://doi.org/10.1007/s00521-020-05101-4
21. S. Mirjalili, A. H. Gandomi, S. Z. Mirjalili, S. Saremi, H. Faris, S. M. Mirjalili, Salp swarm algorithm: a bio-inspired optimizer for engineering design problems, *Adv. Eng. Software*, **114** (2017), 163−191. https://doi.org/10.1016/j.advengsoft.2017.07.002
22. J. Kennedy, R. Eberhart, Particle swarm optimization, in *Proceedings of ICNN'95 - International Conference on Neural Networks*, 1995. https://doi.org/10.1109/ICNN.1995.488968
23. S. Mirjalili, S. M. Mirjalili, A. Lewis, Grey wolf optimizer, *Adv. Eng. Software*, **69** (2014), 46–61. https://doi.org/10.1016/j.advengsoft.2013.12.007
24. S. Mirjalili, A. Lewis, The whale optimization algorithm, *Adv. Eng. Software*, **95** (2016), 51−67. https://doi.org/10.1016/j.advengsoft.2016.01.008
25. S. Mirjalili, SCA: a sine cosine algorithm for solving optimization problems, *Knowledge-Based Syst.*, **96** (2016), 120−133. https://doi.org/10.1016/j.knosys.2015.12.022
26. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, *Bmc Bioinf.*, **9** (2008), 559. https://doi.org/10.1186/1471-2105-9-559
27. B. Zhang, S. Horvath, A general framework for weighted gene co-expression network analysis, *Stat. Appl. Genet. Mol. Biol.*, **4** (200), 17. https://doi.org/10.2202/1544-6115.1128

28. H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Trans. Pattern Anal. Mach. Intell.*, **27** (2005), 1226−1238. https://doi.org/10.1109/TPAMI.2005.159
29. Available from: https://csse.szu.edu.cn/staff/zhuzx/Datasets.html.
30. A. K. Shukla, P. Singh, M. Vardhan, An adaptive inertia weight teaching-learning-based optimization algorithm and its applications, *Appl. Math. Modell.*, **77** (2020), 309−326. https://doi.org/10.1016/j.apm.2019.07.046
31. M. Rostami, S. Forouzandeh, K. Berahmand, M. Soltani, M. Shahsavari, M. Oussalah, Gene selection for microarray data classification via multi-objective graph theoretic-based method, *Artif. Intell. Med.*, **123** (2021), 102228. https://doi.org/10.1016/j.artmed.2021.102228
32. B. Nouri-Moghaddam, M. Ghazanfari, M. Fathian, A novel bio-inspired hybrid multi-filter wrapper gene selection method with ensemble classifier for microarray data, *Neural Comput. Appl.*, **2021** (2021), 1−31. https://doi.org/10.1007/s00521-021-06459-9
33. O. A. Alomari, S. N. Makhadmeh, M. A. Al-Betar, Z. A. A. Alyasseri, I. A. Doush, A. K. Abasi, et al., Gene selection for microarray data classification based on Grey Wolf Optimizer enhanced with TRIZ-inspired operators, *Knowledge-Based Syst.*, **223** (2021), 107034. https://doi.org/10.1016/j.knosys.2021.107034
34. G. Zhang, J. Hou, J. Wang, C. Yan, J. Luo, Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm, *Interdiscip. Sci. Comput. Life Sci.*, **12** (2020), 288−301. https://doi.org/10.1007/s12539-020-00372-w

Appendix

Figures A1 and A2 show WGCNA results. Figure A3 shows the random forest results of the other eight datasets. Tables A1−A4 show the classification results of WGCNA, RF and mRMR feature selection. Tables A5−A8 show the accuracy of different intelligent algorithms in four other classification methods. Figures A4−A11 show the box diagram comparison of different classification methods and different intelligent optimization algorithms for the other eight data sets.

