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 and provide us the chance to modify our manuscript again.

Our responses to the review comments are in blue.

Yours sincerely,

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 very much for providing us this valuable opportunity to revise our manuscript again. We have carefully revised the previous manuscript considering all the review comments. The point-to-point response to each comment of the reviewer 2 in the detail are provided as follows.

**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.

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 very much for your constructive and valuable comments. We also do appreciated your patient and detailed explanation on the issue that we did not understand well, which not only provide us the great help in this process of our revision, but also in our future research. As you said, in the previous revision, we indeed did not addressed your queries well due to our incorrect understanding. In this revision, we have supplemented more experiments and revised our manuscript again. We sincerely hope you can provide us another chance to review our revision.

**# Response** for previous comment 3:

According to your suggestion, we have totally added 10-group cross-validation experiments on independent datasets (**Table 1**), the later seven validation datasets are from GEO data repository. For each experiment, firstly, the selected gene probe features from the first dataset were transformed into the official gene symbols; secondly, the obtained gene symbols were transformed into corresponding gene probe Ids in the second dataset; thirdly, 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 three different classifiers, 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.

The later seven validation datasets are retrieved from GEO and named as their GEO accessions.

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.

1. Dataset GSE8511 has three kinds of samples: “Benign Prostate”, “Local Prostate Cancer” and “Metastatic Prostate Cancer”. The later two kinds of samples are combined together as “Prostate Cancer” samples in validation.
2. Dataset GSE44076 has three kinds of samples: “Mucosa sample from healthy Normal donnor”, “Normal paired sample from patient” and “Tumor sample from M2075 patient”. The first two kinds of samples are combined together as “Normal” samples in validation.
3. On dataset GSE8511, there are total seven gene probe features mapped from Prostate dataset, but two of them contain Null values and have been abandoned. Thus, only five gene features used on GSE8511.
4. On dataset GSE2604, there are 36 samples in total, but 22 samples contain Null values and have been removed. Thus, only 14 samples are used on GSE2604.

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

**# Response** for previous comment 4:

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 SVM-RFE method by Isabelle Guyon indeed has many theoretical results and very fast. But the evolutionary-computation based methods are also one of major branch.*

*Recently, many evolutionary-based methods for gene selection have been proposed.*

*Recently, the novel evolutionary-based methods for gene selection were constantly published [1-4], which still caused high concern.*

*Compared to SVM-RFE based method, the time consuming is relatively higher, but the classification accuracy of generated gene subset is usually higher.*

The currently published leading methods of gene selection in microarray are usually base on swarm intelligence algorithms [1, 2, 5].

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 [6]. 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*. [7] 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.

Let’s go back to the beginning and rethink the motivation for identifying compact discriminatory gene features in the microarray.

1. Structural risk minimization (SRM). Structural risk minimization is an inductive principle for model selection used for learning from finite training data sets. It describes a tradeoff between the empirical error in training data and hypothesis space complexity of a learning model. On microarray data, there are usually several thousand to tens of thousands of gene features but only dozens or hundreds of samples. Thus, the features used by the prediction model must be limited to control the model complexity. By selecting relatively small number of gene features, the learnt model could avoid the overfitting problem and have better generalization ability in unseen data. In the Recursive Feature Elimination process, the number of gene features is reduced step by step, thus the corresponding learnt models are arranged in order of decreasing complexity. In minimizing both the empirical error and capacity of a model, the idea of SRM is clearly embodied.
2. Finding disease related potential biomarkers. The selected minimal discriminatory gene subset has high correlation with the disease phenotype on microarray data. Thus, they are potential biomarker candidates for the specific disease and may provide researchers with insights into the genetic nature of the disease and the mechanism behind it. Therefore, the discriminatory genes are worth of further biological analysis.

SVM-RFE by Isabelle Guyon followed the SRM principal and has been widely recognized as the classical feature selection method on microarray data and other problems. Its main procedure is briefly described as follows.

SVM-RFE has below advantages.

Meanwhile, it should be noted that there are some shortcomings of SVM-RFE.

Essentially, as pointed by Isabelle Guyon, SVM-RFE remains a greedy sub-optimal feature selection method.

The original idea of our algorithm come from combining the RFE progress with GA.

