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 four reviewers. We greatly appreciate the Referee’s comments on our manuscript. The following is our point-by-point response to each comment of the four reviewers. Furthermore, I would like to take this opportunity to thank you for handling the review of ourmanuscript.

*Our responses to the review comments are in blue and italic.*

Sincerely yours,

Ying Li, Ph.D.

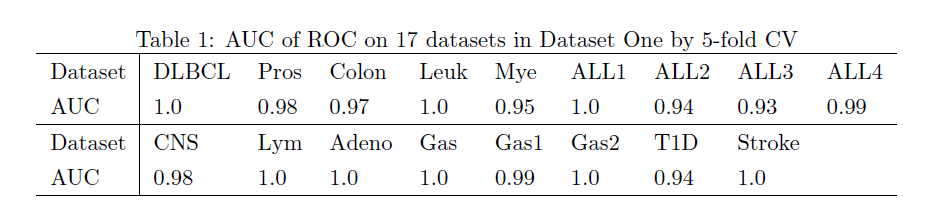
---------------------------------------------------------------------------

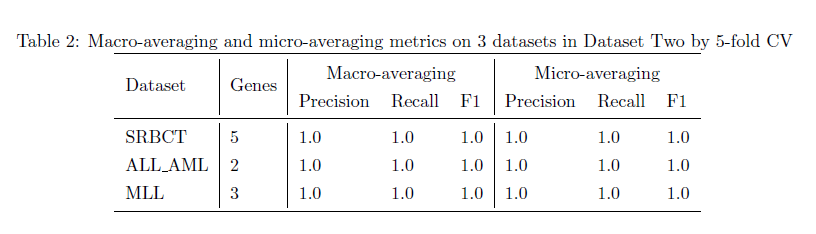
----------------------- REVIEW 1 ---------------------

------------------ Overall evaluation ------------------

In this manuscript, authors suggest the algorithm related to cancer classification. They appropriately described the design and result of their computation, although the evaluation was not extensive enough.

***Response****: Thanks for this comment. To better evaluate our method, we further calculate the AUC values on 17 binary datasets in Dataset One, and micro-averaging and macro-averaging metrics on 3 datasets in Dataset Two by 5-fold cross validation, which are listed in the following Table 1 and 2.*



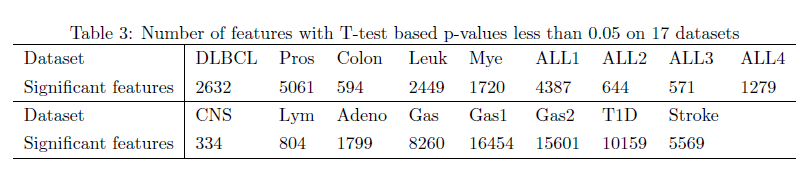


Minor comment

- there is no indication regarding the T-test methodology used for logistic classification and the level of significance regarding to p-value resulting after the T-test.

***Response****: Thanks for this comment. In our proposed method, the T-test is used in the search space reduction stage to cut down the feature dimension. After the T-test, all features will be sorted according to the ascending order of their p-values, then the top significant features with p-values less than 0.05 are preserved. And the upper limit of kept features after the T-test is 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 are kept.*

*We list the number of significant genes according to the T-test on the 17 datasets for logistic classification in Table 3. For datasets of Colon, ALL2, ALL3, CNS and Lym, all features with p-values less than 0.05 are preserved and for the rest datasets only the top 1000 features according to the ascending order of their p-values are kept. Furthermore, after the T-test filter, MIC based feature selection will be carried if the preserved features are more than 500.*



----------------------- REVIEW 2 ---------------------

------------------ Overall evaluation ------------------

This paper presents a gene selection methodology based on genetic algorithm followed by recursive feature elimination. They validated their methodology using various public benchmark datasets and showed improvements compared to the alternatives. However, this reviewer’s concerns about this paper are as follows:

Major Points:

1) If they modified the original embedded genetic algorithm as briefly described in page 7, the additional details and the effect of the modification need to be addressed. The authors also should cite the paper they referenced.