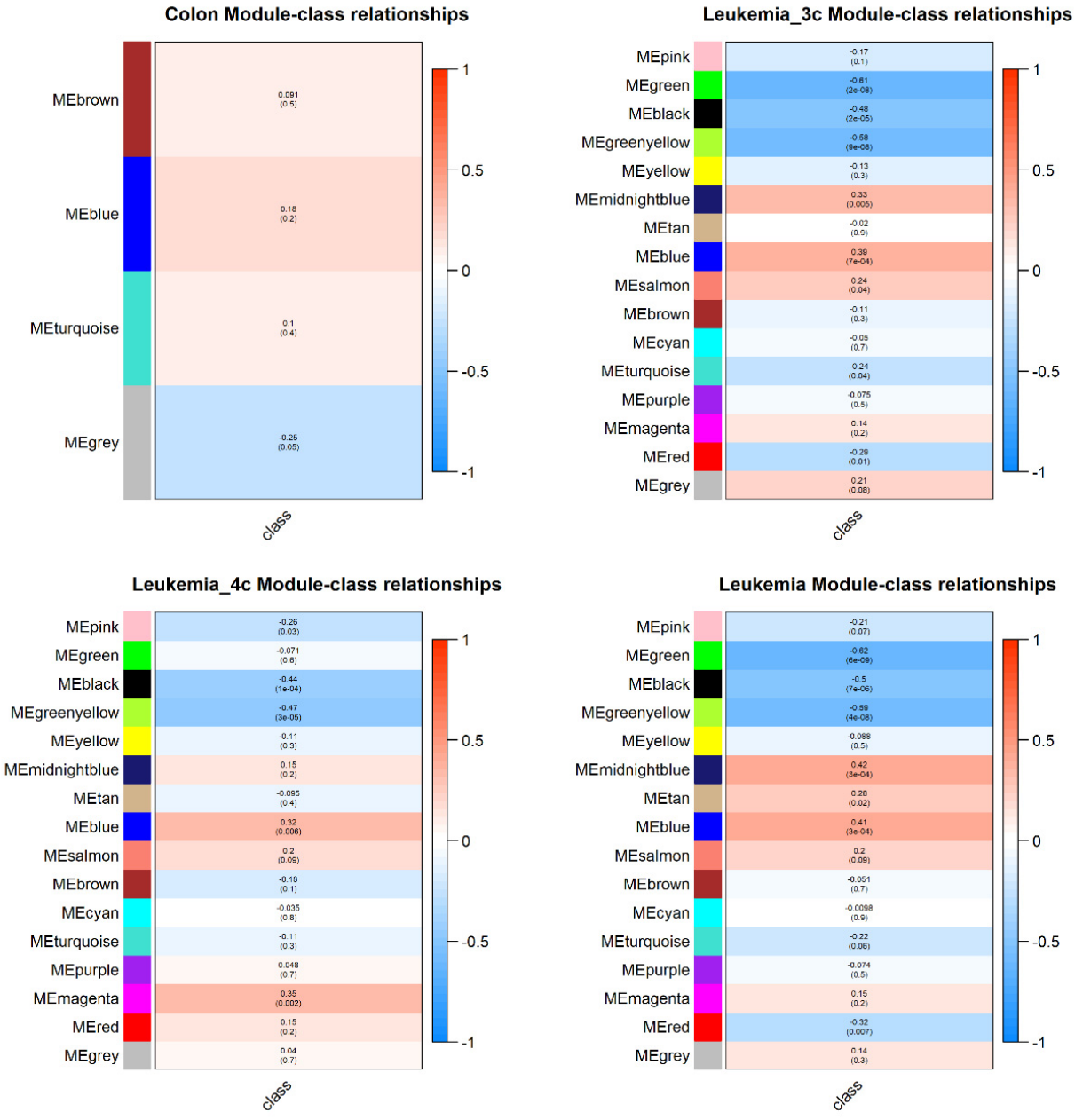

Figure A1. Correlation of modules with categories in Colon, Leukemia_3c, Leukemia_4c, and Leukemia datasets WGCNA.