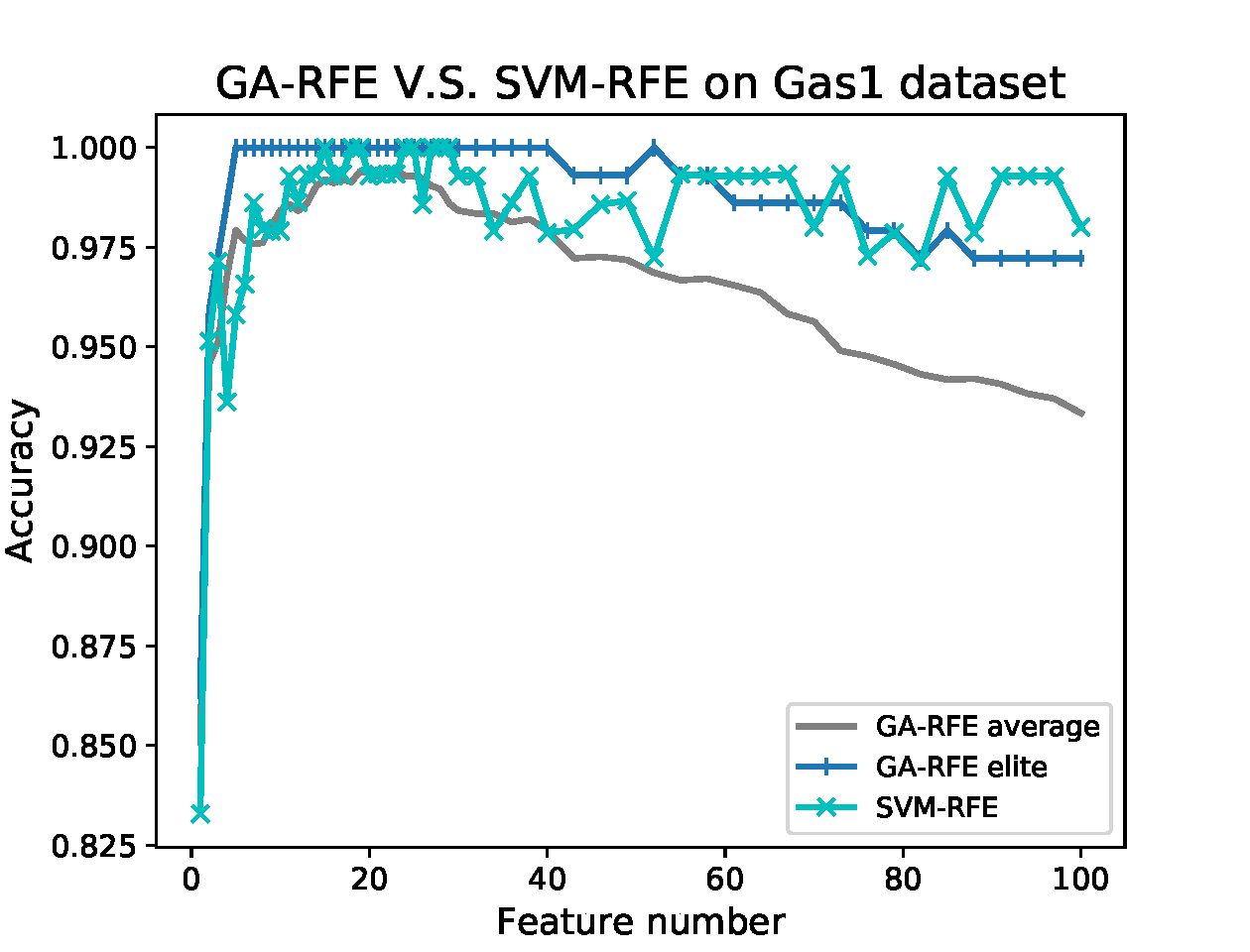
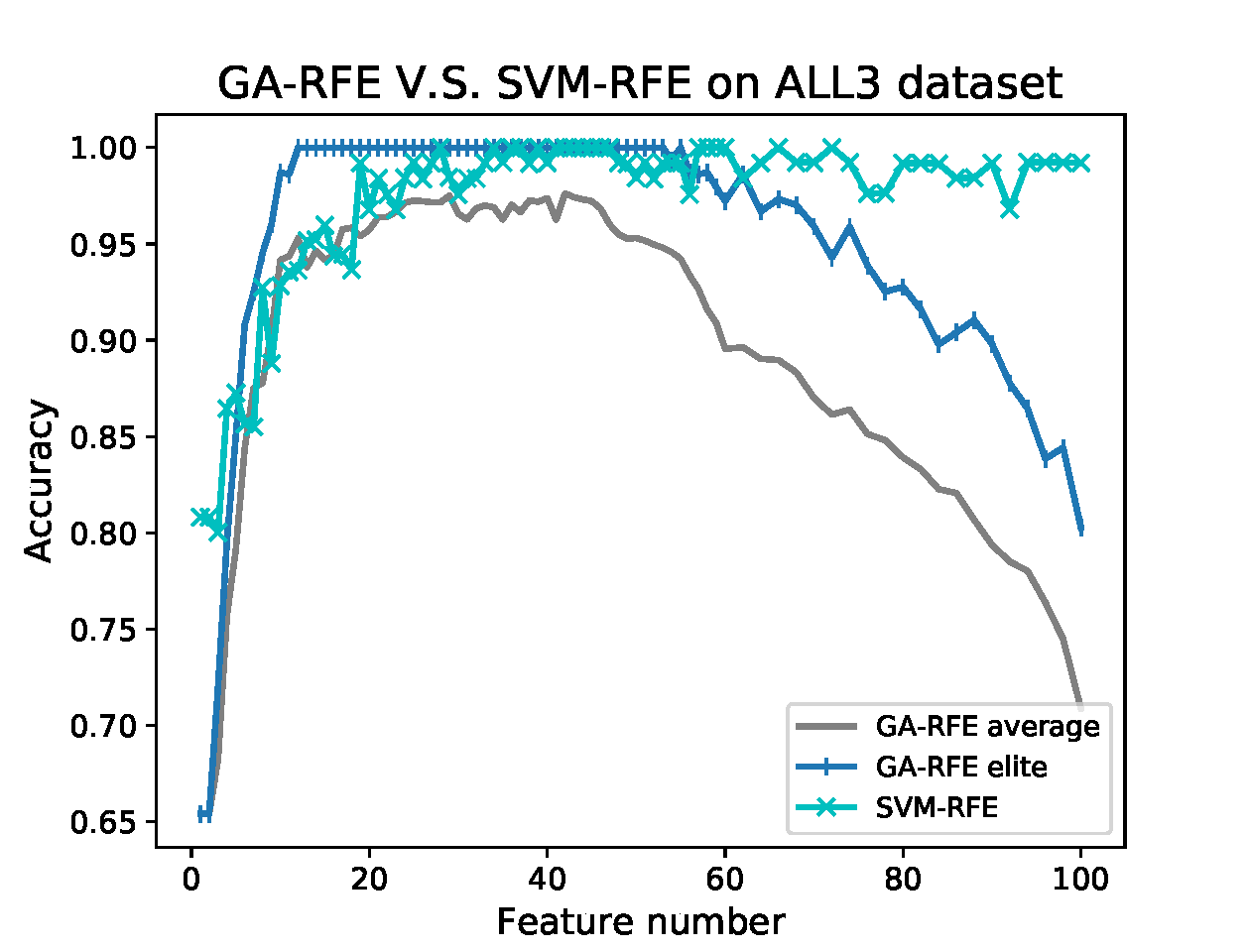
There are two kinds of theoretical analysis method for Genetic Algorithm: and .

