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

Journal:	*IEEE/ACM Transactions on Computational Biology and Bioinformatics*
Manuscript ID	TCBB-2018-05-0199.R1
Manuscript Type:	Regular Paper
Keywords:	Gene selection, Genetic algorithm, Recursive feature elimination, Microarray data, Cancer

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 with a dynamic-length chromosome, 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 significant 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 mostly used in the feature preprocessing stage to generate a candidate gene set in the analysis of microarray. 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, genes in the cell interact with other genes to perform a certain biological function, but the filter methods select genes individually rather than gene combinations [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, which are usually superior to filter algorithms but more time-consuming. A variety of wrapper algorithms have emerged involving simulated annealing, 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 operation like mutation and crossover. Thus, these methods lack the precise control over the gene features in the 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], [24], [27], [28]. 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 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 removing features with the least weights [27], [29], [30], [31].

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 in the search space reduction stage to generate a candidate gene 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 comprehensively compared our proposed MGRFE with the feature selection methods from 20 representative studies on 19 benchmark gene expression datasets including multiclass and imbalanced datasets. The comparison results show that our method outperforms most of state-of-the-art feature selection algorithms. MGRFE can select a smaller gene subset but yield the same or higher classification accuracy than other algorithms. Furthermore, the specific biomedical relevance of the selected genes to the related cancer phenotypes has also been verified by text mining. 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) [32], Prostate (Pros) [33], acute lymphoblastic leukemia (ALL; subdivided into four subtypes based on different phenotypes) [34], central nervous system embryonal tumor (CNS) [35], Lymphoma (Lym) [36], Adenoma (Adeno) [37], Colon [38], Leukaemia (Leuk) [39], Myeloma (Mye) [40], Gastric (Gas) [41], and Gastric1/Gastric2 (Gas1/Gas2) cancer [42] as well as type 1 diabetes (T1D) [43], and Stroke [44]. 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) [45] and mixed lineage leukemia (MLL) [46] and one binary dataset of acute lymphoblastic leukemia and acute myeloid (ALL_AML). Many previous experiments are conducted on these three datasets [45], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61]. 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 on 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 performed 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 GaRFE acting as the feature selection unit in every layer. GaRFE 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, GaRFE 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 which 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.

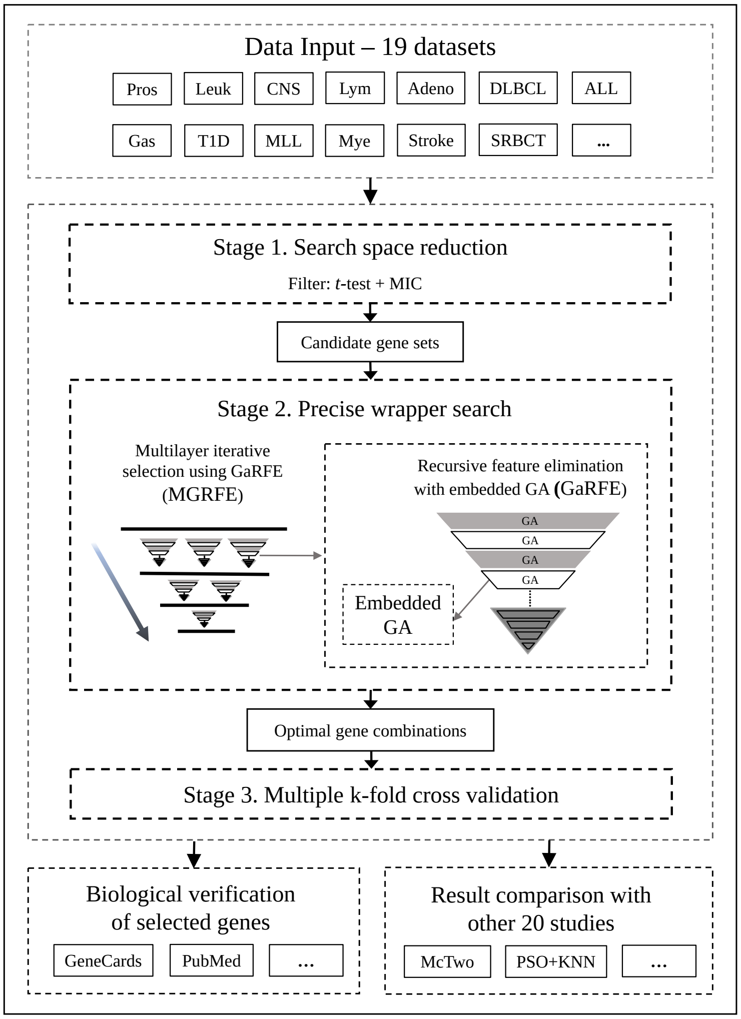
Data Input – 19 datasets
Pros
Leuk
CNS
Lym
Adeno
DLBCL
ALL
Gas
T1D
MLL
Mye
Stroke
SRBCT
...
Stage 1. Search space reduction
Filter: t-test + MIC
Candidate gene sets
Stage 2. Precise wrapper search
Multilayer iterative selection using GaRFE (MGRFE)
Recursive feature elimination with embedded GA (GaRFE)
GA
GA
GA
GA
Embedded GA
Optimal gene combinations
Stage 3. Multiple k-fold cross validation
Biological verification of selected genes
GeneCards
PubMed
...
Result comparison with other 20 studies
McTwo
PSO+KNN
...

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 GaRFE process. GaRFE, 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 GaRFE 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 GaRFE. 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 GaRFE. 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

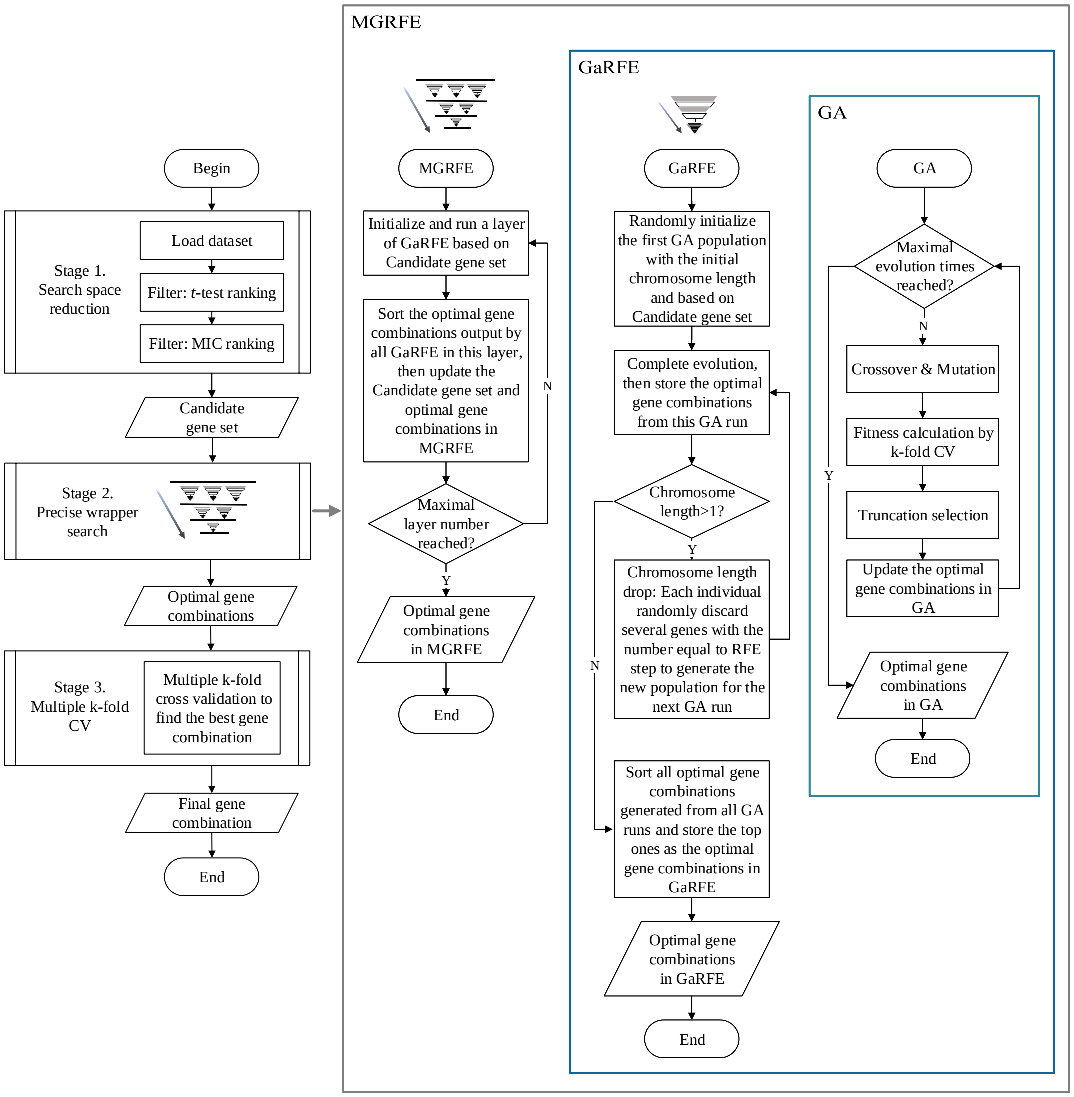
Begin
Stage 1.
Search space reduction
Load dataset
Filter: t-test ranking
Filter: MIC ranking
Candidate gene set
Stage 2.
Precise wrapper search
Optimal gene combinations
Stage 3.
Multiple k-fold CV
Multiple k-fold cross validation to find the best gene combination
Final gene combination
End
MGRFE
MGRFE
Initialize and run a layer of GaRFE based on Candidate gene set
Sort the optimal gene combinations output by all GaRFE in this layer, then update the Candidate gene set and optimal gene combinations in MGRFE
Maximal layer number reached?
N
Y
Optimal gene combinations in MGRFE
End
GaRFE
GaRFE
Randomly initialize the first GA population with the initial chromosome length and based on Candidate gene set
Complete evolution, then store the optimal gene combinations from this GA run
Chromosome length>1?
Y
N
Chromosome length drop: Each individual randomly discard several genes with the number equal to RFE step to generate the new population for the next GA run
Sort all optimal gene combinations generated from all GA runs and store the top ones as the optimal gene combinations in GaRFE
Optimal gene combinations in GaRFE
End
GA
GA
Maximal evolution times reached?
N
Y
Crossover & Mutation
Fitness calculation by k-fold CV
Truncation selection
Update the optimal gene combinations in GA
Optimal gene combinations in GA
End

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: GaRFE and embedded modified GA.

variable-length integer-coding technique for the chromosome in a GA individual, and each individual has 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 [62], which simply ranks all individuals and selects the top ones to form the next generation. 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):

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

, 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 Naive Bayes classifier (NBayes) [63]. 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 GaRFE also use the above-mentioned sorting rule to rank different gene combinations.