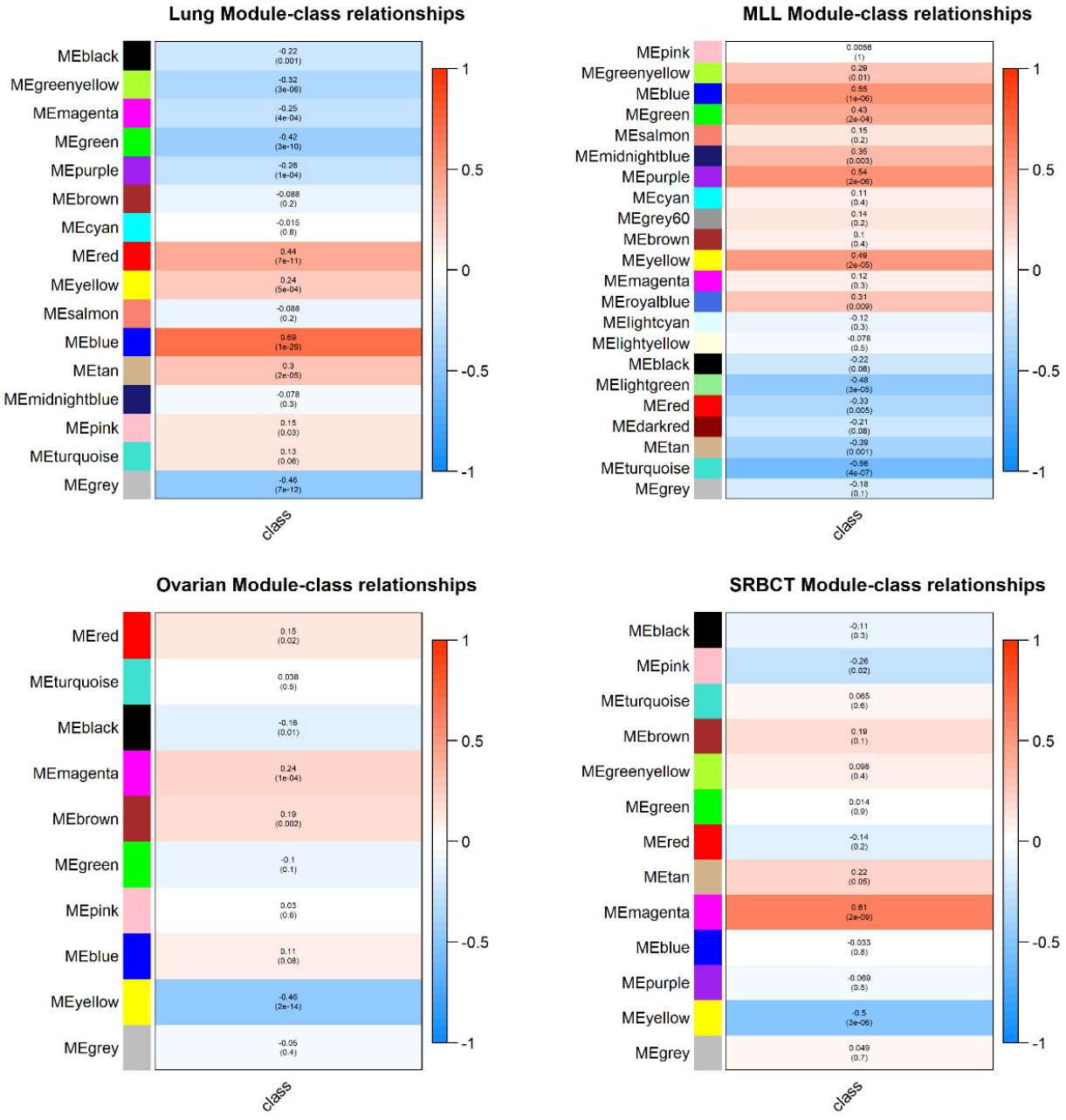

Figure A2. Correlation of modules with categories in Lung, MLL, Ovarian and SRBCT datasets WGCNA.

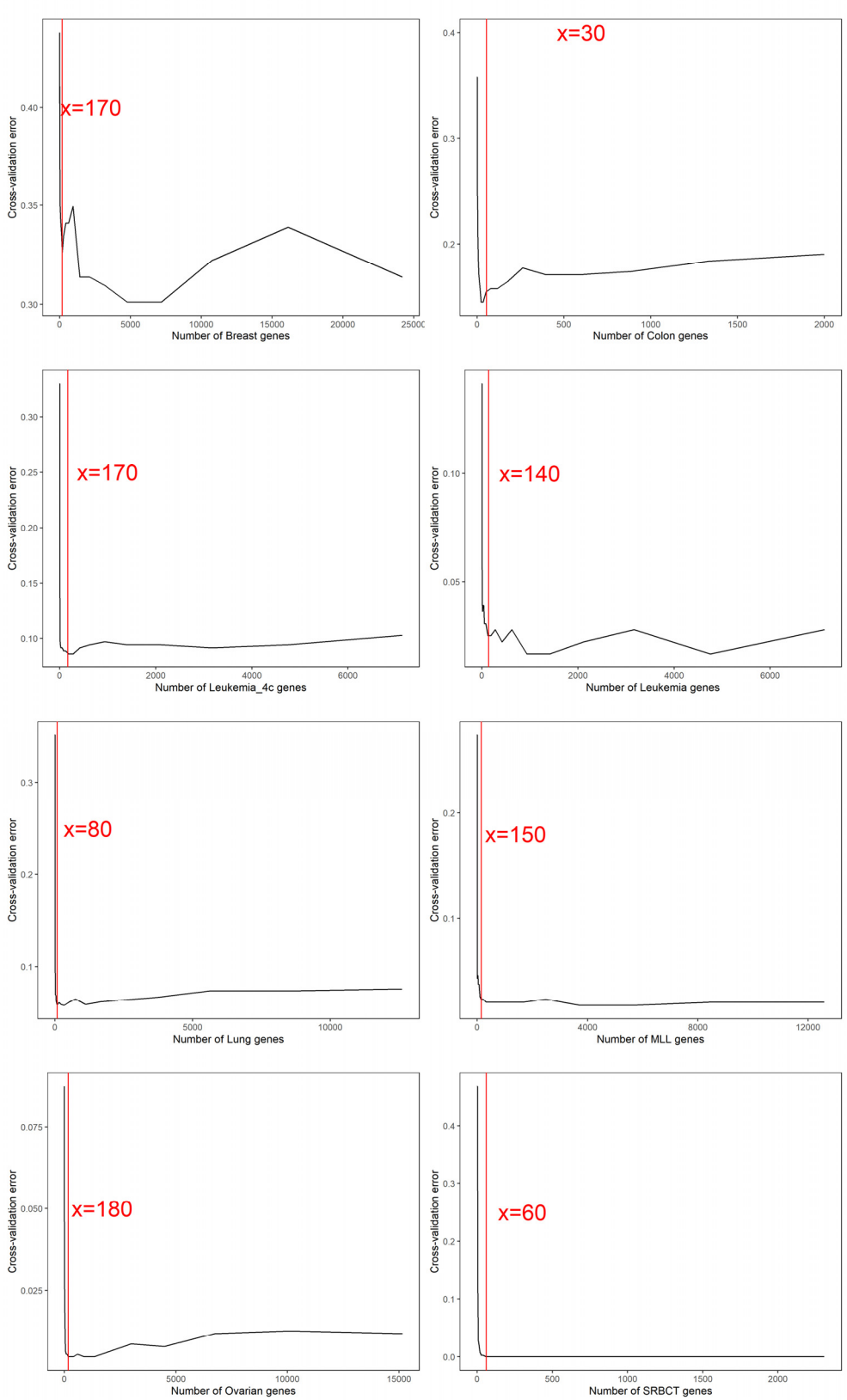

Figure A3. Error rate and number of features of random forest method for ten-fold cross-validation.

