MGRFE: multilayer recursive feature elimination based on an embedded genetic algorithm for cancer classification

Cheng Peng, Xinyu Wu, Wen Yuan, Xinran Zhang, Yu Zhang, and Ying Li

Abstract—Microarray gene expression data have become a topic of great interest for cancer classification and for further research in the field of bioinformatics. Nonetheless, due to the "large p, small n" paradigm of limited biosamples and high-dimensional data, gene selection is becoming a demanding task, which is aimed at selecting a minimal number of discriminatory genes associated closely with a phenotype. Feature or gene selection is still a challenging problem owing to its nondeterministic polynomial time complexity and thus most of the existing feature selection algorithms utilize heuristic rules. A multilayer recursive feature elimination method based on an embedded integer-coded genetic algorithm, MGRFE, is proposed here, which is aimed at selecting the gene combination with minimal size and maximal information. On the basis of 19 benchmark microarray datasets including multiclass and imbalanced datasets, MGRFE outperforms state-of-the-art feature selection algorithms with better cancer classification accuracy and a smaller selected gene number. MGRFE could be regarded as a promising feature selection method for high-dimensional datasets especially gene expression data. Moreover, the genes selected by MGRFE have close biological relevance to cancer phenotypes. The source code of our proposed algorithm and all the 19 datasets used in this paper are available at https://github.com/Pengeace/MGRFE-GaRFE.

Index Terms—Gene selection, Genetic algorithm, Recursive feature elimination, Microarray data, Cancer classification.

1 INTRODUCTION

ONE chief challenge in bioinformatics is the "large p small n" paradigm [1], on account of ever-increasing high-dimensional data and limited available experimental samples. In particular, for gene expression data, the sample number is distinctively small compared with several thousand to tens of thousands of genes. For the analysis of high-dimensional data, feature selection is essential, which is designed to remove irrelevant and redundant features, thus cutting down the dimensionality and improving the predictive performance and model interpretability. On the other hand, due to its nondeterministic polynomial (NP) time complexity, feature selection is still a challenging and extensively studied problem in the machine learning and data mining fields. As for the field of bioinformatics, there are numerous high-dimensional biological data in sequence analysis, microarray analysis, and spectral analysis. This situation makes feature selection more important and challenging. On the basis of the process of choosing features for classification, feature selection methods can be roughly subdivided into three categories: filter, wrapper, and hybrid techniques [2].

Filter algorithms generally evaluate features according to the inherent characteristic of a dataset, then rank all the features and preserve only an optimal subset of the original features. Up to now, lots of filter algorithms have been designed, such as the methods based on the *t*-test [3], χ^2 test [4], mutual information [5], maximal information coefficient (MIC) [6], and signal-to-noise ratio [7]. The *t*-test is the frequently used and efficient statistical approach to detecting differentially expressed genes in microarray analysis [8], [9], [10], [11], [12], [13], [14]. MIC is an information theory-based measurement for capturing a wide range of associations, which has shown excellent performance on detecting novel associations in large datasets [15]. The recent study of McTwo [16], which is based on MIC for selection of a gene subset in a microarray data, has outperformed most of existing algorithms. In addition, MIC may offer more convenience

- *C. Peng, X. Wu, W. Yuan, X. Zhang, Y. Zhang, and Y. Li are with the College of Computer Science and Technology, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China.*
- *Y. Li is the correspondence author. Email: liying@jlu.edu.cn.*

in dealing with multiclass datasets. Hence, the *t*-test and MIC are practicable and qualified approaches for selecting statistically significant discriminative genes and thus are used in the feature preprocessing stage in this study to generate the candidate gene sets. Because there is no classification algorithm involved in the filter algorithm, its computational speed is high and suitable for large datasets. On the other hand, the filter techniques for gene feature selection also have some limitations. First, filter methods are likely to add redundant features into the chosen subsets, which will lead to inaccessibility of optimal results. Second, the mutual information between features is ignored for the implicit orthogonality assumption of features [17]. Third, the features top-ranked by a filter algorithm are not always the best features for classification [16].

Wrapper algorithms usually employ classification models and contain heuristic rules to select feature subsets guided by the classification performance on the feature subsets being used. These methods are usually superior to filter algorithms but more time-consuming. A variety of wrapper algorithms have emerged involving randomized hill climbing [18], regularized random forest (RRF) [19], particle swarm optimization (PSO) [20], [21], and genetic algorithm (GA) [22]. With the rapid development of heuristic rules and evolutionary strategies commonly present in wrapper techniques, various swarm intelligence algorithms have been applied to the optimization of feature selection. Kar *et al.* have proposed a particle swarm optimization method based on adaptive K-nearest neighborhood (KNN) to identify a minimum meaningful gene subset [23]. Moosa *et al.* have presented a modified artificial bee colony algorithm (ABC) to select a minimum number of genes with high predictive accuracy for cancer classification [24]. Oreski *et al.* have designed a hybrid GA with neural networks to identify an optimal feature subset with high classification accuracy and scalability for credit risk assessment [25]. Jung and Zscheischler have described a guided hybrid GA to minimize the number of cost function evaluations [26]. Nevertheless, all these feature selection methods based on swarm intelligence algorithms use the binary encoding method and lack an explicit reduction in the feature number. The feature number only changes in the randomized evolution operators like mutation and crossover. Thus, these methods lack the precise control over the gene features in individuals and can not explicitly remove genes to decrease the feature number. Meanwhile, it has been verified that only a minimal number of informative genes is enough for effective diagnosis of different phenotypes in microarray gene datasets [16], [17], [24], [27]. The feature selection using binary encoding has three main shortcomings in finding an optimal gene combination: (1) The fixed chromosome length for the encoding length must be equal to the gene range to represent all the genes. This arrangement can result in impossibility of the explicit reduction in the gene number and unnecessary space occupation when there are only several 1s among lots of 0s. (2) There are different numbers of actual existing genes in different individuals. Because there are different numbers of 1s, the actual number of genes varies among individuals and cannot be controlled precisely. (3) The convergence speed is usually low and the time cost is high to generate the minimal informative gene combination. The sizes of the optimal gene combinations in most of datasets are usually below 10. The evolution-based feature selection algorithms using binary encoding lack of an explicit feature reduction mechanism, which results in low probability and high time cost to generate the optimal minimal gene combination among the several thousand to tens of thousands of genes in each dataset. Recursive feature elimination (RFE) is a popular strategy that yields an explicit recursive feature reduction by recursively removing features with the least weights [17], [28], [29], [30].

Hybrid algorithms are the combination of filter and wrapper strategies [2]. First, the filter algorithms are applied to remove irrelevant features and narrow the search space. Second, the wrapper algorithms are performed on the pre-selected subsets to accomplish optimal feature selection. Hybrid algorithms can take advantage of both filter and wrapper techniques.

A multilayer recursive feature elimination method with an embedded integer-coded genetic algorithm, MGRFE, is proposed here, which can be categorized into a hybrid algorithm. On the one hand, MGRFE uses the *t*-test and MIC to obviously reduce the feature range and generate a candidate feature set. On the other hand, MGRFE combines the advantages of both evolution calculation of GA and the explicit feature elimination of RFE to achieve the minimum discriminative gene subset with optimal classification ability. To validate the performance of the proposed method, we performed comprehensive experiments on 19 benchmark gene expression datasets including multiclass and imbalanced datasets and compared the performance with other various feature selection methods. The comparison results show that our method outperforms most of state-of-the-art feature selection algorithms for selecting a smaller gene subset but yielding the same or higher classification accuracy. By validation experiments on independent datasets, the generalization ability of selected feature subset by MGRFE is verified. Furthermore, the specific biomedical relevance of the selected genes to the related cancer

phenotypes has also been confirmed by text mining in Pubmed. The whole work flow of this study is presented in Fig. 1.

2 Materials and Methods

2.1 Materials

This study involves 19 benchmark microarrays including binary, multiclass, balanced, and imbalanced datasets, which are subdivided into two large datasets. Dataset One consists of the 17 binary classification datasets used in ref. [16], which includes diffuse large B-cell lymphoma (DLBCL) [31], Prostate (Pros) [32], acute lymphoblastic leukemia (ALL; subdivided into four subtypes based on different phenotypes) [33], central nervous system embryonal tumor (CNS) [34], Lymphoma (Lym) [35], Adenoma (Adeno) [36], Colon [37], Leukaemia (Leuk) [38], Myeloma (Mye) [39], Gastric (Gas) [40], and Gastric1/Gastric2 (Gas1/Gas2) cancer [41] as well as type 1 diabetes (T1D) [42], and Stroke [43]. Among them, DLBCL, Colon, Leukaemia, Myeloma, ALL1-4, and CNS datasets are imbalanced. Dataset Two is composed of the three typical benchmark microarray datasets used in ref. [23], including two multiclass datasets of small-round blue-cell tumor (SRBCT) [44] and mixed lineage leukemia (MLL) [45] and one binary dataset of acute lymphoblastic leukemia and acute myeloid (ALL_AML). Many previous experiments are conducted on these three datasets [44], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60]. The performance comparison between these methods and MGRFE will be provided in the Results section. It should be mentioned that the widely used benchmark Leukaemia was tested in both refs. [16] and [23] named as Leuk and ALL_AML respectively, but they are same actually. The information of the 19 datasets in the two large datasets is given in Tables 1 and 2 in Supplementary Material. All these datasets can be downloaded directly from https://github.com/Pengeace/MGRFE-GaRFE.

