Reviewers report

Title: Deep learning-based ovarian cancer subtypes identification using multi-omics data

Version: 0 Date: 16 Jun 2020

Reviewer: Heejoon Chae

Reviewers report:

This article proposed a deep-learning based framework for identifying ovarian cancer subtypes. Based on the denoising autoencoder, multi-omics dataset composed of mRNA, DNA methylation, miRNA and CNV dataset were integrated and representative features were extracted. K-means clustering was applied for subtype identification and authors provided biomarkers associated with ovarian cancer by the classification model with L1-penalized logistic regression. The authors showed the performance evaluation to provide reliable results, however, I have major and minor concerns that should be addressed for publication.

1. The aim of this paper is not clear. This paper proposed a framework for ovarian (OV) cancer subtype identification, however, in the methods section, this framework only cluster OV cancer datasets into two subtypes named low risk and high risk. Is OV cancer originally composed of only two subtypes? For subtype identification, authors should at least perform estimation of the number of clusters by showing the performance measurement of clustering, whether the two clusters are the optimized number of clusters. There is no chance of finding new novel subtypes of OV cancer in this framework. Moreover, if we set the number of clusters for two from the beginning, why do we need to use this method for cancer subtype identification?

2. Description of the method part is insufficient. To integrate four omics datasets by denoising autoencoder, are all the datasets concatenated and then used for the input? Also, after applying the denoising autoencoder, are the extracted features used for the K-means clustering? There is no any other description of K-means clustering part in the methods section.

3. In the methods section, after obtaining the labels clustered by k-means, the proposed framework performs logistic regression-based classification to provide biomarkers for the labelled subtypes using mRNA features.

Do the mRNA features refer to gene in the dataset? In this case, is the multi-omics dataset only used for subtype clustering and the gene expression dataset only used for biomarker detection? What is the difference between using multi-omics dataset for subtype identification and using mRNA dataset for subtype identification, if the paper provides biomarkers based on the mRNA dataset?

3-1. Three GEO datasets of gene expression profiling were used for testing the logistic regression classification model with lasso. However, the classification model is constructed based on the subtype labels obtained from the previous denoising autoencoder (DAE) and K-means clustering steps. How did those geo datasets obtain the subtype labels, as they did not have other 3 omics dataset (CNV, DNA methylation, miRNA) to be used for DAE?
4. In the section of evaluations of ovarian cancer subtypes identification (Methods), the authors wrote that they presented performance comparison of different cluster methods. However, I think the description should be modified. They used k-means clustering for all experiments and used different dimensionality reduction methods. (e.g. PCA, kPCA, AE, DAE). This experiment shows the performance evaluation of different dimensionality reduction methods, to validate the performance of DAE. Please modify the description in the section.

4-1. For the evaluations of OV cancer subtype identification, the authors should add the results of comparing the previously presented OV cancer subtype identification methods or at least widely used general cancer subtype identification methods such as iCluster or SparseK.

5. This paper presented the cancer subtype identification framework using multi-omics dataset. The performance comparison between using each single omics should be added in the results section to show the advantage of using multi-omics dataset.

6. In the results section, more details of experiments should be added to help reader for better understanding.

6-1. Please add the results of classification accuracy from the logistic regression classification model.

6-2. Please provide the number of features in each omics dataset, and how many features were selected after the logistic regression classification model with lasso.

6-3. In Table 2, what does the number of in the table represent? Is it the number of samples for each subtype? Also, with my understanding, the authors clustered the samples into the low risk and high risk group. What do the "Censored" and "Uncensored" used for?

6-4. In figure 3, are the clustered modules from WGCNA represented by the colors? Then add the description of it, to make readers understand.

7. The architecture parameters are introduced in the methods section. How were these parameters selected? Were they randomly selected?

**Level of interest**
Please indicate how interesting you found the manuscript:

An article of limited interest

**Quality of written English**
Please indicate the quality of language in the manuscript:

Needs some language corrections before being published
Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal

Publons Reviewer Recognition. Springer Nature can send verification of this review directly to Publons (a subsidiary of Clarivate Analytics). If you would like to take advantage of this service, please click on the “Yes” option below. Your name, email address, title of the reviewed manuscript, name of the journal, and date of your review submission (the “Review Data”) will then be transmitted to Publons upon publication of the manuscript. If you have already registered at Publons, they will notify you of the receipt of this review and update your profile as per your settings and their policy. If you are not registered with Publons, you will receive an email from them asking you to register in order for them to be able to recognize your review on your new profile page. Publons may use the Review Data to generate derivative metadata for the benefit of Publons and you as a reviewer, carefully considering the sensitivity of such information. For
example, Publons may verify your record as a reviewer by updating your profile published on its webservice if you have registered for such service or help editors to identify candidate reviewers. Please find the details of processing in Publons’ privacy policy https://publons.com/about/terms

No