Recursive feature elimination with embedded GA:

GaRFE 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 GaRFE. In every iteration layer, the GaRFE 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. 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, our proposed MGRFE is comprehensively compared with the state-of-the-art algorithms of gene selection on 19 benchmark datasets. Additionally, the biological verification of the selected genes is discussed.

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 stage. Moreover, MGRFE achieves stable classification performance in 10 repetitions of 10-fold CV as depicted in Fig. 3. Besides, we also validated the selected gene features of Leuk, Gas1 and Gas2 on independent datasets in the "S8" section of Supplementary Material.

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]

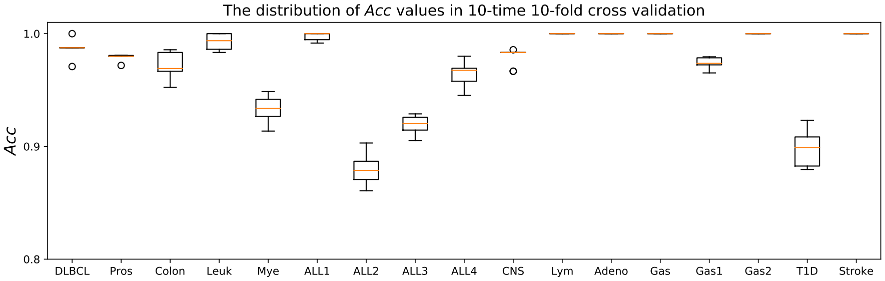

Fig. 3. The distribution of Acc values in 10-time 10-fold CV for the selected gene combinations of 17 datasets in Dataset One.

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. For a more intuitive comparison, Fig. 4 offers the line chart of maximal Acc achieved by the two algorithms on 17 datasets, where MGRFE obviously outperforms McTwo with a higher Acc line. 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 obviously. 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. Fig. 5 offers three instances of GaRFE processes in the first layer of MGRFE for these datasets, where we can notice that the Acc values of the best GA individuals are kept at 1.0 in the majority of gene number ranges and begin to drop only 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. 6. 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. These data confirm that MGRFE has high classification stability.

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.

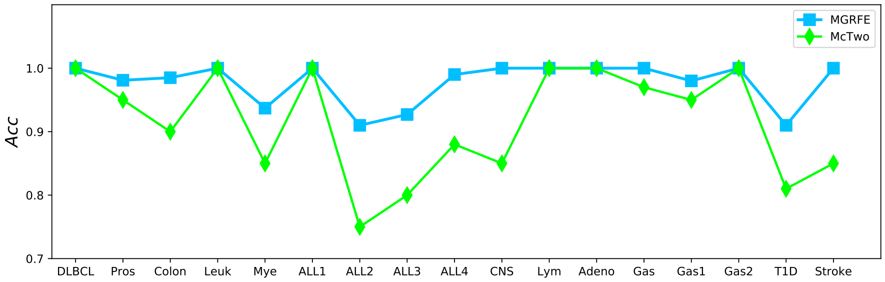

Fig. 4. The line plots of overall maximal accuracy for MGRFE and McTwo on 17 datasets in Dataset One.

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

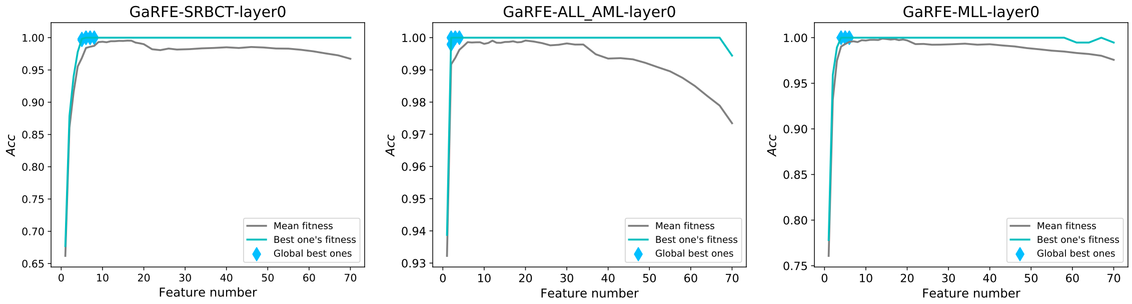

Fig. 5. Three typical GaRFE processes on the three benchmark datasets from Dataset Two.

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

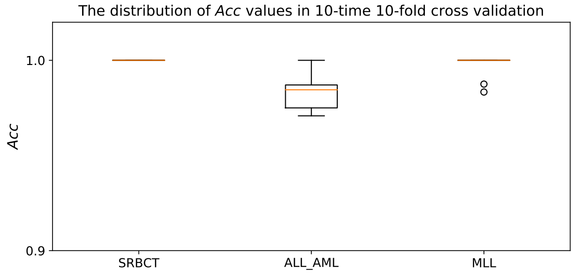

Fig. 6. The distribution of Acc values in 10-time 10-fold CV for the selected gene combinations in Dataset Two.

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.* [45] have applied an artificial neural network (ANN) and selected 96 genes to achieve 100% *Acc*. Tibshirani *et al.* [58] have used the nearest shrunken centroid-based method (NSC) and achieved 100% *Acc* by means of 43 genes. Fu and Fu-Liu [49] employed support vector machine (SVM)-RFE and achieved 100% *Acc* by means of 19 genes. Pal *et al.* [55] 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.* [54] 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.* [56] 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 [49] have achieved 100% train *Acc* by means of 19 genes via SVM-RFE. Yang *et al.* [60] 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.* [54] have selected two genes to reach 100% CV *Acc* based on improved binary PSO. Dashtban and Balafar [28] 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.* [56] have selected four genes with 100% train and test *Acc* based on SFS, LDA, and NCC. Mohamad *et al.* [54] have selected four genes with 100% CV *Acc* based on improved binary PSO. Dashtban and Balafar [28] 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 Biological inferences of the genes selected by MGRFE

The genes selected by MGRFE also have close relevance to the phenotypes in gene expression datasets. The genes selected by MGRFE with the 100% 5-fold CV *Acc* on datasets Leuk, Gas, and ALL1 are only two genes, three genes, and one gene, 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 9th July, 2018. The literature-mining results on these three datasets are shown in Table 7. 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 ref. [65], 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 with a dynamic-length chromosome. MGRFE is aimed at selecting minimal discriminatory gene features associated closely with a phenotype. MGRFE is designed as a complementary feature selection algorithm for high-dimensional data especially for gene expression data analysis. Through the comprehensive comparison on 19 benchmark datasets with other state-of-the-art algorithms, MGRFE can be successfully applied to cancer diagnosis and further biomedical research.

The main innovation of MGRFE is an effective combination of the advantages of evolution calculation of the embedded GA with an explicit feature reduction of the RFE process in the basic feature search unit GaRFE. Therefore, our developed MGRFE can perform explicit feature elimination along with the evolution optimization search and achieve

TABLE 4
Performance comparison among the methods on the SRBCT dataset

Experiments	Methods	Genes				CV *Acc*(%)				Train *Acc*(%)	Test *Acc*(%)
Khan *et al.* (2001) [45]	ANN	96				-				100	100
Tibshirani *et al.* (2002) [58]	NSC	43				-				100	100
Fu and Fu-Liu (2005) [49]	SVM-RFE	19				-				100	100
Yang *et al.* (2006) [60]		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) [55]	FSMLP+NERFCM	7				-				100	100
Li and Shu (2009) [53]	KLLE+LLE+PCA	20				-				100	100
Ji *et al.* (2011) [51]	PLSVIP	24				-				100	100
	PLSVEG	15				-				100	100
Mohamad *et al.* (2011) [54]	IBPSO	6				100				-	-
Zainuddin and Ong (2011) [61]	MSFCM+WNN	10				10CV				-	-
						100					
Lee *et al.* (2011) [52]	AGA+KNN	14				-				100	100
Sharma *et al.* (2012) [56]	SFS+LDA with NCC	4				-				100	100
	SFS+Bayes classifier	4				-				100	90
	SFS+NNC	4				-				100	95
Chen *et al.* (2014) [64]	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) [28]	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.

