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.

$$Sn = \frac{TP}{TP+FN}, Sp = \frac{TN}{TN+FP},$$

$$Acc = \frac{TP+TN}{P+N}, Avc = \frac{Sn+Sp}{2},$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}.$$

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.