Table A1. Classification accuracy of RF method.

Dataset	All	WGCNA	RF	mRMR	Union
Breast	0.63	0.64	0.53	0.79	0.89
CNS	0.58	0.58	0.92	0.75	0.83
Colon	0.85	0.85	0.77	0.85	0.77
Leukemia_3c	0.8	0.67	0.93	0.93	0.87
Leukemia_4c	0.67	0.67	0.8	0.87	0.80
Leukemia	0.87	0.93	0.87	0.93	0.93
Lung	0.85	0.95	0.90	0.95	0.93
MLL	0.93	0.86	0.93	0.93	0.93
Ovarian	0.96	0.73	0.96	0.96	0.98
SRBCT	0.94	0.94	0.94	1.00	1.00

Table A2. Classification accuracy of SVM method.

Dataset	All	WGCNA	RF	mRMR	Union
Breast	0.47	0.69	0.84	0.58	0.53
CNS	0.58	0.58	0.75	0.58	0.67
Colon	0.92	0.77	0.92	0.85	0.77
Leukemia_3c	0.73	0.73	0.93	0.87	0.93
Leukemia_4c	0.53	0.73	0.73	0.93	0.93
Leukemia	0.73	0.93	0.87	0.93	0.93
Lung	0.83	0.83	0.88	0.93	0.88
MLL	0.79	0.86	0.93	0.93	0.93
Ovarian	0.98	0.78	1.00	0.98	0.98
SRBCT	0.94	0.88	0.94	0.94	1.00

Table A3. Classification accuracy of MLP method.

Dataset	All	WGCNA	RF	mRMR	Union
Breast	0.58	0.64	0.84	0.79	0.84
CNS	0.58	0.58	0.67	0.75	0.58
Colon	0.92	0.92	0.85	0.92	0.77
Leukemia_3c	0.8	0.87	0.93	0.80	0.93
Leukemia_4c	0.67	0.73	0.67	0.93	0.93
Leukemia	0.8	0.8	0.80	0.93	0.87
Lung	0.90	0.80	0.85	0.93	0.88
MLL	0.79	0.93	0.93	0.71	0.93
Ovarian	0.88	0.73	1.00	0.98	0.98
SRBCT	0.94	0.88	0.94	1.00	1.00

Table A4. Classification accuracy of KNN method.

Dataset	All	WGCNA	RF	mRMR	Union
Breast	0.58	0.63	0.84	0.84	0.79
CNS	0.58	0.42	0.92	0.67	0.58
Colon	0.77	0.62	0.92	0.92	0.77
Leukemia_3c	0.80	0.80	0.93	0.87	0.93
Leukemia_4c	0.8	0.73	0.87	0.47	0.93
Leukemia	0.58	0.87	0.87	0.93	0.93
Lung	0.83	0.90	0.93	0.93	0.98
MLL	0.79	0.93	0.86	0.93	0.86
Ovarian	0.98	0.67	0.62	0.96	0.96
SRBCT	0.94	0.94	0.94	0.94	0.94

Table A5. Classification results of different intelligent algorithms on RF.

Dataset	Performance	PSO	GWO	WOA	SCA	SSA	ISSA
Breast	Mean	0.842	0.847	0.862	0.809	0.852	0.938
	Var	± 0.0044	± 0.0062	± 0.0044	± 0.0081	± 0.060	± 0.0007
CNS	Mean	0.774	0.841	0.9	0.809	0.86	0.936
	Var	± 0.0188	± 0.0040	± 0.0106	± 0.0049	± 0.0110	± 0.0011
Colon	Mean	0.894	0.884	0.914	0.915	0.862	0.952
	Var	± 0.0039	± 0.0054	± 0.0018	± 0.0030	± 0.0090	± 0.0017
Leukemia_3c	Mean	0.966	0.987	0.979	0.993	0.953	0.993
	Var	± 0.0022	± 0.0017	± 0.0011	± 0.0005	± 0.0038	± 0.0005
Leukemia_4c	Mean	0.927	0.933	0.899	0.91	0.913	0.967
	Var	± 0.0032	± 0.0028	± 0.0071	± 0.0038	± 0.0038	± 0.0012
Leukemia	Mean	0.972	0.993	0.986	0.986	0.986	0.993
	Var	± 0.0013	± 0.0011	± 0.0009	± 0.0009	± 0.0009	± 0.0005
Lung	Mean	0.992	0.967	0.983	0.988	0.976	0.992
	Var	± 0.0015	± 0.0012	± 0.00113	± 0.0010	± 0.0006	± 0.0004
MLL	Mean	1	0.979	0.986	0.986	0.993	1
	Var	0	± 0.0022	± 0.0009	± 0.0009	± 0.0005	0
Ovarian	Mean	0.988	0.992	0.996	0.996	0.986	1
	Var	± 0.0004	± 0.0004	± 0.00007	± 0.00007	± 0.0004	0
SRBCT	Mean	1	0.988	0.994	1	0.994	1
	Var	0	± 0.0006	± 0.0004	0	± 0.0004	0

Table A6. Classification results of different intelligent algorithms on SVM.