2.2 Classification performance measurements

On the 17 binary classification datasets, we employed six widely used measurements to compare the performance: Accuracy (Acc), Sensitivity (Sn), Specificity (Sp), Average accuracy (Avc), Matthews Correlation Coefficient(MCC), and AUC. AUC is the area under the receiver operating characteristic (ROC) curve, and the formulas of the other five measurements are presented in Equation (1). In Equation (1), P and N represent the numbers of positive and negative samples; TP and TN mean the number of correctly predicted positive and negative samples; and FP and FN denote the wrongly predicted positive and negative samples, respectively. For the two multiclass datasets, for consistency and convenience, only Acc is used.

$$\begin{aligned} Sn &= \frac{TP}{TP+FN}, \quad Sp = \frac{TN}{TN+FP}, \\ Acc &= \frac{TP+TN}{P+N}, \quad Avc = \frac{Sn+Sp}{2}, \\ MCC &= \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{aligned} \tag{1}$$

2.3 Method

As shown in Fig. 2, the proposed MGRFE is divided into three stages, which are (1) Search space reduction; (2) Precise wrapper search and (3) Multiple *k*-fold cross-validation (CV). MGRFE is a multilayer iterative feature selection method with GA-RFE acting as the feature selection unit in every layer. GA-RFE is a recursive feature elimination process including embedded GA. The Pseudocodes 1, 2 and 3 in Supplementary Material provide the pseudocodes of the processes MGRFE, GA-RFE and embedded GA.

2.3.1 Stage 1: Search space reduction

At Stage 1, two filter methods, the *t*-test and MIC, are used to decrease the gene range and offer a candidate gene set for later precise wrapper search stage. First, we perform the *t*-test on all genes and subject them to ascending sorting according to their p values, then the top-ranked statistically significant features with p values less than 0.05 are preserved. Next, the upper limit of the features kept after the *t*-test is set to 1000, that is, when there are more than 1000 features having p values less than 0.05, only the top 1000 with lower p values would be kept. If the preserved features after the *t*-test screening are fewer than 500, they are all kept directly and definitively to form the candidate gene set without MIC screening; otherwise, the MIC-based selection will be followed. Second, we carry out MIC calculation on the preserved genes and resort them according to their MIC values, then the candidate gene set is generated from the top 500 genes with higher MIC values. For the two multiclass datasets, a candidate gene set is generated based only on the descending order of MIC values of all genes for the multivariate *t*-test cannot be performed directly. In the Table 5 in Supplementary Material, the number of statistically significant genes with *t*-test-based *p*-values less than 0.05 on each binary-class dataset are listed. In the "S8" section of Supplementary Material, we also give a simple comparison of *t*-test+MIC with other filter combinations.

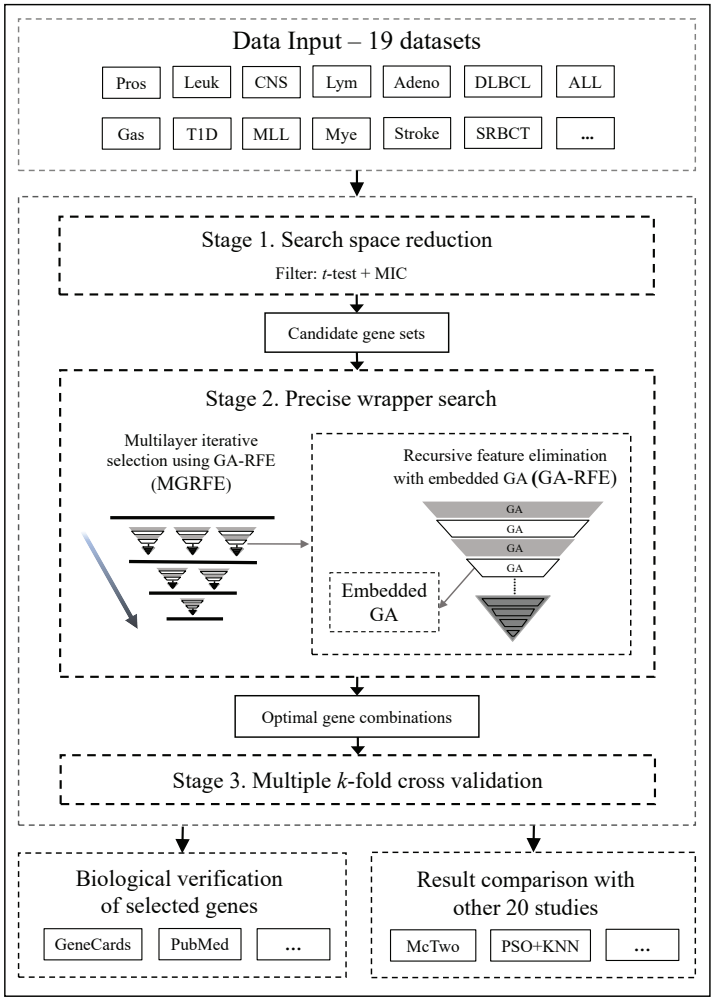

Fig. 1: A flowchart of the whole MGRFE procedure in this study.

2.3.2 Stage 2: Precise wrapper search

At Stage 2, we search the candidate gene set obtained from Stage 1 and compute the optimal gene combinations for further selection at Stage 3. MGRFE is a multilayer iterative feature selection method and its selection unit in each layer is a GA-RFE process. GA-RFE, the inverted triangle in MGRFE as shown in Fig. 1, is the recursive feature elimination process with every stair being embedded GA. Embedded GA is the integer-coded genetic algorithm with a dynamic-length chromosome. The key feature of MGRFE is GA-RFE in each layer, in which embedded GA is responsible for generating optimal gene combinations, and the RFE process is responsible for cutting down the gene number. Therefore, our method can find gene combinations with both significantly reduced sizes and excellent classification performance.

Embedded GA:

In our method, the modified GA using variable-length integer-coded chromosome is embedded in the RFE process as each stair in the inverted triangle of GA-RFE. The embedded GA includes the following steps. First, we initialize the GA population by a certain amount of individuals representing gene combinations with the same sizes. Then, we perform fitness calculation and genetic operators including mutation, crossover, and selection until the stopping criterion is satisfied. In the end, we return the best individuals that represent the best gene combinations to GA-RFE. The stopping criterion of embedded GA is iteration time, which is set to 1 to 3.

To embed GA in the RFE process and achieve the minimal informative genes, some modifications are made in the original GA. The embedded GA uses variable-length integer-coding technique for the chromosome in a GA individual, and each individual has