***Response****: Thank you for this comment. We have added more detailed explanations of our proposed embedded modified genetic algorithm in the sections of method and discussion in the revised manuscript.*

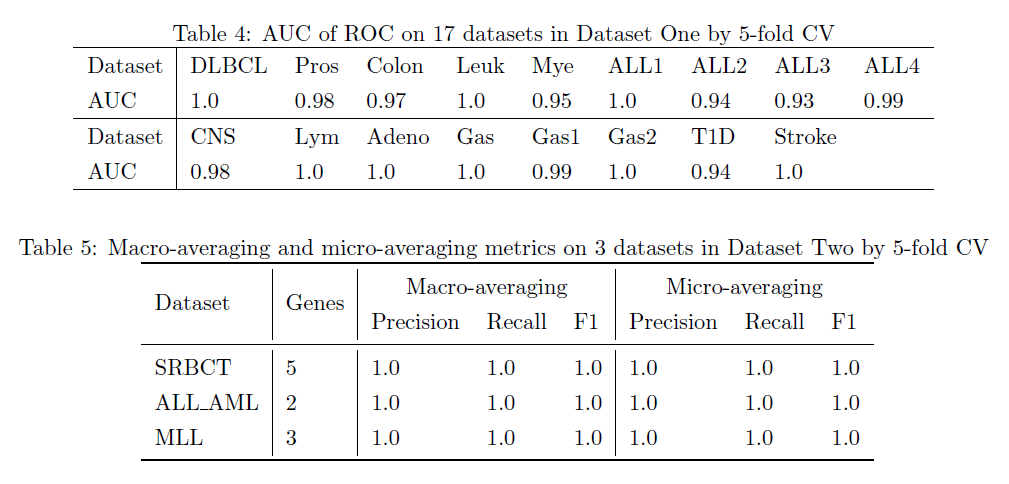
*Genetic algorithms are probabilistic search algorithms inspired by the natural evolution [1] which include genetic operators of mutation, crossover and selection and binary encoding is the most common used solution encoding method for GA and other evolution algorithms [2]. Furthermore, Kim et al. [3] and Lee et al. [4] presented some kinds of variable length encoding methods for GA, and in the GA population where the length of chromosome varies among different individuals to strengthen the represent ability of chromosome for specific problem’s solution. Recently, some evolutionary methods have been used in feature selection, where binary encoding method has been extensively considered for solution encoding in [5][6][7][8][9]. But binary encoding has the shortcomings of the probable existing of irrelevant features in selected feature subset, high time cost for decreasing the feature number and slow convergence speed. Instead, 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. Moreover, our designed method calls GA many times and in each run of GA, all individuals in the population have the fixed length chromosomes and chromosome length only changes between two adjacent GA runs. A similar work by Dashtban and Balafar [10] also provided an integer-coded genetic algorithm with dynamic-length chromosome and intelligent parameter settings for gene selection on microarray data, but their method lacks the recursive feature decline operation. We also made performance comparison between Dashtban’ method and our method in Table 7 and 8 in our manuscript, where the results by Dashtban’ method on datasets SRBCT and ALL\_AML are inferior than the results obtained by our proposed method. Compared to Dashtban’ method, our proposed MGRFE can find much smaller gene subset. In addition, it is also worth mentioning that the RFE process in MGRFE doesn't rank genes to remove the least weighted ones as described in [11], but introduces the random strategy to randomly discard a same number of genes in the chromosome of each individual between two GA runs, which proves to be very effective for the embedded GA by our comprehensive experiments on 19 microarray datasets. As for the selection operator for our embedded GA, the frequently used roulette wheel selection is less efficient than the simple truncation selection [12] currently used in MGRFE based on our extra experiments. 1), the differences between the fitness values of all individuals are just slight in most cases. 2), the gap between fitness values only account for few portion in the whole fitness value. These two points provide all individuals with nearly same area occupations in the roulette wheel and lead to the inefficiency of roulette wheel selection. To the best of our knowledge, none previous studies have designed an evolutionary algorithm using variable length integer encoding method in a recursive manner along with the truncation selection operator to tackle the problem of minimal discriminatory feature selection in high-dimension datasets which is described in this paper.*

2) The partitioning of the datasets into two is not convincing. This reviewer guess the reason is that Ge et al. used one and Kar et al. did the other. The sectioning of dataset one and two needs to be unified and the authors should do the additional necessary experiments on their own for fair comparison.

***Response****: Thank you for this comment. For datasets partitioning, it is indeed just like your suppose and you are insightful. The type of datasets partition is for the sake of consistency and simplicity of results comparison. The Dataset One and Two summarized the most used datasets as benchmarks for performance comparison in this field. Ge et al. and Kar et al. separately performed experiments on Dataset One and Dataset Two, and offered their relative comprehensive results on these two large datasets. Therefore, it is acceptable and convenient to just follow the datasets partition pattern and make the results comparison.*

3) The authors showed the results of five evaluation metrics. To confirm the robustness of the proposed methodology, area under the curve (AUC) could also be compared. In addition, multi-class datasets could be evaluated using other metrics not just the accuracy metric alone. For instance, Asch et al. proposed two metrics, macro-averaging and micro-averaging, because the accuracy metric is vulnerable to imbalanced datasets.