Dataset	Performance	PSO	GWO	WOA	SCA	SSA	ISSA
Breast	Mean	0.793	0.815	0.846	0.851	0.783	0.897
	Var	± 0.0081	± 0.0029	± 0.0049	± 0.0023	± 0.0058	± 0.0018
CNS	Mean	0.9	0.817	0.817	0.901	0.876	0.944

Continued on next page

Dataset	Performance	PSO	GWO	WOA	SCA	SSA	ISSA
CNS	Var	± 0.0076	± 0.0089	± 0.0169	± 0.0060	± 0.0142	± 0.0015
Colon	Mean	0.899	0.93	0.878	0.93	0.921	0.968
	Var	± 0.0038	± 0.0058	± 0.0040	± 0.0058	± 0.0080	± 0.0017
Leukemia_3c	Mean	1	1	1	1	0.979	1
	Var	0	0	0	0	± 0.0011	0
Leukemia_4c	Mean	0.966	0.901	0.94	0.966	0.913	0.972
	Var	± 0.0042	± 0.0040	± 0.0033	± 0.0042	± 0.0059	± 0.0013
Leukemia	Mean	0.986	1	0.973	0.993	0.988	1
	Var	± 0.0009	0	± 0.0042	± 0.0005	± 0.0007	0
Lung	Mean	0.973	0.95	0.982	0.986	0.971	0.982
	Var	± 0.0005	± 0.0006	± 0.0006	± 0.0004	± 0.0008	± 0.0006
MLL	Mean	1	0.993	1	1	1	1
	Var	0	± 0.0009	0	0	0	0
Ovarian	Mean	0.996	0.98	1	0.988	0.989	0.998
	Var	± 0.0002	± 0.0002	0	± 0.0014	± 0.0006	± 0.00004
SRBCT	Mean	1	0.982	1	1	1	1
	Var	0	± 0.0008	0	0	0	0

Table A7. Classification results of different intelligent algorithms on XGBoost.

Dataset	Performance	PSO	GWO	WOA	SCA	SSA	ISSA
Breast	Mean	0.968	0.942	0.942	0.959	0.931	0.975
	Var	± 0.0021	± 0.0041	± 0.0076	± 0.0011	± 0.0045	± 0.0007
CNS	Mean	0.967	0.951	0.951	0.941	0.943	0.992
	Var	± 0.0034	± 0.0064	± 0.0034	± 0.0095	± 0.0062	± 0.0006
Colon	Mean	0.953	0.969	0.923	0.924	0.915	0.992
	Var	± 0.0029	± 0.0055	± 0.0064	± 0.0050	± 0.0030	± 0.0006
Leukemia_3c	Mean	0.986	0.985	0.993	1	0.96	0.986
	Var	± 0.0009	± 0.0010	± 0.0005	0	± 0.0051	± 0.0009
Leukemia_4c	Mean	0.946	1	0.972	0.993	0.993	0.993
	Var	± 0.0098	0	± 0.0013	± 0.0005	± 0.0005	± 0.0005
Leukemia	Mean	1	1	1	1	1	1
	Var	0	0	0	0	0	0
Lung	Mean	0.992	0.942	0.967	0.984	0.977	0.994
	Var	± 0.0001	± 0.0012	± 0.0004	± 0.0004	± 0.0004	± 0.00009
MLL	Mean	0.993	0.951	1	1	0.993	1
	Var	± 0.0005	± 0.0022	0	0	± 0.0005	0
Ovarian	Mean	1	0.98	1	1	1	1
	Var	0	± 0.0006	0	0	0	0
SRBCT	Mean	1	0.927	1	1	1	1
	Var	0	± 0.0058	0	0	0	0

Table A8. Classification results of different intelligent algorithms on KNN.

Dataset	Performance	PSO	GWO	WOA	SCA	SSA	ISSA
Breast	Mean	0.979	0.947	0.969	0.793	0.985	1
	Var	± 0.0014	± 0.0032	± 0.0014	± 0.0028	± 0.0006	0
CNS	Mean	0.901	0.943	0.95	0.942	0.968	0.985
	Var	± 0.0060	± 0.0032	± 0.0051	± 0.0048	± 0.0017	± 0.0011
Colon	Mean	0.968	0.938	0.976	0.848	0.93	0.968
	Var	± 0.0017	± 0.0036	± 0.0015	± 0.0052	± 0.0031	± 0.0017
Leukemia_3c	Mean	1	1	1	1	1	1
	Var	0	0	0	0	0	0
Leukemia_4c	Mean	0.945	0.946	0.96	0.973	0.947	0.979
	Var	± 0.0017	± 0.0047	± 0.0031	± 0.0022	± 0.0036	± 0.0017
Leukemia	Mean	0.993	1	1	0.972	1	1
	Var	± 0.0005	0	0	± 0.0013	0	0
Lung	Mean	0.967	0.94	0.974	0.927	0.981	0.986
	Var	± 0.0011	± 0.0015	± 0.0005	± 0.0012	± 0.0004	± 0.0004
MLL	Mean	0.993	0.965	1	1	1	1
	Var	± 0.0022	± 0.0035	0	0	0	0
Ovarian	Mean	0.996	0.98	0.998	0.984	0.987	0.99
	Var	± 0.00009	± 0.0004	± 0.00005	± 0.0005	± 0.0005	± 0.0001
SRBCT	Mean	1	0.994	1	0.994	1	1
	Var	0	± 0.0004	0	± 0.0004	0	0