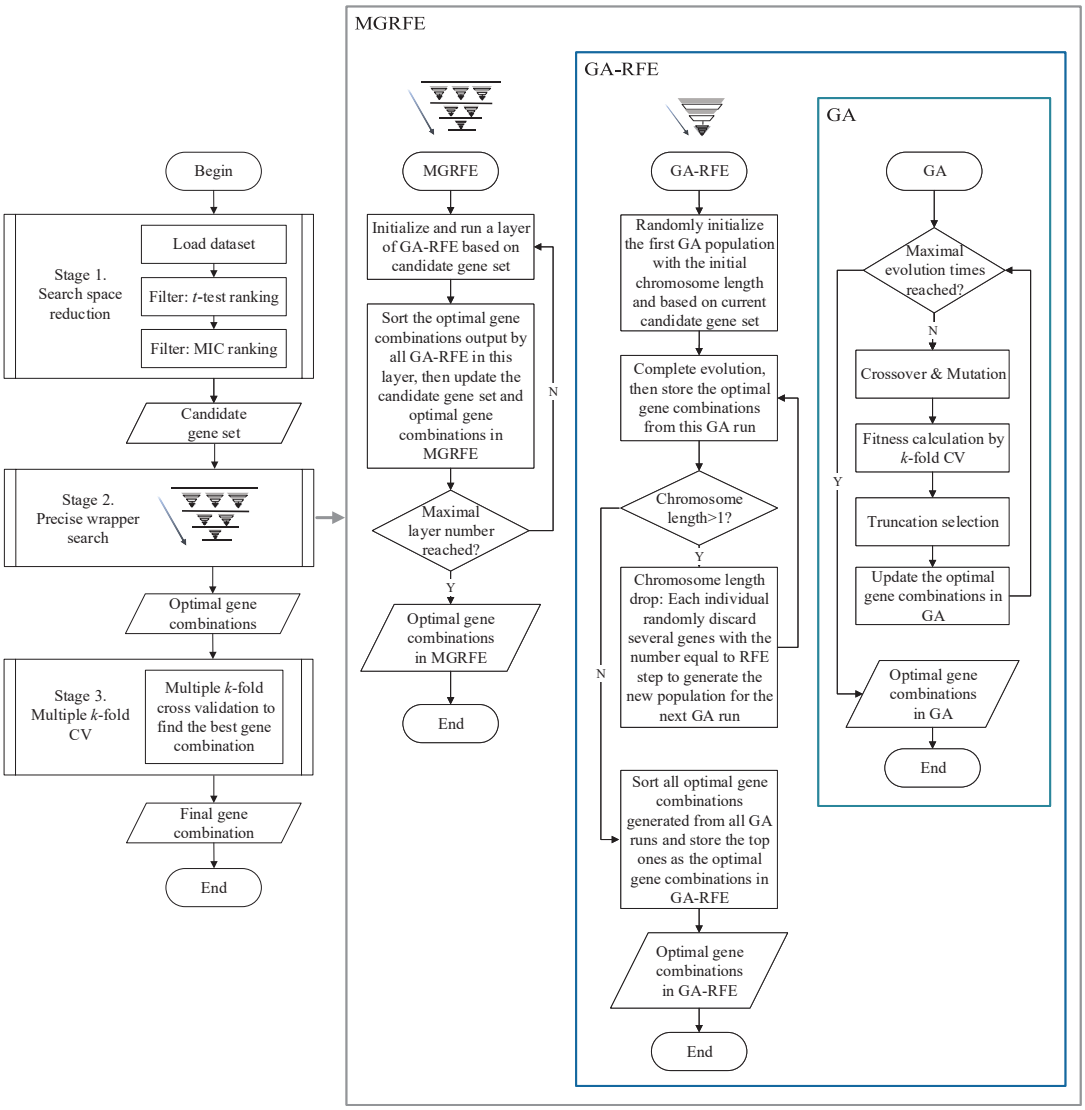

Fig. 2: The flowchart of MGRFE, which is divided into 3 stages: search space reduction, precise wrapper search, and multiple *k*-fold CV. Stage 2 is the core of MGRFE, which includes two key processes: GA-RFE and embedded modified GA.

a set of integers representing different genes to make up a gene combination. In every run of GA, the gene combinations represented by different individuals all have a fixed size. Between two adjacent GA runs in the RFE process, every individual sheds the same number of genes from its chromosome.

A truncation selection method is used as the selection operator in embedded GA [61], which simply ranks all individuals and selects the top ones to form the next generation. The elitism preservation mechanism is used to save the currently generated best individuals. The mutation and crossover operators for generating new individuals are adjusted to adapt to the variable-length integer-encoding technique. One main challenge that should be addressed in these two processes is avoidance of duplicated genes in every individual, which leads to the decline of the number of actual existing genes. Based on our encoding technique, the mutation operation for an individual is randomly changing some genes to others. It should be ensured that new genes do not exist in this individual previously to avoid repetitions. Crossover in embedded GA also involves single-point crossover which is the most widely used crossover method in binary encoding. Specifically, a random position is selected in the chromosome, and two parent individuals split themselves at this crossover point and then exchange chromosome tails to generate children individuals. After the crossover, the potential duplicate genes in the children individuals are replaced with other genes

from their parents to avoid decreasing the gene number. Fitness (F) of an individual is defined in Equation (2):

$$F = \begin{cases} Acc & \text{, balanced dataset} \\ \alpha Acc + (1-\alpha)Avc & \text{, imbalanced dataset} \end{cases} \quad (2)$$

, where α is an adjustment coefficient to deal with the imbalanced datasets. For imbalanced datasets, fitness defined as $\alpha Acc + (1-\alpha)Avc$ can adjust the trend of predicting samples as abundant classes for $Avc = (Sn + Sp)/2$ takes the correct prediction proportion of both sample classes into consideration. In our experiments, we take α 0.6 for imbalanced datasets. For balanced datasets, F is simply defined as Acc. F is calculated by 5-fold CV, and the employed classifier is Gaussian Naive Bayes classifier (NB) [62]. We sort different GA individuals based on two metrics, F and gene number. The individual with higher F is superior. For two individuals with the same F values, the one with a smaller gene number is superior. MGRFE and GA-RFE also use the above-mentioned sorting rule to rank different gene combinations.

Recursive feature elimination with embedded GA:

GA-RFE as shown in Fig. 2 is designed as an explicit recursive feature elimination process with embedded GA to find minimal discriminatory gene combinations. First, we randomly generate the initial GA population based on a certain candidate gene set and chromosome length. Then, we implement a chromosome length drop and a GA run in turns until the chromosome length in GA drops to 1. Finally, we sort the optimal gene combinations from all GA runs and then return the overall top-ranked gene combinations to MGRFE. The chromosome length drop means that every individual in the current GA population randomly discards the same number of genes to generate the new GA population for the next run. The number of discarded genes between two GA runs, the RFE step, is set from 1 to 3 according to the current chromosome length. A larger decline step is set for larger chromosome length to avoid time cost and a smaller decline step set for smaller chromosome length to do precise searching.

Multilayer iterative selection:

MGRFE is designed as a multilayer iterative feature selection method with the selection unit in each layer being GA-RFE. In every iteration layer, the GA-RFE processes analyze the current candidate gene set and return their obtained optimal gene combinations. Then the candidate gene set is reduced and subjected to the next layer of iterative selection. The candidate gene set used by the first layer of MGRFE is from the search space reduction stage. After each iteration layer, all optimal gene combinations in MGRFE will be sorted and the top-ranked ones will form the updated reduced candidate gene set. After the specified layers of iteration, MGRFE sorts all the optimal gene combinations and provides the top-ranked gene combinations for Stage 3 to execute further validation.

2.3.3 Stage 3: Multiple k-fold CV to select the final gene combination

Stage 3 is aimed at finding the optimal gene combination with the best classification performance and minimal variance among different CV processes. The *k*-fold CV is used for calculating the fitness of a GA individual. Multiple *k*-fold CV based on different random seeds is performed to further validate and select the final optimal gene combination.

3 RESULTS

In this section, comprehensive experiments on total 19 datasets are performed to validate the performance of MGRFE. Furthermore, the independent validation and biological verification of the selected genes are provided.

3.1 Results on Dataset One

The results of MGRFE on Dataset One including 17 binary datasets are given in Table 1, where six measurements calculated by 5-fold CV and the *t*-test and MIC-based gene rankings are listed. For 17 datasets, Acc values are all above 0.9 within 10 genes. Moreover, for 8 of 17 datasets (DLBCL, Leuk, ALL1, Lym, Adeno, Gas, Gas2, and Stroke), Acc reached 1.0 with gene number less than 5. MGRFE also show the strong robustness in dealing with imbalanced datasets like DLBCL, Colon, Leuk, ALL1, ALL4, and CNS, for which Sn, Sp, Avc, MCC, and AUC are all above 0.95 without being influenced by the data imbalance. According to the *t*-test and MIC-based gene ranking, the best gene feature subset is not always the highest-ranked features in the filter method, thus the filter algorithm alone cannot generate the optimal feature combination. It could be noted that the relative positions of selected genes in the two ranking methods are consistent on most datasets. The top-ranked genes in the *t*-test are also top-ranked in the MIC sorting (e.g. the selected gene on ALL1 is the top one in both *t*-test and MIC ranking). For 5 of 17 datasets, the top one gene according to the *t*-test appeared in the final selected gene subsets. Generally, the selected informative genes are top-ranked by the *t*-test and MIC methods. Therefore, the filter techniques are qualified for the search space reduction task. Moreover, MGRFE achieves relatively stable classification performance in 10 repetitions of 10-fold CV as depicted in Fig. 3.