***Response****: Thank you for this comment. For merely accuracy metric is vulnerable for imbalanced datasets and multiclass datasets, we compute the AUC values on 17 binary datasets in Dataset One, and micro-averaging and macro-averaging metrics on 3 datasets in Dataset Two by 5-fold cross validation, which are listed in the following Table 4 and 5. The Python codes of the above computations and their original generated results can be available at* <https://github.com/Pengeace/MGRFE-GaRFE>*.*



4) In table 7-9, the accuracies of the proposed methodology are 98.8 (CV), 98.3 (CV), 99.7 (CV), 100 (train), and 100 (test). The results do not correspond to each other.

***Response****: Thank you for this comment. In table 7-9 in our manuscript, the accuracies of selected gene combinations are 98.8 (CV), 98.3 (CV) and 99.7 (CV) on 3 datasets, which are calculated by 10 times 10-fold CV to validate the classification stability of selected genes. But just in 1 time 5-fold CV, these selected gene combinations can all reach 100% classification accuracies for both the train samples and test samples in most cases.*

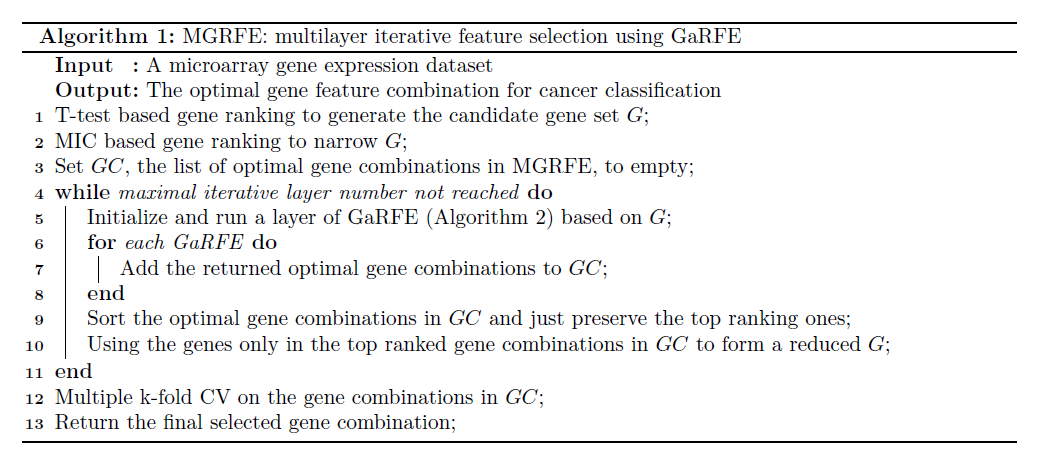
5) There are numerous English grammatical errors in this paper. They should consult with a professional proofreading service.

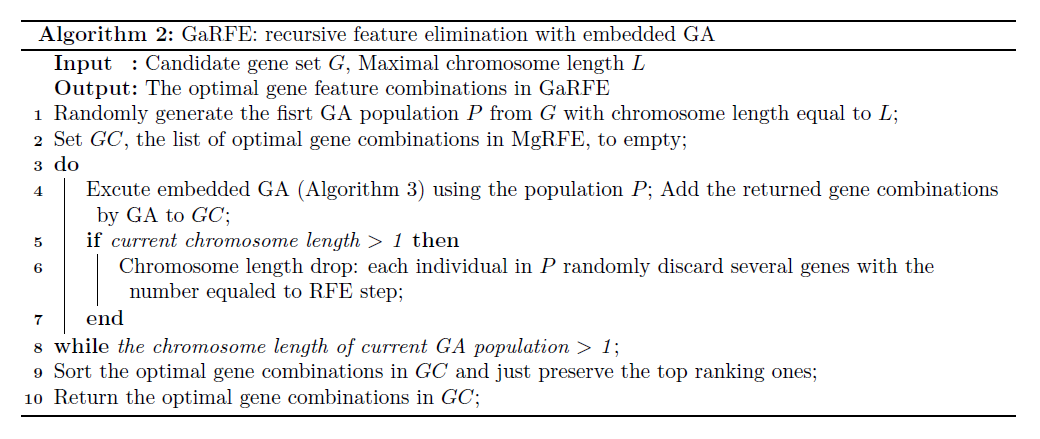
***Response****: Thanks for your comment and suggestion. In the revised manuscript, we have made great effort to correct English grammatical errors and improve the writing by following the professional suggestions.*