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], [28], [66], [67], [68]. 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 [28], 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 [27], MGRFE has better performance. SVM-RFE ranks all gene features by the weight vector from SVM with a linear kernel and removes the features with the smallest weights recursively. Nonetheless, SVM-RFE has the following three limitations: 1) The weight ranking can not completely reflect gene importance levels; 2) the top-ranked genes do not always form the optimal gene subset; 3) when a gene is removed, it has no opportunity to appear again. By contrast, the RFE process in MGRFE does not rank genes but introduces a random strategy to randomly discard the same number of genes in each individual between two GA runs, and when a gene is removed from an individual, it can still be present in other individuals. This arrangement leads to greater tolerance and stability. Besides, MGRFE selects gene combinations based on the sorting rule of a GA individual rather than selecting a gene individually, which has an advantage in finding the optimal gene subset. 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 Accs, respectively [49]. But MGRFE selected only five and two genes to attain 100% Accs in 5-fold CV for the same datasets. As for the selection operator of our embedded GA, the widely used roulette wheel selection [62] is inferior to our currently implemented truncation selection technique in this gene selection problem. The fitness gaps between different GA individuals are usually

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) [49]	SVM-RFE	4				-				100	97.06
Yang *et al.* (2006) [60]		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) [57]	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) [59]	Probabilistic mechanism	-				SVM	KNN			-	-
						97.38	98.21				
Ji *et al.* (2011) [51]	PLSVIP	9				-				100	100
	PLSVEG	8				-				100	100
Mohamad *et al.* (2011) [54]	IBPSO	2				100				-	-
Zainuddin and Ong (2011) [61]	MSFCM+WNN	10				10CV				-	-
						98.61					
Chandra and Gupta (2011) [48]	RNBC	-				10CV				-	-
						RNBC	NBC	KNN			
						94.29	84.29	85.71			
Kumar *et al.* (2012) [50]	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) [28]	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) [60]		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) [54]	IBPSO	4				100				-	-
Chandra and Gupta (2011) [48]	RNBC	-				10CV				-	-
						RNBC	NBC	KNN			
						87.14	80	68.57			
Sharma *et al.* (2012) [56]	SFS+LDA with NCC	4				-				100	100
	SFS+Bayes classifier	4				-				100	100
	SFS+NNC	4				-				100	93
Chen *et al.* (2014) [64]	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**				

TABLE 7
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)

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 algorithms from 20 other studies, 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 deal with the problem of minimal discriminatory feature selection in high-dimension datasets, 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 data. 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 phenotypes and developing potential early detection and molecularly-targeted therapies for cancer diseases. Moreover, 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 61472158, 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] Z. Liu, L. S. Magder, T. Hyslop, and L. Mao, "Survival associated pathway identification with group l p penalized global auc maximization," *Algorithms for Molecular Biology*, vol. 5, no. 1, p. 30, 2010.

[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] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] T. Blickle and L. Thiele, "A comparison of selection schemes used in genetic algorithms," 1995.

[63] 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.

[64] 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.

[65] L. Wong, "Lecture 4: Gene expression analysis," 2012.

[66] 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

[67] 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

[68] 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. His current research interests include machine learning and bioinformatics.

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.

Supplementary For 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

S1. The datasets used in this study

This study used total 19 benchmark microarrays subdivided into two large Datasets to validate the performance of our proposed MGRFE. In Tables 1 and 2, we provide the brief description of each dataset in Dataset One and Dataset Two.

S2. Pseudocode of the proposed MGRFE

In the main manuscript, we provided the flow chart of the proposed MGRFE in Fig. 2. Here, we supplement the pseudocodes of our methodology. Pseudocode 1 describes the complete procedure of MGRFE. Pseudocodes 2 and 3 explain the two key processes of MGRFE: GaRFE and embedded GA.

S3. Implementation notes and Computation time

We implemented the proposed MGRFE in Python version 3.6.0 environment (https://www.python.org/) on a common laptop computer with Intel(R) Core(TM) i5-4210U CPU and 8G memory. The Python SciPy package version 0.19.0 [3] was involved in the *t*-test process, and minepy package version 1.2.0 [4] was used to perform the MIC calculation. Some parameter settings about MGRFE: 1) the evolution iteration number of embedded GA was dynamically set to 1 to 3 (smaller iteration number used for larger chromosome length to save time); 2) the reduced feature number between two GA runs, the RFE step, was dynamically set to 1 to 3 (larger reduction step used for larger chromosome length to save time); and 3) the iterative layer number of MGRFE being 3 with three, two and

Pseudocode 1: MGRFE: multilayer iterative feature selection using GaRFE

```
Input  : A microarray gene expression data
Output: The optimal gene feature combination
        for phenotype classification
The $t$-test-based gene ranking to generate the
  candidate gene set $G$;
The MIC-based gene ranking to narrow $G$;
Set $GC$, the list of optimal gene combinations
  in MGRFE, to empty;
while the maximal iterative layer number not
  reached do
    Initialize and run a layer of GaRFE
      (Pseudocode 2) based on $G$;
    for each GaRFE do
        Add the returned optimal gene
          combinations to $GC$;
    Sort the optimal gene combinations in $GC$
      and only preserve the top ranked ones;
    Use the genes in the top ranked gene
      combinations in $GC$ to form a reduced $G$;
Multiple k-fold CV on the gene combinations
  in $GC$;
Return the final selected gene combination;
```

- *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.*

TABLE 1
Summary of the 17 binary classification datasets in Dataset One from ref. [1]

ID	Dataset	Samples	Features	Summary
1	DLBCL[1]	77	7 129	DLBCL patients (58) and follicular lymphoma (19)
2	Pros(Prostate)[1]	102	12 625	prostate (52) and non-prostate (50)
3	Colon[2]	62	2 000	tumour (40) and normal (22)
4	Leuk(Leukaemia)[2]	72	7 129	ALL (47) and AML (25)
5	Mye(Myeloma)[3]	173	12 625	presence (137) and absence (36) of focallesions of bone
6	ALL1[1]	128	12 625	B-cell (95) and T-cell (33)
7	ALL2[1]	100	12 625	patients that did (65) and did not (35) relapse
8	ALL3[1]	125	12 625	with (24) and without (101) multidrug resistance
9	ALL4[1]	93	12 625	with (26) and without (67) the t(9;22) chromosome translocation
10	CNS[1]	60	7 129	medulloblastoma survivors (39) and treatment failures (21)
11	Lym(Lymphoma)[1]	45	4 026	germinalcentre (22) and activated B-like DLBCL (23)
12	Adeno(Adenoma)[1]	36	7 457	colon adenocarcinoma (18) and normal (18)
13	Gas(Gastric)[3]	65	22 645	tumors (29) and non-malignants (36)
14	Gas1(Gastric1)[3]	144	22 283	non-cardia (72) of gastric and normal (72)
15	Gas2(Gastric2)[3]	124	22 283	cardia (62) of gastric and normal (62)
16	T1D[3]	101	54 675	T1D (57) and healthy control (44)
17	Stroke[3]	40	54 675	ischemic stroke (20) and control (20)

In Tables 1 and 2, ”Samples” and ”Features” indicate the total sample number and feature number of each dataset, and ”Summary” column describes the sample classes and the related sample numbers in parenthesis.

[1] These datasets were retrieved from http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi.

[2] Colon and Leuk datasets were downloaded from the R/Bioconductor packages *colonCA* and *golubEsets*, respectively.

[3] These datasets were downloaded from https://www.ncbi.nlm.nih.gov/geo/.

TABLE 2
Summary of the 3 classification datasets in Dataset Two from ref. [2]

ID	Dataset	Classes	Samples	Features	Summary
1	SRBCT[1]	4	88	2 308	EWS (29), NHL (11), NB (18) and RMS (25)
2	ALL_AML[2]	2	72	7 129	ALL (47) and AML (25)
3	MLL[3]	3	72	12 582	ALL (24), MLL (20) and AML (28)

[1] SRBCT dataset was downloaded from http://research.nhgri.nih.gov/microarray/Supplement/. This dataset includes 88 samples totally, but five of them are irrelevant and thus only 83 samples were used.

[2] ALL_AML in Dataset Two and Leuk in Dataset One are the same dataset in actual.

[3] MLL dataset was retrieved from http://portals.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?paper_id=63.

Pseudocode 2: GaRFE: recursive feature elimination with embedded GA

```
Input  : Candidate gene set $G$, Maximal
         chromosome length $L$
Output: The optimal gene feature
         combinations in GaRFE
Randomly generate the first GA population $P$
  from $G$ with chromosome length equal to $L$;
Set $GC$, the list of optimal gene combinations
  in GaRFE, to empty;
do
    Execute embedded GA (Pseudocode 3)
      using the population $P$;
    Add the returned gene combinations by
      GA to $GC$;
    if the current chromosome length $> 1$ then
        Chromosome length drop: each
          individual in $P$ randomly discard
          several genes with the number equal
          to the RFE step;
while the current chromosome length $\geqslant 1$;
Sort the optimal gene combinations in $GC$ and
  only preserve the top ranked ones;
Return the optimal gene combinations in $GC$;
```

Pseudocode 3: Embedded GA

```
Input  : GA population $P$, Maximal evolution
         times $T$
Output: Updated $P$, The optimal gene feature
         combinations in GA
Set $GC$, the list of optimal gene combinations
  in GA, to empty;
while the maximal evolution times $T$ not reached
  do
    Perform mutation operator;
    Perform crossover operator;
    Fitness calculation of each GA individual
      by k-fold CV;
    Truncation selection to form the updated $P$;
Sort the GA individuals in $P$ and select the top
  ones to form $GC$;
Return the updated population $P$ and the
  optimal gene combinations in $GC$;
```

one GaRFE processes at each layer is usually enough. In the experiments, we limited the size of the final selected gene combination in each dataset to below 10 genes. According to the experiment records, in each of the 19 microarray datasets, the running time of MGRFE is commonly between 500 seconds (8.33 minutes) and 900 seconds (15 minutes). The running time has included the whole filter screen and later wraper search processes. Additionally, it is well worthy to mention that the final chosen gene subset in each dataset might be already found by the first GaRFE process in the first layer of MGRFE, which just costs 2∼3 minutes. More implementation details and experiment results of MGRFE in the 19 datasets are available at https://github.com/Pengeace/MGRFE-GaRFE.