TABLE 1: Results of MGRFE on 17 datasets in Dataset One

Datasets	Pos/Neg	Genes/Total	Sn	Sp	Acc	Avc	MCC	AUC	t-test/MIC-based gene rankings
DLBCL	58/19	**3**/7129	1.0	1.0	**1.0**	1.0	1.0	1.0	[13/8, 39/24, 54/52]
Pros	52/50	**4**/12625	0.980	0.982	**0.981**	0.981	0.963	0.98	[1/1, 15/47, 74/49, 694/618]
Colon	40/22	**6**/2000	1.0	0.960	**0.985**	0.980	0.969	0.97	[15/6, 58/21, 176/297, 225/80, 240/555, 495/482]
Leuk	47/25	**2**/7129	1.0	1.0	**1.0**	1.0	1.0	1.0	[4/3, 7/5]
Mye	137/36	**7**/12625	0.963	0.839	**0.937**	0.901	0.816	0.95	[3/3, 15/103, 83/142, 143/13, 378/217, 404/644, 569/707]
ALL1	95/33	**1**/12625	1.0	1.0	**1.0**	1.0	1.0	1.0	[1/1]
ALL2	65/35	**8**/12625	0.914	0.908	**0.910**	0.911	0.829	0.94	[1/80, 52/395, 78/3040, 80/1297, 522/2448, 687/2038, 737/920, 760/1449]
ALL3	24/101	**8**/12625	0.830	0.950	**0.927**	0.890	0.785	0.93	[4/500, 52/3437, 75/3010, 142/393, 488/443, 510/795, 715/1551, 770/1321]
ALL4	26/67	**6**/12625	1.0	0.986	**0.990**	0.993	0.978	0.99	[1/2, 6/45, 39/356, 282/226, 535/497, 754/1377]
CNS	39/21	**7**/7129	1.0	1.0	**1.0**	1.0	1.0	0.98	[9/907, 53/542, 130/620, 131/519, 272/57, 273/454, 520/49]
Lym	22/23	**3**/4026	1.0	1.0	**1.0**	1.0	1.0	1.0	[4/7, 5/4, 669/135]
Adeno	18/18	**1**/7457	1.0	1.0	**1.0**	1.0	1.0	1.0	[468/27]
Gas	29/36	**3**/22645	1.0	1.0	**1.0**	1.0	1.0	1.0	[22/1, 77/32, 306/36]
Gas1	72/72	**3**/22283	0.986	0.973	**0.980**	0.980	0.961	0.99	[132/74, 248/167, 717/500]
Gas2	62/62	**2**/22283	1.0	1.0	**1.0**	1.0	1.0	1.0	[38/6, 89/62]
T1D	57/44	**7**/54675	0.911	0.912	**0.911**	0.912	0.826	0.94	[14/2229, 25/1579, 113/1287, 559/1282, 578/353, 680/426, 978/1728]
Stroke	20/20	**4**/54675	1.0	1.0	**1.0**	1.0	1.0	1.0	[1/3, 23/115, 129/543, 276/539]

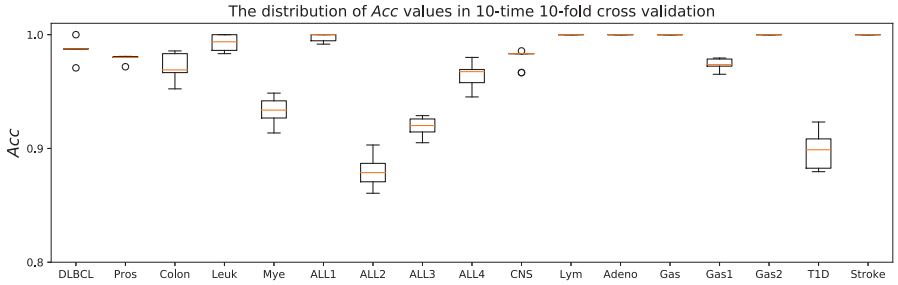

Fig. 3: The distribution of Acc values in 10-time 10-fold CV for the selected gene combinations of 17 datasets in Dataset One.

3.2 Comparison with other methods on Dataset One

McTwo [16] thoroughly tested all the datasets in Dataset One and demonstrated satisfactory performance. Here, we present the performance comparison between McTwo and MGRFE. Table 2 lists the overall maximal Acc and numbers of selected genes on total 17 datasets for MGRFE and McTwo. On all 17 datasets, MGRFE can obtain equal or better prediction performance compared to McTwo. On five datasets ALL2, ALL3, ALL4, Stroke and CNS, MGRFE achieves distinctly better classification performance than McTwo with relatively more genes. For a fairer and more specific comparison, the Acc values of the two algorithms are listed when the gene number of MGRFE is equal to McTwo as shown in Table 3. The results indicate that MGRFE still outperforms McTwo. Nonetheless, the Acc values associated with the usage of the gene numbers fall behind our optimal Acc values on these datasets. Thus, MGRFE selected somewhat more genes to achieve the optimal results.

3.3 Results on Dataset Two

Here we present the results of MGRFE on Dataset Two including three benchmark datasets, where two datasets are multiclass datasets. MGRFE selects five, two, and three genes in SRBCT, ALL_AML, and MLL respectively and the overall maximal Accs are all 1.0 in 5-fold CV. In our experiments, we notice that the Acc values of the best GA individuals are kept at 1.0 in the majority of gene number ranges and only begin to drop when the gene number is significantly reduced. We also carried out 10 repetitions of 10-fold CV to further validate the final selected gene combinations in Dataset Two as shown in Fig. 4. The mean of Accs for SRBCT, ALL_AML, and MLL are 1.0, 0.982, and 0.997, respectively, with standard deviations being 0.0, 0.008, and 0.006. The results confirm that MGRFE has high classification stability.

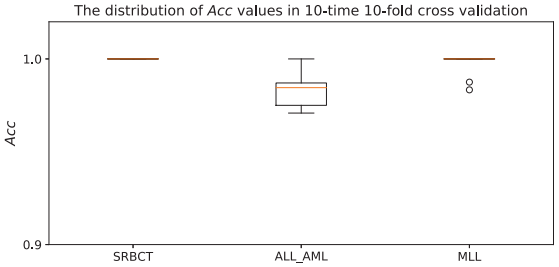

Fig. 4: The distribution of Acc values in 10-time 10-fold CV for the selected gene combinations in Dataset Two.

TABLE 2: Performance comparison between McTwo and MGRFE on 17 datasets in Dataset One

	DLBCL	Pros	Colon	Leuk	Mye	ALL1	ALL2	ALL3	ALL4	CNS	Lym	Adeno	Gas	Gas1	Gas2	T1D	Stroke
MGRFE *Acc*	1.0	0.981	0.985	1.0	0.937	1.0	0.910	0.927	0.990	1.0	1.0	1.0	1.0	0.980	1.0	0.911	1.0
McTwo *Acc*	1.0	0.95	0.9	1.0	0.85	1.0	0.75	0.8	0.88	0.85	1.0	1.0	0.97	0.95	1.0	0.81	0.85
MGRFE Genes	3	4	6	2	7	1	8	8	6	7	3	1	3	3	2	7	4
McTwo Genes	4	3	6	2	7	1	2	5	2	4	4	2	3	4	2	6	1

TABLE 3: Performance comparison on five datasets between MGRFE and McTwo when MGRFE uses the same gene numbers as McTwo does

Datasets	Methods	Genes	*Acc*
ALL2	MGRFE	2	0.760
	McTwo	2	0.75
ALL3	MGRFE	5	0.874
	McTwo	5	0.8
ALL4	MGRFE	2	0.896
	McTwo	2	0.88
CNS	MGRFE	4	0.921
	McTwo	4	0.85
Stroke	MGRFE	1	0.825
	McTwo	1	0.75

3.4 Comparison with other methods on Dataset Two

The performance comparison based on *Acc* and the gene number with other state-of-the-art algorithms of feature selection on the three benchmark datasets are presented in Tables 4, 5, and 6, respectively.

For the SRBCT dataset, MGRFE selected five genes and achieved 100% *Acc* in both 5-fold and 10-time 10-fold CV. In our computational experiments, combinations of four genes can reach 100% train and test *Acc* in 5-fold CV, but these gene combinations did not show classification stability in 10-time 10-fold CV. On the SRBCT dataset, Khan *et al.* [44] have applied an artificial neural network (ANN) and selected 96 genes to achieve 100% *Acc*. Tibshirani *et al.* [57] have used the nearest shrunken centroid-based method (NSC) and achieved 100% *Acc* by means of 43 genes. Fu and Fu-Liu [48] employed support vector machine (SVM)-RFE and achieved 100% *Acc* by means of 19 genes. Pal *et al.* [54] have applied feature selection multilayered perceptron (FSMLP) and non-Euclidean relational fuzzy c-means clustering (NERFCM) and found seven genes important for 100% *Acc*. Mohamad *et al.* [53] carried out improved binary PSO, and six genes were selected. Kar *et al.* [23] applied PSO and KNN and six genes were selected too. Moosa *et al.* [24] have achieved 100% *Acc* with the modified artificial bee colony algorithm (ABC) by means of five genes. Sharma *et al.* [55] have applied successive feature selection (SFS) with linear discriminant analysis (LDA) and nearest centroid classifier (NCC) and achieved 100% train and test *Acc* using four genes.

For the ALL_AML dataset, MGRFE selected two genes and achieved 100% 5-fold *Acc* and 98.2% 10-time 10-fold CV *Acc*. On this dataset, Fu and Fu-Liu [48] have achieved 100% train *Acc* by means of 19 genes via SVM-RFE. Yang *et al.* [59] have employed a gene-scoring technique and SVM, and four genes were selected to achieve 98.6% *Acc* in leave one out cross-validation (LOOCV). Mohamad *et al.* [53] have selected two genes to reach 100% CV *Acc* based on improved binary PSO. Dashtban and Balafar [27] have applied integer-encoding GA and SVM and selected 15 genes with 100% *Acc*. Ge *et al.* [16] have designed a two-step MIC-based method, and two genes were selected to reach 100% *Acc*.