Experimental analysis counts.

**Table 2**. Both GA-RFE and SVM-RFE can achieve 100% cross validation accuracy on 17 binary classification datasets, but GA-RFE used more compact gene subsets with smaller sizes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | GA-RFE | | SVM-RFE | |
| #Genes | Accuracy | #Genes | Accuracy |
| DLBCL | 3 | 1.0 | 8 | 1.0 |
| Pros | 6 | 1.0 | 11 | 1.0 |
| Colon | 5 | 1.0 | 5 | 1.0 |
| Leuk | 2 | 1.0 | 2 | 1.0 |
| Mye | 11 | 1.0 | 18 | 1.0 |
| ALL1 | 1 | 1.0 | 2 | 1.0 |
| ALL2 | 10 | 1.0 | 17 | 1.0 |
| ALL3 | 10 | 1.0 | 28 | 1.0 |
| ALL4 | 5 | 1.0 | 12 | 1.0 |
| CNS | 6 | 1.0 | 8 | 1.0 |
| Lym | 3 | 1.0 | 3 | 1.0 |
| Adeno | 1 | 1.0 | 1 | 1.0 |
| Gas | 3 | 1.0 | 4 | 1.0 |
| Gas1 | 5 | 1.0 | 15 | 1.0 |
| Gas2 | 2 | 1.0 | 4 | 1.0 |
| T1D | 14 | 1.0 | 21 | 1.0 |
| Stroke | 3 | 1.0 | 6 | 1.0 |

**Figure 1**. Sizes of selected gene subsets by GA-RFE and SVM-RFE for achieving 100% cross validation accuracy on 17 datasets.



**Figure 2**. Performance comparison of GA-RFE and SVM-RFE with varying feature number. The accuracy is calculated in 5-fold cross validation. These two methods have been provided with an initial feature set containing 500 genes generated by feature filter process on each dataset. Then, GA-RFE begin its evolution from some randomly sampled feature subsets 100 in length to save time. SVM-RFE starts with the exactly provided feature set.

**Reference**

1. 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.

2. 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.

3. Jain, I., V.K. Jain, and R. Jain, *Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification.* Applied Soft Computing, 2018. **62**: p. 203-215.

4. Ghosh, M., et al., *Recursive memetic algorithm for gene selection in microarray data.* Expert Systems with Applications, 2019. **116**: p. 172-185.

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, 2015. **14**(1): p. 85-96.

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

7. 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.