Because Kar *et al.* also employed an evolutionary-computation method PSO, which is similar to GA, to select minimal informative genes in microarray and provided their program running time records on three datasets SRBCT, ALL_AML, and MLL [2], here, we offer a simple running time comparison between their method and MGRFE. Their PSO-based method cost 2.7956, 2.7906 and 7.1488 hours on SRBCT, ALL_AML and MLL respectively to find their optimal gene subsets. In contrast, MGRFE merely used 10.8230, 9.0108 and 8.8739 minutes respectively in the same three datasets and thus showed much higher converge speed. Moreover, according to Tables 4, 5 and 6 in the main manuscript, the gene subsets selected by MGRFE had smaller sizes but higher classification accuracies compared with Kar *et al.*'s method. We noted that Kar *et al.* didn't employ the filter techniques to cut down the feature search space and thus their binary-coded PSO demanded high time cost to converge when dealing with the thousands of genes in each microarray.

S4. Performance of MGRFE in 10-time 10-fold CV

In the main manuscript, the performance of MGRFE on the two large Datasets in 10-time 10-fold cross validation (CV) are shown in the box-plot form. Here, we supplement the detailed mean accuracies ($Mean\ Acc$s) and standard deviations ($S.D.$s) of MGRFE on all the 19 datasets. In each dataset, 10-fold CV is repeated 10 times based on different random seeds. In each 10-fold CV, the mean accuracy value in 10-fold is calculated and recorded. Then after 10 repetitions of the 10-fold CV, the $Mean\ Acc$ and $S.D.$ of MGRFE in a dataset is calculated from the recorded total 10 mean accuracy values.

TABLE 3
$Mean\ Acc$ and $S.D.$ of MGRFE on 19 benchmark datasets in 10-time 10-fold CV

Dataset	DLBCL	Pros	Colon	Leuk	Mye	ALL1	ALL2	ALL3	ALL4	CNS
$Mean\ Acc$	0.987	0.979	0.971	0.982	0.933	0.998	0.880	0.920	0.963	0.980
$S.D.$	0.007	0.003	0.012	0.007	0.011	0.004	0.013	0.007	0.010	0.007
Dataset	Lym	Adeno	Gas	Gas1	Gas2	T1D	Stroke	SRBCT	MLL	
$Mean\ Acc$	1.000	1.000	1.000	0.974	1.000	0.897	1.000	1.000	0.997	
$S.D.$	0.000	0.000	0.000	0.004	0.000	0.014	0.000	0.000	0.006	

Note that the $Mean\ Acc$ and $S.D.$ of MGRFE in dataset ALL_AML are same as the records in Leuk for these two are the same dataset in actual.

TABLE 4
The gene probes finally selected by MGRFE on the 19 microarray datasets

Dataset	Probe number	Gene probes
DLBCL	3	[X69433_at, Z84497_s_at, M15205_at]
Pros	4	[37639_at, 38634_at, 1909_at, 37537_at]
Colon	6	[Hsa.36952, Hsa.36696, Hsa.94, Hsa.442, Hsa.5226, Hsa.5756]
Leuk	2	[M23197_at, M31523_at]
Mye	7	[35977_at, 33130_at, 31366_at, 34571_at, 38013_at, 1368_at, 41150_r_at]
ALL1	1	[38319_at]
ALL2	8	[37502_at, 39885_at, 1291_s_at, 39408_at, 1838_g_at, 819_at, 31331_at, 39336_at]
ALL3	8	[38907_at, 38478_at, 34284_at, 37693_at, 201_s_at, 34497_at, 37809_at, 41259_at]
ALL4	6	[39631_at, 38119_at, 36795_at, 36873_at, 39905_i_at, 1265_g_at]
CNS	7	[S76475_at, M96739_at, X64624_s_at, X93511_s_at, K01911_at, S78693_f_at, X78565_at]
Lym	3	[GENE3332X, GENE3261X, GENE1191X]
Adeno	1	[D43636]
Gas	3	[225571_at, 236118_at, 237466_s_at]
Gas1	3	[213125_at, 41037_at, 208897_s_at]
Gas2	2	[212344_at, 210766_s_at]
T1D	7	[1566232_at, 215728_s_at, 215612_at, 226585_at, 239474_at, 219870_at, 244223_at]
Stroke	4	[1567009_at, 240084_at, 239389_at, 233835_at]
SRBCT	5	[245330.0, 784257.0, 43733.0, 784224.0, 295985.0]
MLL	3	[38242_at, 37710_at, 1389_at]

S5. The gene probes selected by MGRFE

The gene probes finally selected by MGRFE on all the 19 datasets are listed in Table 4. These differentially expressed genes could be potential biomarker candidates that are useful to related phenotype researches.

S6. The statistically significant genes in *t*-test

In *t*-test, this study adopt the widely used $p = 0.05$ significance threshold to select the differentially expressed genes with *p*-values lower than the threshold. The chosen of *p*=0.05 has also been experimentally validated by the sample distribution condition of 17 binary classification datasets and our final experiment results on these datasets. Table 5 illustrates the number of significant genes with *p*-value less than 0.05 in the *t*-test. From Table 5 we can note that, *p*=0.05 is a relatively accommodative condition on 17 binary-class datasets, which can not only identify the most differentially expressed genes, but also avoid the inappropriately exclusion of too many genes.

S7. Compare other filter methods with the *t*-test and MIC combination

In the feature space reduction stage, the study used *t*-test and MIC for their efficiency and convenience in gene filtering process. The *t*-test has been widely used

TABLE 5
Number of statistically significant features with *t*-test-based *p*-values less than 0.05 on 17 binary classification datasets.

Dataset	DLBCL	Pros	Colon	Leuk	Mye	ALL1	ALL2	ALL3	ALL4
Significant features	2632	5061	594	2449	1720	4387	644	571	1279
Total features	7129	12625	2000	7129	12625	12625	12625	12625	12625
Dataset	CNS	Lym	Adeno	Gas	Gas1	Gas2	T1D	Stroke	
Significant features	334	804	1799	8260	16454	15601	10159	5569	
Total features	7129	4026	7457	22645	22283	22283	54675	54675	

and validated for detecting differentially expressed genes in microarray [5], [6], [7]. But *t*-test has limitation in dealing with multi-class dataset for multivariate *t*-test can't be performed directly. The recently proposed MIC shows excellent performance in detecting a wide range of associations in large datasets including microarray [1], [8], and MIC can cope with multi-class dataset. Thus, we combined *t*-test and MIC to complete the feature screen task.

For the execution order, we perform the *t*-test first and then MIC. By the p=0.05 significance threshold in the *t*-test, we can quickly find the statistically significant genes, which could notably reduce the gene feature range. Besides, it has been noticed that the MIC calculation is kind of time-consuming compared with *t*-test, thus it is suitable to perform *t*-test first to decrease the gene number.

We also compared the performance of other filter method combinations with the *t*-test+MIC combination, the result are shown in Table 6.

Firstly, the combination of first Anova then Fold change (FC), Anova+FC. The experiment was carried on 3 balanced datasets (Adeno, Gas1 and Pros) and 3 imbalanced datasets (DLBCL, Leuk, and CNS) using 5-flod cross validation. For Anova, the *p*-value threshold was also set as 0.05 as in *t*-test. From Table 6, it can notice that with the combination of Anova+FC, the sizes of finally selected genes are 2, 4, and 8 on datasets Adeno, Leuk and CNS, respectivily. But by the *t*-test+MIC combination, simply 1, 2 and 7 genes are needed to achieve the same performance on the 3 datasets. On the rest of 3 datasets, the two filter method combinations have similar performance. Thus, the filter method combination of Anova+FC is little inferior to the combination of *t*-test+MIC in finding the minimal discriminative gene subset.

Secondly, the combination of first "volcano plot" then MIC, Volcano plot+MIC. The "volcano plot" can combine the advantages of *t*-test and fold-change, thus we use the "volcano plot" to replace the *t*-test. The experiments were performed on 2 balanced datasets (Adeno and Pros) and 2 imbalanced datasets (DLBCL and Leuk) by 5-fold cross validation. According to experiment results in Table 6, these two methods select same size of genes and achieve similar performance on all the tested 4 datasets. In the experiments, we noted one defect of "volcano plot" for gene selection in a range of different microarray datasets. When we use the "volcano plot" to selected informative genes, in each microarray dataset, we need to hand-tune the *p*-value threshold in *t*-test and fold-change threshold value in FC to obtain the satisfactory result. For example, on dataset Adeno, there are 894 informative genes have FC value larger or equal to 2; but on dataset Pros, the highest FC value in all genes is just 1.46. Thus, for different datasets, the threshold values in "volcano plot" should be different. In fact, for each of the tested 4 datasets, the threshold values in "volcano plot" have been hand-tuned individually and the finally assigned threshold values vary among different datasets. This situation pose difficulty for building automatically application on a wide range of microarray datasets. In contrast, the *t*-test has the consistent *p*-value setting in all the 17 datasets in experiments and shows more convenience.

To conclude, the filter method combination of *t*-test+MIC are more efficient and convenient than the Anova+FC or Volcano plot+MIC combinations for the feature range reduction task in our study.