For the MLL dataset, MGRFE selected three genes and achieved 100% 5-fold *Acc* and 99.7% *Acc* for 10-time 10-fold CV. On this dataset, Sharma *et al.* [55] have selected four genes with 100% train and test *Acc* based on SFS, LDA, and NCC. Mohamad *et al.* [53] have selected four genes with 100% CV *Acc* based on improved binary PSO. Dashtban and Balafar [27] have applied integer-encoding GA and SVM and selected 15 genes with 100% *Acc*. Kar *et al.* [23] have employed PSO and KNN to select four genes with 100% train and test *Acc* and 92.5% CV *Acc*.

3.5 Independent validation of selected features

We performed totally 10-group validation experiments on independent datasets to verify the generalization ability of selected gene subsets by MGRFE. The results are shown in Table 7. For each experiment, firstly, the selected gene probe features from the first dataset were transformed into the official gene symbols; secondly, the obtained gene symbols were transformed into corresponding gene probe Ids in the second dataset; thirdly, three kinds of classifier were used to perform 10 times *k*-fold cross validation using the samples and selected gene probe features on the second dataset. Particularly, no feature mapping between Gas1 and Gas2 for they are generated simultaneously and have identical feature set.

On tested datasets GSE56315 and GSE2604 with gene features from DLBCL and ALL1 respectively, NB, SVM and RF (Random Forest) classifiers all achieved 1.0 cross validation accuracy in each test. In particular, there are only 14 samples totally on GSE2604, which means the classifiers were trained on merely about 10 samples in each 5-fold cross validation. Thus, the

TABLE 4: Performance comparison among the methods on the SRBCT dataset

Experiments	Methods	Genes				CV *Acc*(%)				Train *Acc*(%)	Test *Acc*(%)
Khan *et al.* (2001) [44]	ANN	96				-				100	100
Tibshirani *et al.* (2002) [57]	NSC	43				-				100	100
Fu and Fu-Liu (2005) [48]	SVM-RFE	19				-				100	100
Yang *et al.* (2006) [59]		5CV		LOOCV		5CV		LOOCV			
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM		
	GS1	88	93	57	34	98	97.9	98.8	98.8	-	-
	GS2	90	99	77	96	98.1	99	98.8	100	-	-
	Cho's	98	98	82	80	90.2	94.3	92.8	98.8	-	-
	F-test	90	95	89	78	98	99.2	98.8	100	-	-
Pal *et al.* (2007) [54]	FSMLP+NERFCM	7				-				100	100
Li and Shu (2009) [52]	KLLE+LLE+PCA	20				-				100	100
Ji *et al.* (2011) [50]	PLSVIP	24				-				100	100
	PLSVEG	15				-				100	100
Mohamad *et al.* (2011) [53]	IBPSO	6				100				-	-
Zainuddin and Ong (2011) [60]	MSFCM+WNN	10				10CV				-	-
						100					
Lee *et al.* (2011) [51]	AGA+KNN	14				-				100	100
Sharma *et al.* (2012) [55]	SFS+LDA with NCC	4				-				100	100
	SFS+Bayes classifier	4				-				100	90
	SFS+NNC	4				-				100	95
Chen *et al.* (2014) [63]	PSODT	-				5CV				-	-
						92.94					
Kar *et al.* (2015) [23]	PSO+KNN	6				98.0159				100	100
Moosa *et al.* (2016) [24]	ABC	5				-				100	100
Dashtban and Balafar (2017) [27]	GA+SVM	18				-				100	100
This paper	**MGRFE**	**5**				**5CV**	**10-10CV**			**100**	**100**
						100	**100**				

In Tables 4, 5, and 6, 5CV means 5-fold cross validation; 10CV means 10-fold cross validation; 10-10CV represents 10-time 10-fold cross validation; and LOOCV represents leave one out cross validation.

TABLE 5: Performance comparison among the methods on the ALL_AML (Leukaemia) dataset

Experiments	Methods	Genes				CV *Acc*(%)				Train *Acc*(%)	Test *Acc*(%)
Fu and Fu-Liu (2005) [48]	SVM-RFE	4				-				100	97.06
Yang *et al.* (2006) [59]		5CV		LOOCV		5CV		LOOCV			
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM		
	GS1	100	93	60	4	97.9	97.9	98.6	98.6	-	-
	GS2	85	98	10	25	97.1	97.4	98.6	98.6	-	-
	Cho's	100	98	9	80	96.8	97	97.2	98.6	-	-
	F-test	96	99	25	33	97.4	97.5	98.6	98.6	-	-
Shen *et al.* (2008) [56]	Stepwise	3				-				90.83	88.14
	Pure TS	5				-				95.83	94.24
	Pure PSO	7				-				94.75	94.19
	HPSOTS	7				-				98.08	95.81
Wong and Liu (2010) [58]	Probabilistic mechanism	-				SVM	KNN			-	-
						97.38	98.21				
Ji *et al.* (2011) [50]	PLSVIP	9				-				100	100
	PLSVEG	8				-				100	100
Mohamad *et al.* (2011) [53]	IBPSO	2				100				-	-
Zainuddin and Ong (2011) [60]	MSFCM+WNN	10				10CV				-	-
						98.61					
Chandra and Gupta (2011) [47]	RNBC	-				10CV				-	-
						RNBC	NBC	KNN			
						94.29	84.29	85.71			
Kumar *et al.* (2012) [49]	GSA	10				100				-	-
Kar *et al.* (2015) [23]	PSO+KNN	3				95.8868				100	97.0588
Ge *et al.* (2016) [16]	McTwo	2				-				100	100
Dashtban and Balafar (2017) [27]	GA+SVM	15				-				100	100
This paper	**MGRFE**	**2**				**5CV**	**10-10CV**			**100**	**100**
						100	**98.2**				

TABLE 6: Performance comparison among the methods on the MLL dataset

Experiments	Methods	Genes				CV *Acc*(%)				Train *Acc*(%)	Test *Acc*(%)
Yang *et al.* (2006) [59]		5CV		LOOCV		5CV		LOOCV			
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM		
	GS1	29	99	97	56	94.8	95.2	97.2	97.2	-	-
	GS2	91	87	90	91	94.9	94.7	97.2	97.2	-	-
	Cho's	93	89	23	44	96	95.5	97.2	95.8	-	-
	F-test	99	100	65	31	95.4	94.8	95.8	95.8	-	-
Mohamad *et al.* (2011) [53]	IBPSO	4				100				-	-
Chandra and Gupta (2011) [47]	RNBC	-				10CV				-	-
						RNBC	NBC	KNN			
						87.14	80	68.57			
Sharma *et al.* (2012) [55]	SFS+LDA with NCC	4				-				100	100
	SFS+Bayes classifier	4				-				100	100
	SFS+NNC	4				-				100	93
Chen *et al.* (2014) [63]	PSODT	-				5CV				-	-
						100					
Kar *et al.* (2015) [23]	PSO+KNN	4				92.5439				100	100
This paper	**MGRFE**	**3**				**5CV**	**10-10CV**			**100**	**100**
						100	**99.7**				

selected unique gene *CD3D* is one ideal discrimination for B-cell acute lymphoblastic leukemia (ALL) and T-cell ALL. On GSE2685, the sample number is merely 30, and only contains one gene probe Id mapped from the *LIFR*, one of three selected genes on Gastric dataset. But NB and SVM still achieved acceptable cross validation accuracies over 0.8. Except three tested datasets of GSE8511, GSE2685 and GSE8514 with samples less than 50, the prediction accuracies of three classifiers are above 0.9 on all other seven datasets in cross validation. The independent validation results proved that the selected genes features by MGRFE in each dataset have strong association with the disease phenotype and can be selected as biomarker candidates.

