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

RE: TCBB-2018-05-0199, "MGRFE: multilayer recursive feature elimination based on embedded genetic algorithm for cancer classification"

Manuscript Type: Survey/Tutorial

Dear Dr. Li,

We have completed the review process of the above referenced paper for the IEEE/ACM Transactions on Computational Biology and Bioinformatics and recommend that your paper undergo a Major Revision.

Enclosed are your reviews. If you should choose to revise your paper, please prepare a separate document describing in detail how each of the reviewers' comments are responded to in your revision and submit it before 30-Nov-2018.

To revise your manuscript, log into https://mc.manuscriptcentral.com/tcbb-cs and enter your Author Center, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision.

\*\*\*\*\*Text in any color other than black is not acceptable.\*\*\*\*\*

Your revised paper must include the following:

-abstract

-index terms

-author affiliation information

-main text

-references

-figure captions

-table titles

-brief author biographies

(biographies are not required for short papers or comments papers)

Once the revised manuscript is prepared, you can upload it and submit it through your Author Center.

When submitting your revised manuscript, you will be able to respond to the comments made by the reviewer(s) in the space provided. You can use this space to document any changes you make to the original manuscript. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the reviewer(s)’ questions and comments. You may also upload your responses as separate files for review along with your revision. If you choose to do this, please choose “Summary of Changes” as the file designation.

When the submission process is complete, you will receive an automated confirmation email immediately. If you did not receive that email, your submission is not yet complete.

The journal’s Administrator will contact you should we have any concerns or questions regarding your revision. Otherwise, your revision will be forwarded to the assigned Associate Editor to begin a second round of reviews.

Our page limitation and formatting guidelines for TCBB can be found on:

http://www.computer.org/portal/web/peerreviewjournals/author#manuscript

Thank you for your contribution to TCBB, and we look forward to receiving your revised manuscript.

Sincerely,

Aidong Zhang, EIC

IEEE/ACM Transactions on Computational Biology and Bioinformatics

azhang@buffalo.edu

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

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.

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

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 autors 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.

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?

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.

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.

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.

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

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.

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.

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.

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

Additional Questions:

1. Which category describes this manuscript?: Research/Technology

2. How relevant is this manuscript to the readers of this periodical? Please explain your rating under Public Comments below.: Very Relevant

1. Please explain how this manuscript advances this field of research and/or contributes something new to the literature.: This manuscript has proposed a new hybrid methodology towards gene selection from microarray gene expression data. The work is interesting. It presents a low computation time for selecting a useful gene subset. Moreover, the manuscript compared their findings with other related works. The biological significance of the selected genes have also been discussed in the manuscript.

2. Is the manuscript technically sound? Please explain your answer under Public Comments below.: Yes

1. Are the title, abstract, and keywords appropriate? Please explain under Public Comments below.: Yes

2. Does the manuscript contain sufficient and appropriate references? Please explain under Public Comments below.: References are sufficient and appropriate

3. Does the introduction state the objectives of the manuscript in terms that encourage the reader to read on? Please explain your answer under Public Comments below.: Could be improved

4. How would you rate the organization of the manuscript? Is it focused? Is the length appropriate for the topic? Please explain under Public Comments below.: Could be improved

5. Please rate the readability of the manuscript. Explain your rating under Public Comments below.: Easy to read

6. Should the supplemental material be included? (Click on the Supplementary Files icon to view files): Yes, as part of the digital library for this submission if accepted

7. If yes to 6, should it be accepted: After revisions. Please include explanation under Public Comments below.

Please rate the manuscript. Please explain your answer.: Good

Reviewer: 2

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

Comments:

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.

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.

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.

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.

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.

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.

Additional Questions:

1. Which category describes this manuscript?: Research/Technology

2. How relevant is this manuscript to the readers of this periodical? Please explain your rating under Public Comments below.: Relevant

1. Please explain how this manuscript advances this field of research and/or contributes something new to the literature.: 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.

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.

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.

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.

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.

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.

2. Is the manuscript technically sound? Please explain your answer under Public Comments below.: No

1. Are the title, abstract, and keywords appropriate? Please explain under Public Comments below.: Yes

2. Does the manuscript contain sufficient and appropriate references? Please explain under Public Comments below.: References are sufficient and appropriate

3. Does the introduction state the objectives of the manuscript in terms that encourage the reader to read on? Please explain your answer under Public Comments below.: Could be improved

4. How would you rate the organization of the manuscript? Is it focused? Is the length appropriate for the topic? Please explain under Public Comments below.: Could be improved

5. Please rate the readability of the manuscript. Explain your rating under Public Comments below.: Readable - but requires some effort to understand

6. Should the supplemental material be included? (Click on the Supplementary Files icon to view files): Does not apply, no supplementary files included

7. If yes to 6, should it be accepted:

Please rate the manuscript. Please explain your answer.: Poor