S8. Independent validation of selected features

In this section, we validate the selected gene subsets from Leuk, Gas1 and Gas2 on independent datasets using 10-time 10-fold cross validation.

The datasets Leuk and MLL both contain the sample data of leukemia subtypes ALL (acute lymphoblastic leukemia) and AML(acute myeloid). We use the dataset MLL to validate the selected genes in Leuk. On Leuk the selected gene probes are [*M23197*,*M31523*], and the related genes are [*CD33*, *TCF3*]. For these two genes, the coresponding gene probes in MLL are [*32874_at*, *36802_at*, *1373_at*, *1374_g_at*]. Thus, we test the classification performance of the obtained 4 gene probes on the ALL and AML samples in MLL dataset.

TABLE 6
The performance comparison among different filter method combinations by 5-fold cross validation

Filter Methods	Dataset	Genes	Sn	Sp	Acc	Avc	MCC	AUC
t-test+MIC	Adeno	1	1.0	1.0	1.0	1.0	1.0	1.0
	Gas1	3	0.986	0.973	0.980	0.980	0.961	0.99
	Pros	4	0.980	0.982	0.981	0.981	0.963	0.98
	DLBCL	3	1.0	1.0	1.0	1.0	1.0	1.0
	Leuk	2	1.0	1.0	1.0	1.0	1.0	1.0
	CNS	7	1.0	1.0	1.0	1.0	1.0	1.0
Anova+FC	Adeno	2	1.0	1.0	1.0	1.0	1.0	1.0
	Gas1	3	0.987	0.973	0.980	0.980	0.960	0.979
	Pros	4	0.980	0.982	0.981	0.981	0.963	0.982
	DLBCL	3	1.0	1.0	1.0	1.0	1.0	1.0
	Leuk	4	1.0	1.0	1.0	1.0	1.0	1.0
	CNS	8	1.0	1.0	1.0	1.0	1.0	1.0
Volcano plot+MIC	Adeno	1	1.0	1.0	1.0	1.0	1.0	1.0
	Pros	4	0.980	0.982	0.980	0.981	0.963	0.968
	DLBCL	3	1.0	1.0	1.0	1.0	1.0	1.0
	Leuk	2	1.0	1.0	1.0	1.0	1.0	1.0

The datasets Gas1 and Gas2 are both gastric cancer data but related to different gastric cancer subtypes. Gas1 is about non-cardia gastric cancer, while Gas2 is about cardia gastric cancer. These two datasets are both from ref. [9] and have the same gene probes as features. The gene probes selected by MGRFE on Gas1 are [*215380_s_at*, *221928_at*,*214746_s_at*], and the gene probes selected on Gas2 are [*210125_s_at*,*206361_at*]. For Gas1 and Gas2, we both validated the selected gene probe subset on the other dataset.

From Table 7, it can be noted that the selected gene features on Leuk achieved satisfying performance on MLL. The obtained accuracy is 0.963, just slightly lower than the classification accuracies achieved within the datasets MLL or Leuk. The gene subset selected in Gas1 and Gas2 also showed acceptable performance on the other dataset. The different gastric cancer subtypes could partially account for the performance decrease in these two datasets.

TABLE 7
Independent validation of selected features from dataset Leuk, Gas1, and Gas2 by 10-time 10-fold CV.

Feature test on	Feature from	Sn	Sp	Acc	Avc	MCC	AUC
MLL	**Leuk**	0.963	0.96	0.963	0.962	0.934	0.993
MLL	MLL	1.0	1.0	1.0	1.0	1.0	1.0
Leuk	Leuk	0.99	1.0	0.993	0.995	0.987	1.0
Gas1	Gas1	0.984	0.965	0.974	0.974	0.952	0.989
Gas1	**Gas2**	0.917	0.929	0.923	0.923	0.853	0.967
Gas2	**Gas1**	0.933	0.827	0.880	0.880	0.774	0.973
Gas2	Gas2	1.0	1.0	1.0	1.0	1.0	1.0

For dataset MLL, only ALL and AML leukemia samples are taken into consideration in this experiment to stay consistent with Leuk dataset.

References

[1] 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.

[2] 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.

[3] E. Jones, T. Oliphant, and P. Peterson, "Scipy: Open source scientific tools for python," 2014.

[4] D. Albanese, M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman, and C. Furlanello, "minerva and minepy: a c engine for the mine suite and its r, python and matlab wrappers," *Bioinformatics*, vol. 29, no. 3, pp. 407–408, 2013. [Online]. Available: ⟨GotoISI⟩://WOS:000314892000022

[5] 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

[6] 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

[7] 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

[8] 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.

[9] 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.

Dear Editor,

Enclosed please find our substantially revised manuscript “MGRFE: multilayer recursive feature elimination based on embedded genetic algorithm for cancer classification”. In this revised manuscript, we have carefully addressed all the concerns by the two reviewers. We greatly appreciate the Referee’s comments on our manuscript. The following is our point-by-point response to each comment of the reviewers. Furthermore, I would like to take this opportunity to thank you for handling the review of our manuscript.

Our responses to the review comments are in blue.

Sincerely yours,
Ying Li, Ph.D.

College of Computer Science and Technology
Jilin University
Qianjin Street 2699, Changchun, Jilin 130012, P.R.China
Phone: 86-13504319660 (Mobile)

Response to Editor Comments

****

Editor Comments

Associate Editor
Comments to the Author:
This manuscript was reviewed by two experts.

Both of them have concerns on comparison with other methods, ways of computational experiments, and statistical tests.

Furthermore, one reviewer recommends that the type of the paper should be changed to regular one. And, I agree with this opinion.
(For page length/paper type issue, please do not ask me instead ask to the editorial staff or the editor in chief.)

Based on these points, I recommend the authors to revise the manuscript with taking all reviewers' comments into account.
****

Response: Thanks for providing us the opportunity to revise the manuscript. The revised version considering all remarks of the reviewers has just been submitted. We have substantially revised the previous manuscript and made great efforts in responding to the review comments. In addition, we have changed the type of the paper in the online system to a regular research one as you suggested. The response to each comment of the reviewers in the detail are provided as follows.

Response to Comments of Reviewer 1

Reviewer: 1

Recommendation: Author Should Prepare A Minor Revision

Comments:
A multilayer recursive feature elimination technique based on embedded genetic algorithm for cancer classification has been presented. The authors have proposed a hybrid technique comprising both filter and wrapper methods for gene subset selection. The work is interesting and the manuscript is well organized.

1. In the introduction section author has mentioned the phrase "lack an explicit decline of the feature number". The particular phrase is not clear. Please elaborate and explain clearly the lacuna of swarm intelligence based gene selection approach.

Response: Thank you for this comment. In the revised manuscript, we have added more explanations on the phrase "lack an explicit decline of the feature number" to account for the lacuna of swarm intelligence based gene selection approaches.

- The sentence containing the mentioned particular phrase in the "Introduction" section is:

"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."

- We have added more explanations in the revised manuscript after the above sentence:

"The feature number only changes in the randomized evolution operation like mutation and crossover. Thus, these methods lack the precise control over the gene features in the individuals and can not explicitly remove genes to decrease the feature number."

2. In algorithm 1, it has been mentioned to sort the optimal gene combination in GC and to preserve the top ranked genes. On what basis the top ranked gene would be sorted? For sorting what procedure is used?

Response: Thanks for your comment. For the first question, we sort the gene combinations based on two metrics: fitness and gene number. The individual with higher fitness is superior. For two individuals with the same fitness values, the one with a smaller gene number is superior. The fitness of an individual is defined respectively according to different datasets. On imbalanced datasets, the fitness is defined as $\alpha*Acc+(1-\alpha)*Avc$. On balanced datasets, the fitness is simply defined as accuracy *Acc* (more explanations of fitness definition are available in the section 2.3.2 “Stage 2: Precise wrapper search” in the revised manuscript).

For the sorting procedure, we use the TimSort [1] method, which is the default sorting procedure of Python: TimSort is a hybrid stable sorting algorithm derived from merge sorting and insertion sorting, which has O(n log n) time complexity for the worst case and O(n) for the best case scenarios. It performs well on the complex real-world data and can speed up the gene combination sorting process in our experiments.

3. In the search space reduction stage, it has been mentioned that top 1000 genes have been selected by a threshold of 0.05 in *t*-test technique and thereafter MIC has been applied on the 1000 genes to re-rank them. Is there any particular reason of selection 0.05 value as threshold? Is there any mathematical reason for selecting particularly this value for threshold in *t*-test? Or it has been selected experimentally and any other value can also be chosen? Clarify in detail.

Response: Thank you for this comment. The selection 0.05 value as threshold is on the basis of mathematical and statistically theories. It is worth noting that the standard level of significance used to justify a claim of a statistically significant effect is 0.05 and there are many theories to account for the use of 0.05 in denotation statistical significance, which can trace back to the influence of *Statistical Methods for Research Workers* by R.A. Fisher [2]. For better or worse, the term statistically significance has become synonymous with $p \leq 0.05$. Meanwhile, it is convenient to take the value of this threshold as a limit in judging whether a deviation ought to be considered significant or not. Overall, in the majority of analyses, an *p* of 0.05 is used as the cutoff for significance, guaranteeing the feasibility of most researches and studies.