3.6 Biological inferences of the genes selected by MGRFE

The sizes of gene subsets selected by MGRFE with the 100% 5-fold CV *Acc* on datasets Leuk, Gas, and ALL1 are only two, three, and one, respectively. For each gene selected by MGRFE on these three datasets, we surveyed the number of published literatures involving the gene of interest and the related cancer phenotype in PubMed on July 9, 2018. The literature-mining results on these three datasets are shown in Table 8. Moreover, the gene probes finally selected by MGRFE on all 19 datasets are provided in the "S5" section of the Supplementary Material. In the Leukaemia dataset, our selected genes are *CD33* and *TCF3*. In PubMed, there are 3001 published literatures about *CD33*, among which 1753 (58.94%) papers discuss the relevance of *CD33* to leukemia. And there are 569 publications about *TCF3* in PubMed, among which 115 (20.21%) papers confirming the association between *TCF3* and leukemia. According to GeneCards, the E protein encoded by *TCF3* performs a critical function in lymphopoiesis and is necessary for B and T lymphocytes. This gene is related to cancers including ALL (t(1;19), with *PBX1*), childhood leukemia (t(19;19), with *TFPT*), and acute leukemia (t(12;19)). In the Gastric dataset, genes *COL8A1*, *SEMA6D*, and *LIFR* are selected by MGRFE, and there are 187 publications in PubMed confirming their relevance to cancer, but only five papers reveal their relations with gastric cancer. According to the excellent classification performance of these three genes on gastric cancer, they could be novel biomarker candidates for gastric cancer. In the ALL1 dataset, only one gene, *CD3D*, is selected by MGRFE. There are 84 publications in PubMed about *CD3D*, among which 13 (15.48%) papers revealing the relevance of *CD3D* to leukemia. In [64], it has also been pointed out that gene *CD3D* is one ideally discriminatory feature and gave a diagnostic rule when the expression of *CD3D* is below a certain cutoff limit. Regarding *CD3D*, GeneCards explains that this gene is involved in T-cell development and signal transduction, whereas defects in this gene will lead to severe combined immunodeficiency.

4 Discussion

The proposed MGRFE is a novel multilayer recursive feature elimination algorithm based on an embedded integer-coded genetic algorithm. MGRFE is aimed at selecting minimal discriminatory gene feature subset associated closely with the disease phenotype. MGRFE could be regarded as a complementary feature selection algorithm for high-dimensional data especially for gene expression data analysis.

The main innovation of MGRFE is effectively combining the advantages of evolution calculation of the embedded GA with the explicit feature reduction manner of the RFE process in GA-RFE. Therefore, our developed MGRFE can perform explicit feature elimination along with the evolution optimization search and achieve relatively quick convergence speed. First, compared with other evolutionary-computation-based feature selection algorithms, our proposed MGRFE has shown higher convergence speed and obtained a slightly smaller discriminatory gene subset. For selecting informative gene features in a microarray, the state-of-the-art methods are commonly evolutionary-computation based. Meanwhile, almost all the existing evolution-based gene selection methods mainly rely on binary encoding and none of them take advantage of the RFE technique [14], [23], [24], [27], [65], [66], [67]. Nonetheless, the binary encoding has the shortcomings of the probable existing irrelevant features in a selected feature subset and high time cost to converge because there are thousands of genes in a microarray. Meanwhile, the fixed coding length of binary encoding leads to impossibility of explicit recursive feature reduction. Instead, MGRFE utilizes a variable-length integer-encoding technique in embedded GA and cuts down the encoding length recursively in an RFE process, which can quickly remove the irrelevant and redundant features and converge to a minimal informative feature combination. In 2017, Dashtban and Balafar also proposed an integer-coded GA with dynamic coding length for gene selection [27], but they did not employ the recursive feature reduction technique. In fact, their method selected 18 and 15 genes with *Acc* 100% on the SRBCT and ALL_AML datasets, respectively. But MGRFE only needs five and two genes to accomplish the same performance. Second, compared with the original SVM-RFE [17], MGRFE is more flexible and robust. SVM-RFE ranks all gene features by the weight vector from SVM

TABLE 7: Independent validation of selected gene features by MGRFE with 10-time k-fold cross validation.

Feature From / #Features	Feature Tested / #Features	#Samples	Classifier	*Sn*	*Sp*	*Acc*	*Avc*	*MCC*	*AUC*
Leuk / 2	MLL / 4	52	NB	**0.963**	**0.955**	**0.961**	**0.959**	**0.929**	**0.993**
			SVM	0.935	0.887	0.913	0.911	0.844	0.975
			RF	0.960	**0.955**	0.959	0.958	0.925	0.977
Gas1 / 2	Gas2 / 3	124	NB	**0.968**	**0.966**	**0.967**	**0.967**	**0.937**	**0.993**
			SVM	0.952	**0.982**	**0.967**	**0.967**	**0.937**	0.992
			RF	0.957	0.931	0.944	0.944	0.895	0.987
Gas2 / 3	Gas1 / 2	144	NB	**0.949**	0.968	**0.958**	**0.958**	**0.920**	**0.975**
			SVM	0.941	**0.972**	0.956	0.956	0.916	0.970
			RF	0.936	0.958	0.947	0.947	0.900	0.974
DLBCL / 3	GSE56315 / 7	88	NB	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
			SVM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
			RF	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Prostate / 4	GSE8511 / 5	41	NB	**0.884**	**0.852**	**0.870**	**0.868**	**0.753**	**0.935**
			SVM	**0.900**	0.665	0.806	0.783	0.582	0.900
			RF	0.868	0.752	0.822	0.810	0.646	0.917
Gastric / 3	GSE2685 / 1	30	NB	0.919	**0.650**	**0.846**	**0.785**	**0.584**	0.861
			SVM	**0.990**	0.440	0.843	0.715	0.464	**0.865**
			RF	0.862	0.500	0.765	0.681	0.365	0.686
Gastric / 3	GSE66229 / 7	400	NB	0.903	**0.896**	0.902	0.900	0.764	0.961
			SVM	**0.955**	0.864	0.932	0.909	0.823	**0.971**
			RF	0.950	0.894	**0.936**	**0.922**	**0.835**	**0.971**
Adenoma / 1	GSE8514 / 3	15	NB	0.900	0.800	0.867	0.850	0.700	**0.960**
			SVM	0.900	0.500	0.767	0.700	0.400	0.920
			RF	**0.910**	**0.820**	**0.880**	**0.865**	**0.730**	0.950
Colon / 6	GSE44076 / 23	148	NB	**0.988**	0.950	**0.976**	**0.969**	**0.948**	0.996
			SVM	0.969	0.952	0.963	0.961	0.924	0.995
			RF	0.977	**0.960**	0.972	**0.969**	0.940	**0.998**
ALL1 / 1	GSE2604 / 4	14	NB	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
			SVM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
			RF	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

On the tested datasets with sample number greater than 50, 10-time 10-fold CV were performed with different random seeds. Meanwhile, 10-time 5-fold CV were performed on datasets with samples less than 50.
The later seven validation datasets are retrieved from GEO and named as their GEO accessions.
On MLL, only "ALL" and "AML" samples are used. On GSE8511, the "Local Prostate Cancer" and "Metastatic Prostate Cancer" samples are combined together. On GSE44076, the "Mucosa sample from healthy Normal donor", "Normal paired sample from patient" samples are combined together. On GSE8511, the features containing Null values have been abandoned. On GSE2604, the samples with Null values have been removed.
NB, SVM and RF represent Gaussian Naive Bayes, Support Vector Machine and Random Forest classifiers, respectively. The **bold** face values denote the highest performance achieved by classifiers.

TABLE 8: Literature mining in PubMed for the selected genes on Leukaemia, ALL1 and Gastric datasets

Dataset	Probe ID	Gene	PubMed hits for gene of interest	PubMed hits for gene of interest and leukemia[1](*Ratio1)	
Leukaemia	*M23197_at*	CD33 Molecule(*CD33*)	3001	1753(58.41%)	
	M31523_at	Transcription Factor 3(*TCF3*)	569	115(20.21%)	
ALL1	*38319_at*	CD3d molecule(*CD3D*)	84	13(15.48%)	
Dataset	Probe ID	Gene	PubMed hits for gene of interest	PubMed hits for gene of interest and cancer[2](**Ratio2)	PubMed hits for gene of interest and gastric cancer[3](***Ratio3)
Gastric	*226237_at*	collagen type VIII alpha 1 chain(*COL8A1*)	66	15(22.73%)	2(13.33%)
	226492_at	semaphorin 6D(*SEMA6D*)	41	13(31.71%)	1(7.69%)
	227771_at	leukemia inhibitory factor receptor alpha(*LIFR*)	463	159(34.34%)	2(1.26%)