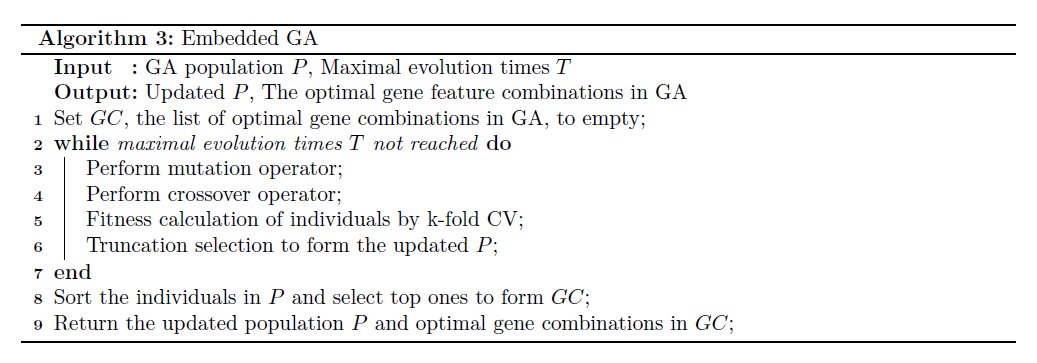
Minor Points:

1. In addition to flowchart on Fig. 2, pseudocode of your methodology could help the readers to understand MGRFE.

***Response****: Thank you for this comment. We have offered three pseudocodes of our method including MGRFE, GaRFE and Embedded GA in the revised manuscript. For convenience to review, the pseudocodes are listed as followed.*







2) Table 7-9 are hard to interpret the meanings especially for the cascaded tables with different number of k-folds, SCV and LOOCV.

***Response****: Thank you for this comment. SCV should be 5CV meaning 5-fold cross validation and LOOCV means leave one out cross validation.*

3) The source code need to be revised to improve the readability and usability. There is no main script to apply their methodology to user’s own datasets.

***Response****: Thank you for this comment. We have completely revised the source code and added sufficient comments to make the code more readable and understandable. The updated code can be available at* <https://github.com/Pengeace/MGRFE-GaRFE>*. In addition, we provide a demo at* <https://github.com/Pengeace/MGRFE-GaRFE/tree/master/Demo> *to show how to use the code to perform the gene feature selection on user’s specified dataset.*

4) The format of this paper does not follow the instruction of Bioinformatics or TCBB.

***Response****: Thank you for this comment. The format of our revised manuscript has been adjusted by following the instruction of Bioinformatics.*

----------------------- REVIEW 3 ---------------------

------------------ Overall evaluation ------------------

This work proposed a new gene feature selection method MGFRE for cancer data classification. An hybrid approach of feature elimination and genetic algorithm optimization was used in a recursive manner, and the result shows comparable or marginal benefit compared to previous gene feature selection methods from benchmark experiments. Overall, the proposed idea is interesting and actually achieved acceptable performances, but its performance gain is rather marginal, without proposing specific types of classification problems that the proposed method can be more beneficial.

***Response****: Thank you for your comment. Our proposed method has showed its great efficiency in gene feature selection through comprehensive experiments on large amount of high-dimensional expression datasets compared with the popular algorithms in this field. Our method could be regarded as a promising method for the classification problems of high-dimensional data especially for the gene expression data. For the analysis of high-dimensional data, feature selection is essential, which aims at removing the irrelevant and redundant features, cutting down the dimensionality, and improving the predictive performance and model interpretability. But due to its 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 of high dimensional data such as gene expression data, which makes feature selection more important and challenging [13]. Furthermore, the* *identification of a small subset of informative predictors is deeply desired in the analysis of high-throughput genomic, proteomic and metabolomic data which is characterized by the“large p small n” paradigm [14]. To conclude, our proposed method could conceivably be applied to the feature selection and classification problems in above high-dimensional datasets due to its excellent performance.*

----------------------- REVIEW 4 ---------------------

------------------ Overall evaluation ------------------

It is hard to find any novel finding emerged from this study. Overly technical. Cancer genome analysis is a mature field; it is not clear how this study can push forward the field.