In addition, the reason why our proposed method MGRFE selected 0.05 as threshold in *t*-test technique is based on both theoretical and experimental foundations.

a) The p=0.5 threshold is widely adopted in *t*-test among researchers owing to the characteristics of the selection 0.05 value mentioned above, forming a common standard in research result explain and comparison processes.

b) The threshold of p=0.05 in *t*-test is experimentally selected based on the data distribution condition of 17 binary classification datasets and then validated by the satisfying experiment results on these datasets. **Table 1** illustrates the number of significant genes with *p*-value less than 0.05 according to the *t*-test. From **Table 1** we can note that, p=0.05 is a relatively accommodative condition on 17 binary-class datasets, which can not only identify the most differentially expressed genes, but also avoid the inappropriately exclusion of too many genes.

This **Table 1** becomes the Table 5 in Supplementary Material.

Table 1. Number of statistically significant features with *t*-test-based *p*-value less than 0.05 on 17 binary classification datasets.

Dataset	DLBCL	Pros	Colon	Leuk	Mye	ALL1	ALL2	ALL3	ALL4
Significant features	2632	5061	594	2449	1720	4387	644	571	1279
Total features	7129	12625	2000	7129	12625	12625	12625	12625	12625

Dataset	CNS	Lym	Adeno	Gas	Gas1	Gas2	T1D	Stroke
Significant features	334	804	1799	8260	16454	15601	10159	5569
Total features	7129	4026	7457	22645	22283	22283	54675	54675

4. What is the rationale for using particularly *t*-test first and then MIC? Can any combination of other two filter methods be used in search space reduction task? Clarify in detail. Use any other combination of two filter methods and compare it to the proposed combination of *t*-test and MIC-based search space reduction.

Response: Thanks for your comment.

(1) The *t*-test and MIC method are used in gene filtering process due to their efficiency and convenience. The *t*-test has been widely used to identify differentially expressed genes in gene expression profiles [3-5]. But *t*-test has its limitation on dealing with multi-class dataset. The multi-variate *t*-test can't be performed directly. The recently proposed MIC shows excellent performance on detecting a wide range of associations in large datasets including microarray [6, 7], which can be used for multi-class dataset directly. Thus, we combined *t*-test and MIC to reduce the search space.

(2) Using *t*-test first primarily because that the *t*-test can quickly reduce the gene number by the commonly used p=0.05 significance threshold. Compared to *t*-test, the calculation of MIC is more time-consuming. Therefore, it is reasonable to perform *t*-test first to quickly reduce the gene number, then use MIC to further select the significant gene on a relatively small search space.

(3) The other combination of two filter methods Anova then Fold change (FC) is used in the search space reduction task. We use the combination of first Anova then Fold change (FC) to compare with the combination of first *t*-test then MIC as shown in **Table 2**. The experiment was carried on 3 balanced datasets (Adeno, Gas1 and Pros) and 3 imbalanced datasets (DLBCL, Leuk, and CNS) using 5-fold cross validation. For Anova, the *p*-value threshold was also set as 0.05. From **Table 2**, for the combination of Anova+FC, the sizes of the finally selected genes are 2, 4, and 8 on datasets Adeno, Leuk and CNS, respectively. But for *t*-test+MIC combination used in our proposed MGRFE, the selected 1, 2 and 7 genes can achieve the same performance on the 3 datasets. On the rest of 3 datasets, the two combinations have similar performance. Thus, the filter method combination of Anova+FC is little inferior to the original combination of *t*-test+MIC in finding the minimal discriminative gene subset.

Due to the length restriction of the main manuscript, we added the filter method combination results to the "S7" section in Supplementary Material. The **Table 2** becomes part of the Table 6 in the supplementary section "S7".

Table 2. Comparison of the combination of Anova and Fold change (FC) with the combination of the *t*-test and MIC on 3 balanced datasets (Adeno, Gas1 and Pros) and 3 imbalanced datasets (DLBCL, Leuk, and CNS) using 5-fold cross validation.

Filter Methods	Dataset	Genes	*Sn*	*Sp*	*Acc*	*Avc*	*MCC*	*AUC*
Anova+FC	Adeno	2	1	1	1	1	1	1
	Gas1	3	0.987	0.973	0.98	0.98	0.96	0.979
	Pros	4	0.98	0.982	0.981	0.981	0.963	0.982
	DLBCL	3	1	1	1	1	1	1

	Leuk	4	1	1	1	1	1	1
	CNS	8	1	1	1	1	1	1
***t*-test+MIC**	Adeno	1	1	1	1	1	1	1
	Gas1	3	0.986	0.973	0.98	0.98	0.961	0.99
	Pros	4	0.98	0.982	0.981	0.981	0.963	0.98
	DLBCL	3	1	1	1	1	1	1
	Leuk	2	1	1	1	1	1	1
	CNS	7	1	1	1	1	1	1

5. The researchers have used *t*-test and then MIC. The gene selected after MIC is used in the proposed MGRFE algorithm. In table 5, why the *t*-test-based gene ranking has been compared? MIC based ranking should also be compared.

Response: Thanks for your suggestion. In the Table 1 in our previous manuscript, we listed the *t*-test ranking results for all selected genes on 17 binary datasets (is table 5 a mistake of Table 1?). In the revised manuscript, we have added the MIC-based gene rankings to the Table 1 and the results explanation are also added in the subsection 3.1 "Results on Dataset One". For convenience to review, the *t*-test and MIC-based gene ranking results are listed in **Table 3**.

From the Table 3, it could be noted that: a) The relative positions of selected genes in the two ranking methods are consistent on the most datasets. For example, on the DLBCL dataset, the selected genes are ranked [13, 39, 54] in the *t*-test sorting. Meanwhile, the MIC-based gene rankings for same genes are [8, 24, 52], keeping the same ascending order as in *t*-test. b) The top-ranked genes in the *t*-test are also top-ranked in the MIC ranking. For example, the selected gene on dataset ALL1 is the top one (with ranking 1) in both *t*-test and MIC sorting process. c) The selected discriminatory genes are usually top-ranked by the *t*-test and MIC methods. For 5 of 17 datasets, the top one gene according to the *t*-test appeared in the final selected gene subsets. The *t*-test and MIC could find the informative genes, which are important for the later feature wrapper search process. Therefore, the employed filter techniques are qualified for the search space reduction stage.

Table 3. The sizes of selected gene subsets for *t*-test and MIC-based gene ranking results on the 17 binary classification datasets.

Datasets	**Genes**	***t*-test/MIC-based gene rankings**
DLBCL	3/7129	[13/8, 39/24, 54/52]
Pros	4/12625	[1/1, 15/47, 74/49, 694/618]
Colon	6/2000	[15/6, 58/21, 176/297, 225/80, 240/555, 495/482]
Leuk	2/7129	[4/3, 7/5]
Mye	7/12625	[3/3, 15/103, 83/142, 143/13, 378/217, 404/644, 569/707]
ALL1	1/12625	[1/1]
ALL2	8/12625	[1/80, 52/395, 78/3040, 80/1297, 522/2448, 687/2038, 737/920, 760/1449]
ALL3	8/12625	[4/500, 52/3437, 75/3010, 142/393, 488/443, 510/795, 715/1551, 770/1321]
ALL4	6/12625	[1/2, 6/45, 39/356, 282/226, 535/497, 754/1377]
CNS	7/7129	[9/907, 53/542, 130/620, 131/519, 272/57, 273/454, 520/49]

Lym	3/4026	[4/7, 5/4, 669/135]
Adeno	1/7457	[468/27]
Gas	3/22645	[22/1, 77/32, 306/36]
Gas1	3/22283	[132/74, 248/167, 717/500]
Gas2	2/22283	[38/6, 89/62]
T1D	7/54675	[14/2229, 25/1579, 113/1287, 559/1282, 578/353, 680/426, 978/1728]
Stroke	4/54675	[1/3, 23/115, 129/543, 276/539]

In the “Genes” column, we list the number of selected gene among the total genes in each dataset. In the “*t*-test/MIC-based gene rankings” column, we list the ranking positions for each gene in the *t*-test and MIC processes.

6. Elaborate the significance of '0' ranked gene in *t*-test.

Response: Thank you for this comment. For the index of the first element in a general array in common programming language is '0', so in our previous manuscript, the value '0' is just assigned to the first gene in the *t*-test gene ranking result.

In the revised manuscript, we have changed the initial index to '1' in order to be understood clearly.

7. The comparative results of SRBCT, ALL-AML and ALL have been shown in table 7, 8 and 9. However, the tables are very similar to work in Kar *et al.* [28]. The similar type of comparison should be given for all the dataset used i.e. 19 dataset in the present work**.**

Response: Thanks for your comment. The chief reason why we have based on the tables of Kar *et al.* to make comparison of results is that their tables have been well organized, readable and understandable. In fact, we have added several new items on the tables in the work by Kar *et al.*, so that it is easy to better compare the performance of our proposed method with the previous work on these three datasets.

For similar comparison of other datasets, the results are consistent. We have added two typical selected gene comparison on DLBCL and Prostate as shown in **Table 4**. Many previous work just split the dataset into simply fixed train and test data, and the performance was recorded on the test data. But the performance of MGRFE is in 5-fold cross validation, thus more stable and convictive.

Table 4. The performance comparison between MGRFE and previous methods on DLBCL and Pros dataset.

Dataset	Criteria	EPSO [8]	IBPSO [9]	TS-BPSO[10]	BPSO-CGA[11]	Random forest	mABC [12]	PLS [13]	DGA [14]	MGRFE
DLBCL	*Acc*	1.0	1.0	1.0	-	0.946	1.0	0.930	1.0	1.0
	#Genes	3	1042	2671	-	21	3	20	18	3
Pros	*Acc*	0.990	0.922	0.955	0.937	-	1.0	0.917	0.963	0.981
	#Genes	5	1294	5320	795	-	5	2	14	4

8. The proposed work has also been compared with Kar *et al.* [28] in computational performance. Kar *et al.* have applied a swarm intelligence-based method to the space of all genes. They have not reduced the search space prior to the optimization task. In contrast, the proposed method have

applied MGRFE technique on the reduced search space. The reduce search space have been constructed by *t*-test and then by MIC technique. In my opinion the search space reduction is fixed. It is done once before the application of MGRFE. In that regard, the comparison of computational time would not significant because it has been computed in the reduced search space. The genes outside the reduced search space could carry valuable information towards classification accuracy.