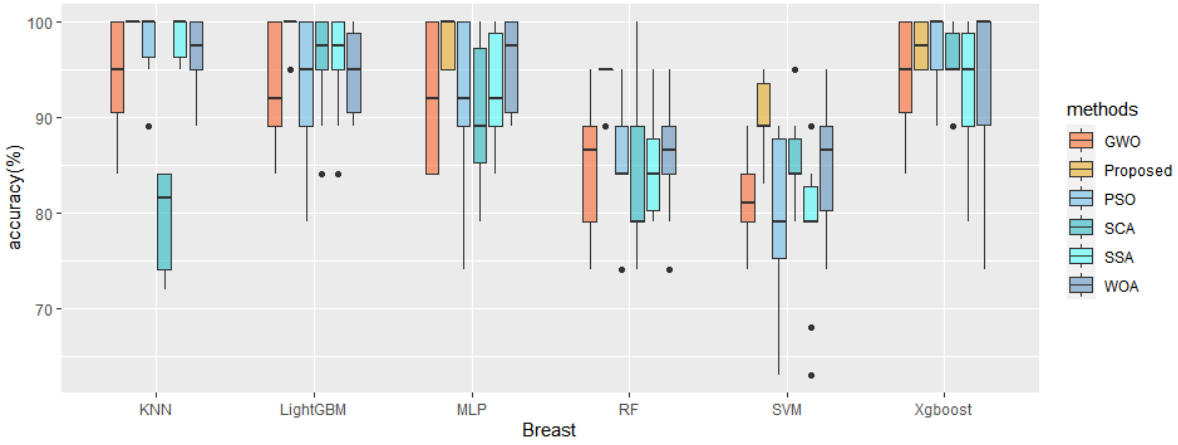

Figure A4. Boxplot comparison of different classification methods and different intelligent optimization algorithms for Breast dataset.

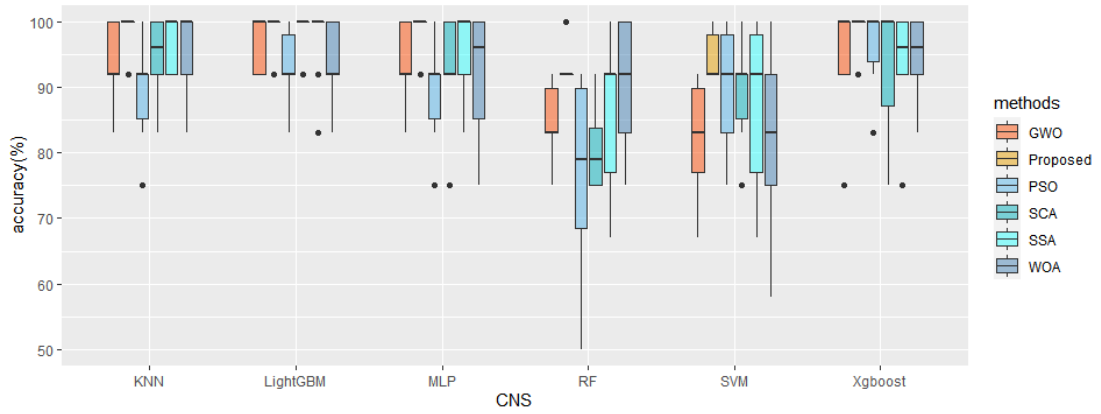

Figure A5. Boxplot comparison of different classification methods and different intelligent optimization algorithms for CNS dataset.

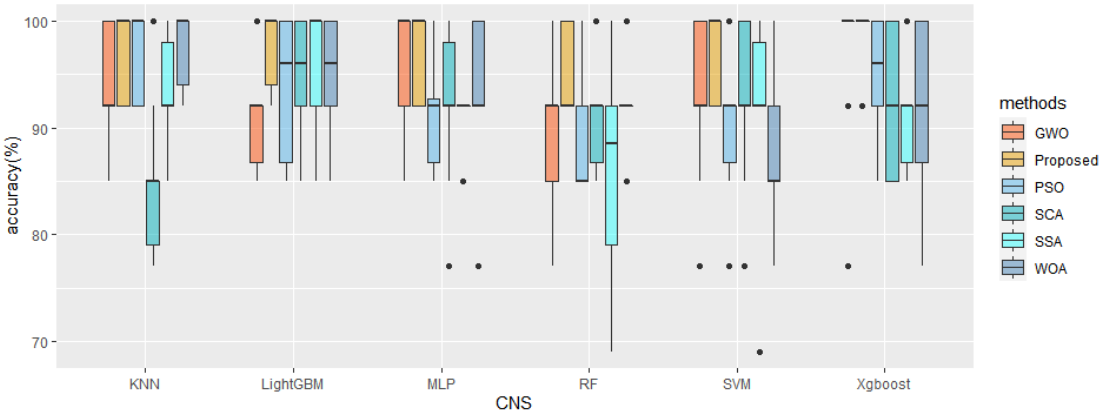

Figure A6. Boxplot comparison of different classification methods and different intelligent optimization algorithms for Colon dataset.