[1] gene of interest [All Fields] AND ("leukemia"[All Fields]).
[2] gene of interest [All Fields] AND ("tumour"[All Fields] OR "neoplasms"[MeSH Terms] OR "neoplasms"[All Fields] OR "tumor"[All Fields] OR "cancer"[All Fields] OR "carcinoma"[All Fields]).
[3] gene of interest [All Fields] AND ("stomach"[All Fields] OR "gastric"[All Fields]) AND ("tumour"[All Fields] OR "neoplasms"[MeSH Terms] OR "neoplasms"[All Fields] OR "tumor"[All Fields] OR "cancer"[All Fields] OR "tumor"[All Fields] OR "carcinoma"[All Fields]).
* Ratio1 = #(gene of interest-leukemia related literatures)/#(gene of interest literatures).
** Ratio2 = #(gene of interest-cancer related literatures)/#(gene of interest literatures).
*** Ratio3 = #(gene of interest-gastric cancer related literatures)/#(gene of interest-cancer related literatures)

with a linear kernel and removes the feature with the smallest weight recursively. Nonetheless, SVM-RFE has the following potential limitations: 1) the greedy feature elimination manner makes it not very robust in achieving the minimal discriminatory feature subset; 2) features are only selected on training dataset, no independent test or validation involved; 3) other classifiers that can be used to replace the linear kernal SVM are limited. In GA-RFE, the population evolution of GA is more robust and fault-tolerant than the feature refining process on a single feature subset in original SVM-RFE. Meanwhile, the generation and evaluation processes of feature subsets are separated in GA-RFE. Thus, independent validation for feature subset could be done and all kinds of classifies could be embedded in MGRFE to obtain their most suitable feature subsets. In 2005, Fu and Fu-Liu evaluated SVM-RFE on datasets SRBCT and ALL_AML and finally selected 19 and four genes to achieve 100% and 97.6% test *Acc*s, respectively [48]. But MGRFE selected only five and two genes to attain 100% *Acc*s in 5-fold CV for the same datasets. As for the selection operator of our embedded GA, we find that the widely used roulette wheel selection [61] is inferior to simple

truncation selection in this gene selection problem. The fitness gaps between different GA individuals are usually slight and account for only a tiny proportion of a fitness value. This situation provides all individuals with nearly the same area occupation in the roulette wheel and leads to the inefficiency of roulette wheel selection.

The 19 popular benchmark microarray datasets including multiclass and imbalanced datasets are employed to validate MGRFE. According to the performance comparison with other various algorithms, our proposed MGRFE is proved to be superior to most of the current state-of-the-art feature selection methods. MGRFE offers smaller informative gene subsets but the same or higher phenotype diagnosis accuracies. Many promising results are obtained by MGRFE on these datasets. MGRFE can reach Acc 100% within only five genes for 10 (52.6%) of 19 datasets, and Acc higher than 90% within 10 genes for all 19 datasets, in 5-fold CV. MGRFE also possesses strong robustness for multiclass datasets and imbalanced datasets according to metrics Sn, Sp, Avc, MCC, and AUC.

To conclude, the chief research contribution in theory is providing a novel feature selection method which combines embedded genetic algorithm with recursive feature elimination process, working as a creative thought for future research. To the best of our knowledge, none previous studies have designed an evolutionary algorithm using variable length integer encoding approach in a recursive manner to select minimal discriminatory feature subset in high-dimension data, which is described in this paper. Meanwhile, through theoretical and experimental comparisons, our proposed MGRFE could outperform mostly other state-of-the-art algorithms for gene selection on microarray. Therefore, the proposed method MGRFE is worthy to be generalized to more feature selection problems on high-dimensional data characterized by the "large p small n" paradigm and applied in several practical fields.

Furthermore, our presented MGRFE would be useful in medical diagnosis as well as further biomedical research. The biological associations with phenotypes using literature mining in PubMed for the selected genes confirmed that the genes selected by MGRFE are biologically relevant to cancer phenotypes. Therefore, the informative genes selected by MGRFE could be novel biomarker candidates that are useful for better understanding the molecule mechanism related to the disease phenotypes and developing potential early detection and molecularly-targeted therapies for cancer diseases. For clinical applications involving microarrays, MGRFE can contribute to the development of a potential simplified procedure for diagnosis of cancer subgroups by selecting the minimal discriminatory gene subsets, which will cut down the cost of medical diagnoses.

ACKNOWLEDGMENTS

The authors would like to thank the National Natural Science Foundation of China [Grant number 61572105, 61872418 and 71774154] and the Natural Science Foundation of Jilin Province (20180101331JC). Also, we are grateful to the two revered reviewers for their constructive comments.

REFERENCES

[1] G. Diao and A. N. Vidyashankar, "Assessing genome-wide statistical significance for large p small n problems," *Genetics*, vol. 194, no. 3, pp. 781–783, 2013.

[2] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on knowledge and data engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[3] N. Zhou and L. Wang, "A modified t-test feature selection method and its application on the hapmap genotype data," *Genomics, proteomics & bioinformatics*, vol. 5, no. 3, pp. 242–249, 2007.

[4] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *Tools with artificial intelligence, 1995. proceedings., seventh international conference on*. IEEE, 1995, pp. 388–391.

[5] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[6] C. Lin, T. Miller, D. Dligach, R. Plenge, E. Karlson, and G. Savova, "Maximal information coefficient for feature selection for clinical document classification," in *ICML Workshop on Machine Learning for Clinical Data. Edingburgh, UK*, 2012.

[7] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, no. 5439, pp. 531–537, 1999.

[8] X. Q. Cui and G. A. Churchill, "Statistical tests for differential expression in cdna microarray experiments," *Genome Biology*, vol. 4, no. 4, 2003. [Online]. Available: ⟨GotoISI⟩://WOS:000182696200003

[9] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe, "A survey on filter techniques for feature selection in gene expression microarray analysis," *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 4, pp. 1106–1119, 2012. [Online]. Available: ⟨GotoISI⟩://WOS:000304147000018

[10] N. Sato, I. M. Sanjuan, M. Heke, M. Uchida, F. Naef, and A. H. Brivanlou, "Molecular signature of human embryonic stem cells and its comparison with the mouse," *Developmental Biology*, vol. 260, no. 2, pp. 404–413, 2003. [Online]. Available: ⟨GotoISI⟩://WOS:000184946000010

[11] P. Baldi and A. D. Long, "A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes," *Bioinformatics*, vol. 17, no. 6, pp. 509–519, 2001. [Online]. Available: ⟨GotoISI⟩://WOS:000169404700004

[12] R. J. Fox and M. W. Dimmic, "A two-sample bayesian t-test for microarray data," *Bmc Bioinformatics*, vol. 7, 2006. [Online]. Available: ⟨GotoISI⟩://WOS:000236547800001

[13] P. Pavlidis, Q. H. Li, and W. S. Noble, "The effect of replication on gene expression microarray experiments," *Bioinformatics*, vol. 19, no. 13, pp. 1620–1627, 2003. [Online]. Available: ⟨GotoISI⟩://WOS:000185310600004

[14] Q. Shen, W. M. Shi, and W. Kong, "Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data," *Computational Biology and Chemistry*, vol. 32, no. 1, pp. 53–60, 2008. [Online]. Available: ⟨GotoISI⟩://WOS:000253028600007

[15] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[16] R. Ge, M. Zhou, Y. Luo, Q. Meng, G. Mai, D. Ma, G. Wang, and F. Zhou, "Mctwo: a two-step feature selection algorithm based on maximal information coefficient," *BMC bioinformatics*, vol. 17, no. 1, p. 142, 2016.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1, pp. 389–422, 2002.

[18] D. B. Skalak, "Prototype and feature selection by sampling and random mutation hill climbing algorithms," in *Proceedings of the eleventh international conference on machine learning*, 1994, pp. 293–301.

[19] H. Deng and G. Runger, "Feature selection via regularized trees," in *Neural Networks (IJCNN), The 2012 International Joint Conference on*. IEEE, 2012, pp. 1–8.

[20] K.-H. Chen, K.-J. Wang, M.-L. Tsai, K.-M. Wang, A. M. Adrian, W.-C. Cheng, T.-S. Yang, N.-C. Teng, K.-P. Tan, and K.-S. Chang, "Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm," *BMC bioinformatics*, vol. 15, no. 1, p. 49, 2014.

[21] C. Jin, S.-W. Jin, and L.-N. Qin, "Attribute selection method based on a hybrid bpnn and pso algorithms," *Applied Soft Computing*, vol. 12, no. 8, pp. 2147–2155, 2012.

[22] X. Li, N. Xiao, C. Claramunt, and H. Lin, "Initialization strategies to enhancing the performance of genetic algorithms for the p-median problem," *Computers & Industrial Engineering*, vol. 61, no. 4, pp. 1024–1034, 2011.

[23] S. Kar, K. D. Sharma, and M. Maitra, "Gene selection from microarray gene expression data for classification of cancer subgroups employing pso and adaptive k-nearest neighborhood technique," *Expert Systems with Applications*, vol. 42, no. 1, pp. 612–627, 2015.

[24] J. M. Moosa, R. Shakur, M. Kaykobad, and M. S. Rahman, "Gene selection for cancer classification with the help of bees," *BMC medical genomics*, vol. 9, no. 2, p. 47, 2016.

[25] S. Oreski and G. Oreski, "Genetic algorithm-based heuristic for feature selection in credit risk assessment," *Expert systems with applications*, vol. 41, no. 4, pp. 2052–2064, 2014.

[26] M. Jung and J. Zscheischler, "A guided hybrid genetic algorithm for feature selection with expensive cost functions," *Procedia Computer Science*, vol. 18, pp. 2337–2346, 2013.

[27] M. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," *Genomics*, vol. 109, no. 2, pp. 91–107, 2017.

