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 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:

The manuscript was reviewed by the original reviewers.

Although Reviewer 1 is satisfied with the revised version, Reviewer 2 gives very critical comments.

Therefore, I recommend the authors to revise the manuscript with taking all comments into account.

Since I understand that giving theoretical justification is difficult,

it is enough to give some discussions.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**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. The response to each comment of the reviewers in the detail are provided as follows.

**Response to Comments of Reviewer 1**

---------------------------------------------------------------------------

Reviewer: 1

Recommendation: Accept With No Changes

Comments:

The revisions are satisfactory.

**Response:** Thank you for reviewing our paper. We deeply appreciate all your provided constructive suggestions.

**Response to Comments of Reviewer 2**

---------------------------------------------------------------------------

Reviewer: 2

Recommendation: Reject

Comments:

The selection of highly informative genes in cancer patients is a standard problem with many techniques in existence. The paper presents yet another approach based on an embedded genetic algorithm. In my previous review I had raised a number of queries, which have essentially been dismissed by the authors in their revised version. My queries have NOT been addressed satisfactorily.

My original comment 3 is that there is no validation on an independent data set. The authors state in their rebuttal that "Thus, the currently published gene selection algorithms on microarrays are commonly validated within each microarray benchmark dataset."

I am sorry to say that this is incorrect. I have published several papers in computational cancer biology, and ALL of them had validations on independent data sets. I am not persuaded by the authors' argument.

"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."

This is PRECISELY the reason why validation on an independent data set is so crucial. It is true that two different databases of the same form of cancer may have different genes under study. The way to handle this is to study only those genes that are common to both databases. One can also convert microarray values to Z-scores by subtracting the sample mean and dividing by the sample variance. The authors don't even try to do this.

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

This is not correct. The authors are simply trying to justify whey they did not do any validation on an independent dataset.

If they have managed to do cross-validation on another dataset for leukemia, then that should be in the main paper, not in the supplementary material.

**Response**: Thank you for this comment.

According to your suggestion, we have totally added 10-group cross-validation experiments on independent datasets (**Table 1**). For each experiment, firstly, the selected gene probe features from the first dataset were transformed into the official gene symbols; secondly, the obtained gene symbols were transformed into the selected gene probe Ids in the second dataset; thirdly, a kind of classifier were used to perform 10 times *k*-fold cross validation using the samples and selected gene probe features on the second dataset; and fourthly, the obtained performance of Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF) were recorded.

**Table 1**. Independent validation of selected gene features by MGRFE with 10-time *k*-fold cross validation.

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

On the tested datasets with sample number greater than 50, 10-time 10-fold cross validation were performed with different random seeds. Meanwhile, 10-time 5-fold cross validation were performed on datasets with samples less than 50.

NB, SVM and RF represent Naive Bayes, Support Vector Machine and Random Forest classifiers, respectively.

The bold face values denote the highest performance achieved by classifiers.

On tested datasets GSE56315 and GSE2604 with gene features from DLBCL and ALL1 respectively, NB, SVM and RF classifiers all achieved 100% accuracy in each test. In particular, there are only 14 samples totally on GSE2604, which means the classifiers were trained on merely about 10 samples in each 5-fold cross validation. Thus, the selected unique gene *CD3D is* one ideal discrimination for B-cell acute lymphoblastic leukemia (ALL) and T-cell ALL. Except the three tested datasets of GSE8511, GSE2685 and GSE8514 with samples less than 50, the prediction accuracy of three classifiers are above 0.9 on all other datasets in cross validation. The comprehensive independent validation results proved that the selected genes features by MGRFE in each dataset have strong association with the disease phenotype.

My comment 4 was that their method lacked theoretical justification and compared it to SVM-RFE. Here again the authors simply explain away my objection. They say that their GA (genetic algorithm) works faster than that of Kar et al. That was not my point at all.

In short, I believe that the authors have not adequately addressed my previous comments. Without either theoretical justification or validation on independent datasets, there is very little merit in the paper.

**Response**: Thank you for this comment.

----------------------------------------------------------------------------------------------------------------------

# Previous Responses:

1. 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.
2. 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 [1], 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) [2] and small round blue cell tumors (SRBCT) [3].
3. 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.

1. We manage to validate the selected gene subsets of Leuk, Gas1 and Gas2 on independent datasets as shown in **Table 2**. The results of independent feature subset validation are added to the “S8” section in Supplementary Material. The **Table 2** in this response document is the Table 7 in the supplementary section “S8”.
2. 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.
3. 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. [4] 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 2.** 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 2**, 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.

1. 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 [5-7].

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 [8]. 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*. [9] 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.

**Reference**

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

2. 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.

3. 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.

4. 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).

5. 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.

6. 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.

7. 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.

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

9. 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.