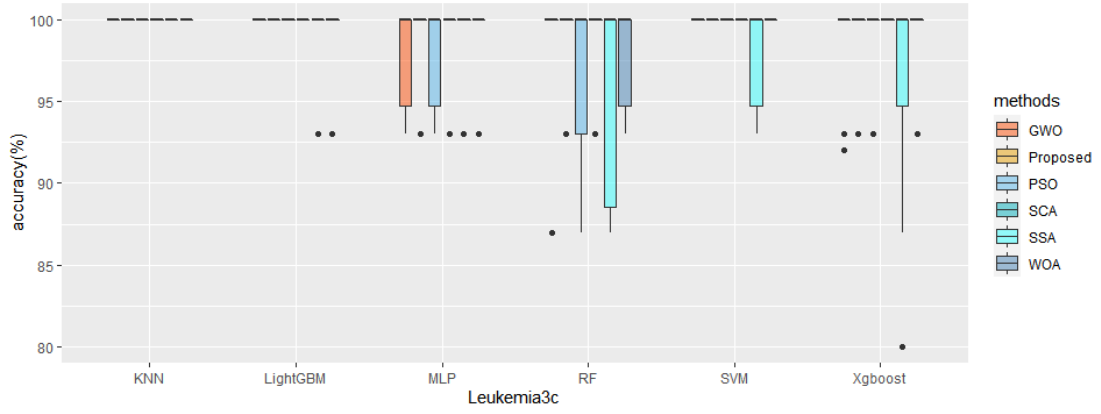

Figure A7. Boxplot comparison of different classification methods and different intelligent optimization algorithms for Leukemia_3c dataset.

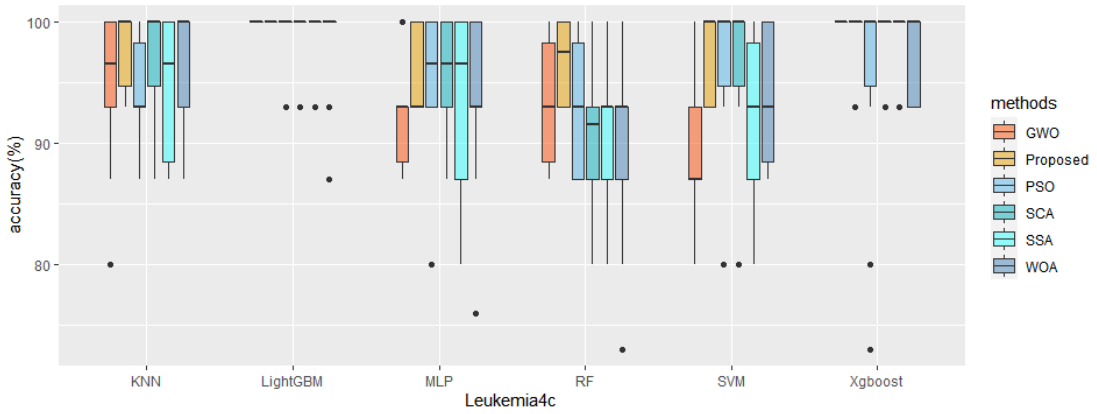

Figure A8. Boxplot comparison of different classification methods and different intelligent optimization algorithms for Leukemia_4c dataset.

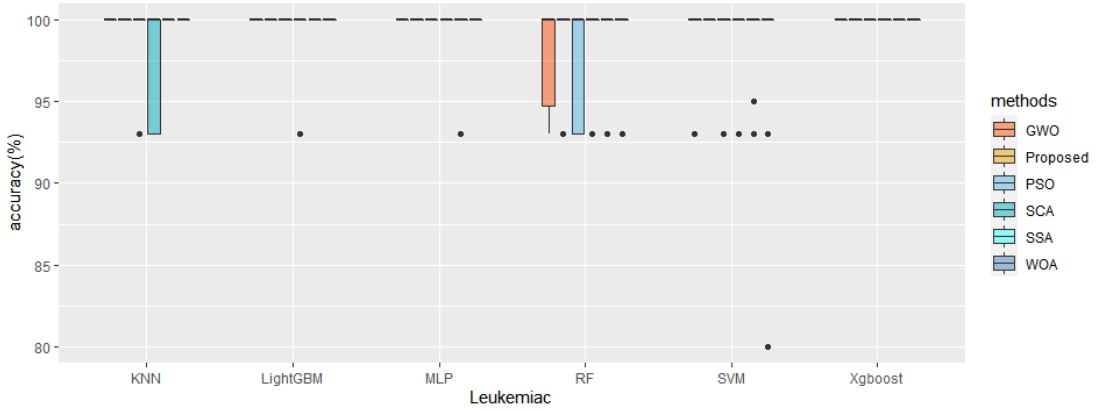

Figure A9. Boxplot comparison of different classification methods and different intelligent optimization algorithms for Leukemia dataset.