Response: Thank you for this comment.

1) Indeed, the lack of filter preprocessing could increase the running time of Kar *et al.*'s method. But it should be noticed that the recursive feature elimination (RFE) manner has dramatically reduced the running time of MGRFE. Both the method by Kar *et al.* and our proposed algorithm MGRFE employed the time-consuming swarm intelligence methods (i.e. PSO and GA). But we combined the GA with RFE, which notably speeded up the convergence process. It should also be pointed out that the computation time of MGRFE includes the whole filter screen and MGRFE wrapper search process, rather than on a fixed reduced feature search space. Take for example the multi-class SRBCT dataset, with 2308 genes in total, Kar *et al.*'s PSO+KNN method took 2.7956 hours to select the final gene subset. However, MGRFE merely demanded 10.828 minutes totally (including whole filter and MGRFE wrapper process) to select the final gene subset, about 6.46% of the time spent by Kar *et al.*'s method. Besides, the 5 genes selected by MGRFE obtained higher classification accuracy than the 6 genes finally selected by *Kar et al.*'s method on dataset SRBCT.
2) For the filter process has already supplied sufficient discriminatory expressed genes for finding the minimal discriminatory gene subset, MGRFE just ignore the rest genes outside the reduced feature space. Our filter process using the combination of *t*-test and MIC selects significant genes related to the phenotypes which are very valuable for the subsequent feature selection process. There are several thousand to tens of thousands of genes in a microarray, among which the most are irrelevant features. If take all gene features into consideration, the irrelevant and redundant features will disturb the wrapper algorithm to select the minimal informative gene subset with higher time cost.

9. In the Conclusion section, the authors will need to clearly address the research contributions in theory. The research contributions in theory must be fully stated in at least one paragraph.

Response: Thanks for your comment. To better illustrate the contributions of our algorithm, we have added more statements in the section of discussion and conclusion in the revised manuscript.

The chief research contribution in theory of this paper 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 been designed as an evolutionary algorithm using variable length integer encoding approach in a recursive process to deal with the problem of minimal discriminatory feature selection for high-dimension datasets. Meanwhile, through the experimental comparisons with the popular feature selection on diverse datasets, our proposed MGRFE could outperform the most other state-of-the-art algorithms for gene selection on microarray data. Therefore, MGRFE is a novel effective and efficient feature selection algorithm to offer the better options for users.

10. In the Conclusion section, the authors need to fully discuss insightful and practical implications.

Response: Thank you for this comment. We have supplemented more practical significance and applications about MGRFE in the section of discussion and conclusion in the revised manuscript

The presented MGRFE would be useful in medical diagnosis as well as further research. The biological associations between selected genes by MGRFE with the related cancer phenotypes are validated by the literature mining results on PubMed. Therefore, the informative genes selected by MGRFE could be novel biomarker candidates that are useful for better understanding the molecule mechanism related to the phenotypes and developing potential early detection and molecularly-targeted therapies for cancer diseases. Besides, in view of the satisfying performance achieved by MGRFE on various microarray datasets, it is worthy to generalize the proposed MGRFE to more feature selection problems on high-dimensional genomic, proteomic and metabolomics data.

Thank you again for reviewing our paper. We deeply appreciate all your provided constructive suggestions.

Response to Comments of Reviewer 2

Reviewer: 2

Recommendation: Author Should Prepare A Major Revision For A Second Review

Comments:

1. First of all, the paper is described as "Survey/Tutorial," but it appears to describe a claimed original contribution by the authors, namely the MGRFE algorithm. The proposed new algorithm is compared against several existing algorithms. Therefore, if at all the paper is to be published, it should be as a regular research paper, and not as a survey/tutorial paper.

Response: Thank you for this comment. For the type of the paper, it is indeed just like your suggestion and you are insightful. We have changed the type of the paper to a regular research one in the online system.

2. The paper is a mixture of techniques that are by now standard in the world of computational biology. Given a very large number of features, first use some pre-filtering to eliminate perhaps 90% to 95% of the features, and then use recursive feature elimination (RFE) on the remaining features. I could not find any compelling evidence that the proposed approach is superior to the existing methods.

Response: Thank you for this comment. Our proposed approach has showed great efficiency in gene selection through experimental comparisons on large amount of high-dimensional expression datasets with the most state-of-the-art algorithms. The chief innovation of MGRFE is combining the evolutionary strategy of GA with recursive feature elimination method, and the results of MGRFE is better than methods based only on recursive feature elimination method or only genetic algorithm respectively.

The RFE feature selection method has advantage of high execution speed, meanwhile limitation of dissatisfactory classification accuracy for gene selection**.** The swarm intelligence based gene selection approaches have advantage of powerful heuristic search ability in finding optimal gene subset, but shortcomings such as slow convergence speed and existing of irrelevant features in selected feature subset. Through combining the evolution calculation of genetic algorithm and the explicit feature elimination of RFE process, the designed MGRFE can take the best points of these two kind of methods and avoid their limitations, thus could find the minimal discriminatory gene subset with high convergence speed.

According to the performance comparison on the 19 benchmark datasets, the proposed MGRFE could offer smaller informative gene subsets but the same or higher phenotype diagnosis accuracies compared with the currently state-of-the-art feature selection methods.

3. The authors claim to compare their method on 17 data sets. But I did not see any evidence that the finally determined feature set is validated on an independent data set of the same form of

cancer for example. All that the authors have done is five-fold cross-validation within the same data set. Without this sort of validation on an independent data set, the claimed performance figures by themselves are not very persuasive. This is because cross-validation within the same data set does not take into account factors such as batch effect, platform variation, and the like.

Response: Thank you for this comment.

1. There are two main difficulties for validate the determined feature set on independent gene expression data set.
   a) The very limited available benchmark datasets for one typical disease. It is difficult to acquire sufficient and appropriate bio-samples due to high expense of micro-array sample collection and other various factors [15], thus the available benchmark datasets are limited and the sample number in each data set is usually small. For many diseases, we just have one widely used microarray benchmark, like the colon cancer (Colon) [16] and small round blue cell tumors (SRBCT) [17].
   b) For microarray benchmark datasets about same disease, the features and sample classes are usually different. Different microarray datasets usually have different gene features for the gene probes vary among different microarray analysis platform. For example, on the leukemia related datasets of Leuk and MLL used in this study, the gene probes are very different for generating from different microarray platforms.

   Thus, the currently published gene selection algorithms on microarrays are commonly validated within each microarray benchmark dataset.

2. We manage to validate the selected gene subsets of Leuk, Gas1 and Gas2 on independent datasets as shown in **Table 5**. The results of independent feature subset validation are added to the "S8" section in Supplementary Material. The **Table 5** in this response document is the Table 7 in the supplementary section "S8".
   a) The datasets Leuk and MLL both contain the sample data of leukemia subtypes ALL (acute lymphoblastic leukemia) and AML(acute myeloid). First, on Leuk the selected gene probes are [*M23197*, *M31523*]. Second, the genes related to these two probes are [*CD33*, *TCF3*]. Third, for these two genes, the corresponding gene probes in MLL are [*32874_at*, *36802_at*, *1373_at*, *1374_g_at*]. Thus, we test the classification performance of the obtained 4 gene probes for ALL and AML samples in MLL dataset.
   b) The datasets Gas1 and Gas2 are both gastric cancer data but related to different gastric cancer subtypes. Gas1 is about non-cardia gastric cancer, while Gas2 is about cardia gastric cancer. These two datasets are both from ref. [18] and have the same gene probes as features. The gene probes selected by MGRFE on Gas1 are [*215380_s_at*, *221928_at, 214746_s_at*], and the gene probes selected on Gas2 are [*210125_s_at*, *206361_at*]. For Gas1 and Gas2, we both validated the selected gene probe subset on the other dataset.

Table 5. Independent validation of features selected on Leuk, Gas1 and Gas2 by 10-time 10-fold cross validation.

Feature test on	Feature from	*Sn*	*Sp*	*Acc*	*Avc*	*MCC*	*AUC*
MLL	**Leuk**	0.963	0.96	0.963	0.962	0.934	0.993

MLL	**MLL**	1	1	1	1	1	1
Leuk	**Leuk**	0.99	1	0.993	0.995	0.987	1
Gas1	**Gas1**	0.984	0.965	0.974	0.974	0.952	0.989
Gas1	**Gas2**	0.917	0.929	0.923	0.923	0.853	0.967
Gas2	**Gas1**	0.933	0.827	0.880	0.880	0.774	0.973
Gas2	**Gas2**	1	1	1	1	1	1

For dataset MLL, only ALL and AML samples are taken into consideration in this experiment to stay consistent with Leuk dataset.

From **Table 5**, it can be noted that the selected gene features on Leuk achieved satisfying performance on MLL. The obtained accuracy is 0.963, just slightly lower than the classification accuracies achieved within the datasets MLL or Leuk. The gene subset in Gas1 and Gas2 also showed acceptable performance on the other dataset. The different gastric cancer subtypes could account for the performance decrease in these two datasets.

4. The authors' preferred method of genetic algorithms is known to lack theoretical foundations, to be very sensitive to various parameters in the algorithm, and to be extremely time consuming. In contrast, the original paper where RFE was proposed, by Isabel Guyon, used the support vector machine (SVM) which is very fast and for which lots of theoretical results are available. This is another reason for my not being overly enthusiastic about the paper.