[28] Y. Ding and D. Wilkins, "Improving the performance of svm-rfe to select genes in microarray data," *BMC bioinformatics*, vol. 7, no. 2, p. S12, 2006.

[29] C. Furlanello, M. Serafini, S. Merler, and G. Jurman, "An accelerated procedure for recursive feature ranking on microarray data," *Neural Networks*, vol. 16, no. 5, pp. 641–648, 2003.

[30] P. Guo, Y. Luo, G. Mai, M. Zhang, G. Wang, M. Zhao, L. Gao, F. Li, and F. Zhou, "Gene expression profile based classification models of psoriasis," *Genomics*, vol. 103, no. 1, pp. 48–55, 2014.

[31] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus *et al.*, "Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature medicine*, vol. 8, no. 1, pp. 68–74, 2002.

[32] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie *et al.*, "Gene expression correlates of clinical prostate cancer behavior," *Cancer cell*, vol. 1, no. 2, pp. 203–209, 2002.

[33] S. Chiaretti, X. Li, R. Gentleman, A. Vitale, M. Vignetti, F. Mandelli, J. Ritz, and R. Foa, "Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival," *Blood*, vol. 103, no. 7, pp. 2771–2778, 2004.

[34] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, C. Lau *et al.*, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.

[35] A. A. Alizadeh, M. B. Elsen, R. E. Davis, C. Ma *et al.*, "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, p. 503, 2000.

[36] D. A. Notterman, U. Alon, A. J. Sierk, and A. J. Levine, "Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays," *Cancer research*, vol. 61, no. 7, pp. 3124–3130, 2001.

[37] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.

[38] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, no. 5439, pp. 531–537, 1999.

[39] E. Tian, F. Zhan, R. Walker, E. Rasmussen, Y. Ma, B. Barlogie, and J. D. Shaughnessy Jr, "The role of the wnt-signaling antagonist dkk1 in the development of osteolytic lesions in multiple myeloma," *New England Journal of Medicine*, vol. 349, no. 26, pp. 2483–2494, 2003.

[40] Y. Wu, H. Grabsch, T. Ivanova, I. B. Tan, J. Murray, C. H. Ooi, A. I. Wright, N. P. West, G. G. Hutchins, J. Wu *et al.*, "Comprehensive genomic meta-analysis identifies intratumoural stroma as a predictor of survival in patients with gastric cancer," *Gut*, pp. gutjnl–2011, 2012.

[41] G. Wang, N. Hu, H. H. Yang, L. Wang, H. Su, C. Wang, R. Clifford, E. M. Dawsey, J.-M. Li, T. Ding *et al.*, "Comparison of global gene expression of gastric cardia and noncardia cancers from a high-risk population in china," *PloS one*, vol. 8, no. 5, p. e63826, 2013.

[42] H. Levy, X. Wang, M. Kaldunski, S. Jia, J. Kramer, S. J. Pavletich, M. Reske, T. Gessel, M. Yassai, M. W. Quasney *et al.*, "Transcriptional signatures as a disease-specific and predictive inflammatory biomarker for type 1 diabetes," *Genes and immunity*, vol. 13, no. 8, p. 593, 2012.

[43] T. Krug, J. P. Gabriel, R. Taipa, B. V. Fonseca, S. Domingues-Montanari, I. Fernandez-Cadenas, H. Manso, L. O. Gouveia, J. Sobral, I. Albergaria *et al.*, "Ttc7b emerges as a novel risk factor for ischemic stroke through the convergence of

several genome-wide approaches," *Journal of Cerebral Blood Flow & Metabolism*, vol. 32, no. 6, pp. 1061–1072, 2012.
[44] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson *et al.*, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature medicine*, vol. 7, no. 6, p. 673, 2001.
[45] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature genetics*, vol. 30, no. 1, p. 41, 2002.
[46] C. Bhattacharyya, L. Grate, A. Rizki, D. Radisky, F. Molina, M. I. Jordan, M. J. Bissell, and I. S. Mian, "Simultaneous classification and relevant feature identification in high-dimensional spaces: application to molecular profiling data," *Signal Processing*, vol. 83, no. 4, pp. 729–743, 2003.
[47] B. Chandra and M. Gupta, "Robust approach for estimating probabilities in naïve–bayes classifier for gene expression data," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1293–1298, 2011.
[48] L. M. Fu and C. S. Fu-Liu, "Evaluation of gene importance in microarray data based upon probability of selection," *BMC bioinformatics*, vol. 6, no. 1, p. 67, 2005.
[49] P. G. Kumar, T. A. A. Victoire, P. Renukadevi, and D. Devaraj, "Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm," *Expert Systems with Applications*, vol. 39, no. 2, pp. 1811–1821, 2012.
[50] G. Ji, Z. Yang, and W. You, "Pls-based gene selection and identification of tumor-specific genes," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 6, pp. 830–841, 2011.
[51] C.-P. Lee, W.-S. Lin, Y.-M. Chen, and B.-J. Kuo, "Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method," *Expert Systems with Applications*, vol. 38, no. 5, pp. 4661–4667, 2011.
[52] X. Li and L. Shu, "Kernel based nonlinear dimensionality reduction for microarray gene expression data analysis," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7644–7650, 2009.
[53] M. S. Mohamad, S. Omatu, S. Deris, and M. Yoshioka, "A modified binary particle swarm optimization for selecting the small subset of informative genes from gene expression data," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 6, pp. 813–822, 2011.
[54] N. R. Pal, K. Aguan, A. Sharma, and S.-i. Amari, "Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering," *BMC bioinformatics*, vol. 8, no. 1, p. 5, 2007.
[55] A. Sharma, S. Imoto, and S. Miyano, "A top-r feature selection algorithm for microarray gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 9, no. 3, pp. 754–764, 2012.
[56] Q. Shen, W.-M. Shi, and W. Kong, "Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data," *Computational Biology and Chemistry*, vol. 32, no. 1, pp. 53–60, 2008.
[57] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6567–6572, 2002.
[58] T.-T. Wong and K.-L. Liu, "A probabilistic mechanism based on clustering analysis and distance measure for subset gene selection," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2144–2149, 2010.
[59] K. Yang, Z. Cai, J. Li, and G. Lin, "A stable gene selection in microarray data analysis," *BMC bioinformatics*, vol. 7, no. 1, p. 228, 2006.
[60] Z. Zainuddin and P. Ong, "Reliable multiclass cancer classification of microarray gene expression profiles using an improved wavelet neural network," *Expert Systems with Applications*, vol. 38, no. 11, pp. 13 711–13 722, 2011.
[61] T. Blickle and L. Thiele, "A comparison of selection schemes used in genetic algorithms," 1995.
[62] H. Zhang, "Exploring conditions for the optimality of naive bayes," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, no. 02, pp. 183–198, 2005.
[63] K.-H. Chen, K.-J. Wang, M.-L. Tsai, K.-M. Wang, A. M. Adrian, W.-C. Cheng, T.-S. Yang, N.-C. Teng, K.-P. Tan, and K.-S. Chang, "Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm," *BMC bioinformatics*, vol. 15, no. 1, p. 49, 2014.
[64] L. Wong, "Lecture 4: Gene expression analysis," 2012.
[65] H. M. Alshamlan, G. H. Badr, and Y. A. Alohali, "Genetic bee colony (gbc) algorithm: A new gene selection method for microarray cancer classification," *Computational Biology and Chemistry*, vol. 56, pp. 49–60, 2015. [Online]. Available: ⟨GotoISI⟩://WOS:000356111800009
[66] B. A. Garro, K. Rodriguez, and R. A. Vazquez, "Classification of dna microarrays using artificial neural networks and abc algorithm," *Applied Soft Computing*, vol. 38, pp. 548–560, 2016. [Online]. Available: ⟨GotoISI⟩://WOS:000366805900040
[67] H. Salem, G. Attiya, and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," *Applied Soft Computing*, vol. 50, pp. 124–134, 2017. [Online]. Available: ⟨GotoISI⟩://WOS:000395834100010

Cheng Peng was born in Shanxi, China in 1996. He was a undergraduate in the College of Computer Science and Technology, Jilin University and received his BE degree in computer science and technology in 2018. He is currently a graduate student in the School of Software, Tsinghua University.

Ying Li was born in Henan, China in 1978. She received the Ph.D. degree in computational mathematics from Jilin University, Changchun, China, in 2004. She is currently an associate professor with the College of Computer Science and Technology. She was a postdoctoral fellow at Tsinghua University from 2005 to 2007. She was a visiting scholar at University of Georgia of United Kingdom from 2011 to 2012. She has published more than 30 journal and conference papers. Her current research interests include machine learning and bioinformatics.

Cheng Peng was born in Shanxi, China in 1996. He was a undergraduate in the College of Computer Science and Technology, Jilin University and received his BE degree in computer science and technology in 2018. He is currently a graduate student in the School of Software, Tsinghua University.