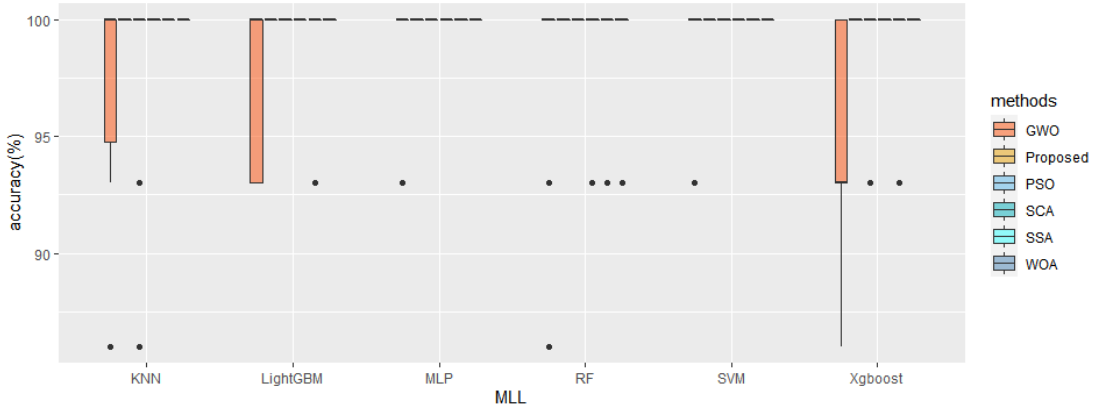

Figure A10. Boxplot comparison of different classification methods and different intelligent optimization algorithms for MLL dataset.

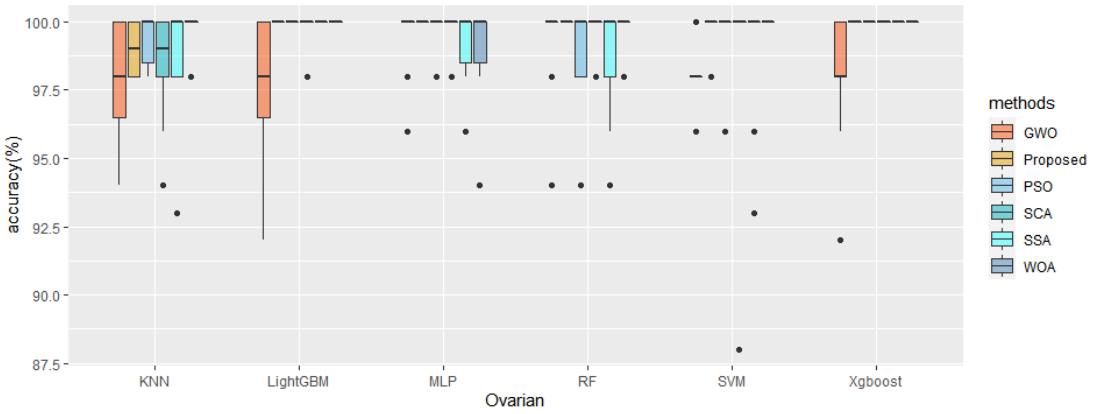

Figure A11. Boxplot comparison of different classification methods and different intelligent optimization algorithms for Ovarian dataset.

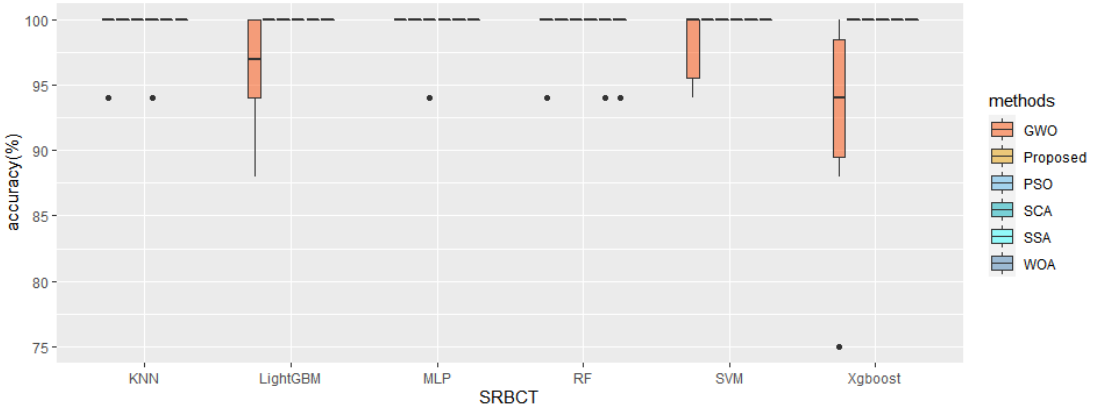

Figure A12. Boxplot comparison of different classification methods and different intelligent optimization algorithms for SRBCT dataset.