Response: Thank you for this comment.

Firstly, for selecting informative gene features in a microarray, the state-of-the art methods are commonly evolutionary-computation based. Although the SVM-RFE method has many theoretical results, the classification accuracy of generated gene subset is likely to be lower than the result of evolutionary-computation based methods. The currently published leading methods of gene selection in microarray are usually base on swarm intelligence algorithms [14, 19, 20].

Secondly, there are several limitations of the RFE method which could not be ignored: a). the weights ranking could not exactly and completely reflect the importance of each gene; b). the top-ranked genes do not mean the best gene subset. Based on our experiment results, genes should be selected in combination but not individually; and c). there is no opportunity for a gene to appear again after being removed. On the contrast, the proposed MGRFE has been well-designed to avoid the above limitations by introducing the evolution computation strategy, thus has more advantages in finding the minimal informative gene subset. Fu and Fu-Liu evaluated SVM-RFE on datasets SRBCT and ALL AML and finally selected 19 and 4 genes to achieve 100% and 97.6% test accuracies, respectively [21]. But MGRFE selected only 5 and 2 genes to attain 100% accuracies in 5-fold CV for the same datasets.

Thirdly, compared with existed GA algorithm, the introduced RFE process has significantly enhanced the convergence speed and reduced running time. Instead of relying on widely used binary encoding, our proposed method utilizes variable length integer encoding in GA and cuts down the encoding length recursively in search process, which could quickly remove the irrelevant and redundant features and converge to the minimal informative feature combination. Kar *et al*. [22] employed the evolutionary computation method PSO to select gene subset on three datasets SRBCT, ALL AML, and MLL. Their PSO-based method cost 2.7956, 2.7906 and 7.1488 hours on the three datasets respectively. In contrast, MGRFE merely used 10.8230, 9.0108 and 8.8739 minutes respectively in the same three datasets. Moreover, the selected gene subsets by MGRFE are smaller but with same or higher classification accuracies compared with Kar *et al.*'s PSO based method.

Fourthly, time complexity is of secondary significance in this issue, what should be prioritized is the discriminating ability of selected gene subset. For each microarray data set, just one running of the feature selection method is enough to generated the informative genes and minimal gene feature subset, which would be used repeatedly in the later classification or clustering applications. Thus, the running time of feature selection method is less important than its ability to locate the discriminatory genes.

5. There are several places where the authors do not appear to be aware of simple statistical facts. For instance, the accuracy is a weighted average of the sensitivity and the specificity. But the authors talk as though they are independent parameters. Equation (1) in the right column of page 1 is too wide.

Response: Thank you for this comment. In the revised manuscript, we have made effort to optimize the expressions about statistical terms. Also, the format of our revised manuscript has been adjusted and we make the equations more organized.

In this study, we used six widely used measurements to compare the method performance: Accuracy (*Acc*), Sensitivity (*Sn*), Specificity (*Sp*), Average accuracy (*Avc*), Matthews Correlation Coefficient(*MCC*), and *AUC* (area under the receiver operating characteristic curve). Sensitivity indicates the proportion of correct prediction on positive samples. Specificity measures the fraction of correctly classified negative samples. Accuracy is the model's overall classification accuracy on two classes.

$$Sn=\frac{TP}{TP+FN}, Sp=\frac{TN}{TN+FP},$$
$$Acc=\frac{TP+TN}{P+N}, Avc=\frac{Sn+Sp}{2},$$
$$MCC=\frac{TP\bullet TN-FP\bullet FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}.$$

6. In Section 2.3.1 the authors use the T-test and MIC to achieve a first-cut reduction in the feature set. I have found that using the so-called "volcano plot" which combines the T-test with a fold-change criterion, works better than just the T-test alone.

Response: Thank you for this comment. Since the "volcano plot" can combine the advantages of *t*-test and fold-change (FC), we use the "volcano plot" to replace the *t*-test and performed comparison experience on 2 balanced datasets (Adeno and Pros) and 2 imbalanced datasets (DLBCL and Leuk) by 5-fold cross validation. The results are shown in **Table 6**. According to experiment results, these two methods select same size of genes and achieve similar performance on all the tested 4 datasets.

In our experiments, we noted one defect of "volcano plot" for gene selection in a number of different microarray datasets. When we use the "volcano plot" to selected informative genes, in each microarray dataset, we need to hand-tune the *p*-value threshold in *t*-test and especially fold-change threshold value in FC to obtain the satisfactory result. For example, on dataset Adeno, there are 894 informative genes have FC value larger or equal to 2; but on dataset Pros, the highest FC value in all genes is just 1.46. Thus, for different datasets, the threshold values in “volcano plot” should be different. In fact, for each of the tested 4 datasets, the threshold values in “volcano plot” have been hand-tuned individually and the finally assigned threshold values vary among different datasets. This situation pose difficulty for building automatically application on the wide range of microarray datasets. In contrast, the *t*-test has the consistent *p*-value setting in all the 17 datasets in experiments and shows more convenience.

Due to the length restriction of the main manuscript, we added the performance comparison results to the “S7” section in Supplementary Material. The **Table 6** becomes part of the Table 6 in the supplementary section “S7”.

Table 6. Performance comparison between filter methods *t*-test and "volcano plot" on 2 balanced datasets (Adeno and Pros) and 2 imbalanced datasets (DLBCL and Leuk) using 5-fold cross validation.

Filter Methods	Dataset	Genes	*Sn*	*Sp*	*Acc*	*Avc*	*MCC*	*AUC*
Volcano plot +MIC	Adeno	1	1	1	1	1	1	1
	Pros	4	0.980	0.982	0.980	0.981	0.963	0.968
	DLBCL	3	1	1	1	1	1	1
	Leuk	2	1	1	1	1	1	1
***t*-test +MIC**	Adeno	1	1	1	1	1	1	1
	Pros	4	0.980	0.982	0.981	0.981	0.963	0.980
	DLBCL	3	1	1	1	1	1	1
	Leuk	2	1	1	1	1	1	1

Thank you so much for your review work. We greatly appreciate your insightful comments.

Reference

1. Auger, N., C. Nicaud, and C. Pivoteau, *Merge Strategies: from Merge Sort to TimSort*. 2015.
2. Fisher, R.A., *Statistical Methods for Research Workers*. 1958: Oliver and Boyd. 66-70.
3. Cui, X.Q. and G.A. Churchill, *Statistical tests for differential expression in cDNA microarray experiments*. Genome Biology, 2003. **4**(4).
4. Lazar, C., et al., *A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis*. Ieee-Acm Transactions on Computational Biology and Bioinformatics, 2012. **9**(4): p. 1106-1119.
5. Shen, Q., 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, 2008. **32**(1): p. 53-60.
6. Ge, R.Q., et al., *McTwo: a two-step feature selection algorithm based on maximal information*

coefficient. Bmc Bioinformatics, 2016. **17**.

7. Reshef, D.N., et al., *Detecting Novel Associations in Large Data Sets.* Science, 2011. **334**(6062): p. 1518.

8. Mohamad, M.S., *An enhancement of binary particle swarm optimization for gene selection in classifying cancer classes.* Algorithms for Molecular Biology, 2013. **8**(1): p. 1-11.

9. Chuang, L., et al., *Improved binary PSO for feature selection using gene expression data.* Computational Biology & Chemistry, 2008. **32**(1): p. 29-38.

10. Li-Yeh, C., Y. Cheng-Huei, and Y. Cheng-Hong, *Tabu search and binary particle swarm optimization for feature selection using microarray data.* Journal of Computational Biology A Journal of Computational Molecular Cell Biology, 2009. **16**(12): p. 1689.

11. Li-Yeh, C., et al., *A hybrid BPSO-CGA approach for gene selection and classification of microarray data.* Journal of Computational Biology A Journal of Computational Molecular Cell Biology, 2012. **19**(1): p. 68.

12. Moosa, J.M., et al., *Gene selection for cancer classification with the help of bees.* Bmc Medical Genomics, 2016. **9**(Suppl 2): p. 47.

13. Li, G.Z., et al. *Partial Least Squares Based Dimension Reduction with Gene Selection for Tumor Classification*. in *IEEE International Symposium on Bioinformatics and Bioengineering*. 2007.

14. Dashtban, M. and M. Balafar, *Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts.* Genomics, 2017. **109**(2): p. 91-107.

15. Dougherty, E.R., *Small sample issues for microarray-based classification.* Comparative and Functional Genomics, 2001. **2**(1): p. 28-34.

16. Alon, U., et al., *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 of the United States of America, 1999. **96**(12): p. 6745-6750.

17. Khan, J., et al., *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.* Nature Medicine, 2001. **7**(6): p. 673-679.

18. Wang, G.S., et al., *Comparison of Global Gene Expression of Gastric Cardia and Noncardia Cancers from a High-Risk Population in China.* Plos One, 2013. **8**(5).

19. Han, F., et al., *A Gene Selection Method for Microarray Data Based on Binary PSO Encoding Gene-to-Class Sensitivity Information.* Ieee-Acm Transactions on Computational Biology and Bioinformatics, 2017. **14**(1): p. 85-96.

20. Motieghader, H., et al., *A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata.* Informatics in Medicine Unlocked, 2017. **9**: p. 246-254.

21. Fu, L.M. and C.S. Fu-Liu, *Evaluation of gene importance in microarray data based upon probability of selection.* Bmc Bioinformatics, 2005. **6**.

22. Kar, S., K. Das 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, 2015. **42**(1): p. 612-627.