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

*Firstly, the innovation of our method is to combine the evolutionary strategy of GA with recursive feature elimination process as the search unit GaRFE to give an explicit feature number decline along with the evolutionary optimization search. The large amount experiments on benchmark datasets have showed great performances on finding the minimal informative genes in microarray datasets. Recently, some evolutionary methods have been used in feature selection and binary encoding method has been extensively considered for solution encoding in [5][6][7]* *[8][9]. But binary encoding has the shortcomings of the probable existing of irrelevant features in selected feature subset, high time cost for decreasing the feature number and slow convergence speed. Instead, 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. In addition, it is also worth mentioning that the RFE process in MGRFE doesn't rank genes to remove the least weighted ones as described in [11], but introduces the random strategy to randomly discard a same number of genes in the chromosome of each individual between two GA runs, which proves to be very effective for the embedded GA by our comprehensive experiments on 19 microarray datasets. As for the selection operator for our embedded GA, the frequently used roulette wheel selection is less efficient than the simple truncation selection [12] currently used in MGRFE based on our extra experiments. 1), the differences between the fitness values of all individuals are just slight in most cases. 2), the gap between fitness values only account for few portion in the whole fitness value. These two points provide all individuals with nearly same area occupations in the roulette wheel and lead to the inefficiency of roulette wheel selection. To the best of our knowledge, none previous studies have designed an evolutionary algorithm using variable length integer encoding method in a recursive manner along with the truncation selection operator to tackle the problem of minimal discriminatory feature selection in high-dimension datasets which is described in this paper.*

*Secondly, the gene selection for cancer classification on the microarray gene data is still a hot topic of present research and the cancer related biomarkers found in the expression datasets are important for clinical medical diagnosis and prognosis. We list some publications on gene selection for cancer classification problem using microarray data in recent years as follows: Kar et al. [8] employed PSO and adaptive K-nearest neighborhood to do gene feature selection in 2015; Moosa et al [9] presented a modified Artificial Bee Colony Algorithm (ABC) for seeking cancer related genes in 2016; Ge et al. [15] posed a two-step feature selection algorithm based on maximal information coefficient in 2016; Dashtban and Balafar [10] proposed a novel evolutionary method based on genetical algorithms and artificial intelligence in 2017.*

**REFERENCES**

[1] Holland J H. Adaptation in natural and artificial systems. An introductory analysis with application to biology, control, and artificial intelligence[J]. Ann Arbor, MI: University of Michigan Press, 1975.

[2] Goldberg D E. Genetic Algorithms in Search, Optimization, and Machine Learning[J]. 1989.

[3] Kim I Y, De Weck O L. Variable chromosome length genetic algorithm for progressive refinement in topology optimization[J]. Structural and Multidisciplinary Optimization, 2005, 29(6): 445-456.

[4] Lee C Y, Antonsson E K. Variable Length Genomes for Evolutionary Algorithms[C]//GECCO. 2000, 2000: 806.

[5] Oreski S, Oreski G. Genetic algorithm-based heuristic for feature selection in credit risk assessment[J]. Expert Systems with Applications, 2014, 41(4):2052-2064.

[6] Jung M, Zscheischler J. A Guided Hybrid Genetic Algorithm for Feature Selection with Expensive Cost Functions[J]. Procedia Computer Science, 2013, 18:2337-2346.

[7] Yang J, Honavar V. Feature Subset Selection Using a Genetic Algorithm[J]. IEEE Intelligent Systems & Their Applications, 2002, 13(2):44-49.

[8] Kar S, Sharma K D, Maitra M. Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K -nearest neighborhood technique[J]. Expert Systems with Applications, 2015, 42(1):612-627.

[9] Moosa J M, Shakur R, Kaykobad M, et al. Gene selection for cancer classification with the help of bees[J]. Bmc Medical Genomics, 2016, 9 Suppl 2(2):47.

[10] Dashtban M, Balafar M. Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts[J]. Genomics, 2017, 109(2):91-107.

[11] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines[J]. Machine learning, 2002, 46(1): 389-422.

[12] Blickle T, Thiele L. A comparison of selection schemes used in genetic algorithms[J]. 1995.

[13] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics[J]. Bioinformatics, 2007, 23(19):2507.

[14] Lin W J, Chen J J. Class-imbalanced classifiers for high-dimensional data[J]. Briefings in Bioinformatics, 2013, 14(1):13.

[15] Ge R, Zhou M, Luo Y, et al. McTwo: a two-step feature selection algorithm based on maximal information coefficient[J]. Bmc Bioinformatics, 2016, 17(1):142.