Revisions Response Letter

Article name: A hybrid metaheuristic-deep learning technique for the pan-classification of cancer based on DNA methylation

Submission ID: 4f4ea643-0140-4dc4-b012-df52e281680a

Authors: Noureldin S. Eissa, Uswah Khairuddin and Rubiyah Yusof

Dear Editor,
Prof Peng Wei,

We would like to thank you for giving us the opportunity to revise and resubmit our manuscript, and for the great and seamless peer review process. We value your comments and sincerely appreciate the constructive feedback that helped us improve our article and highlight the proposed system. We value the time and efforts of the two reviewers and thank them for their insightful and kind comments that helped improve our article.

We have thoroughly revised our manuscript according to your feedback and comments, as well as those provided by the reviewers. According to the given feedback, we modified the manuscript in the following ways:

1- We revised the definitions for the DNA methylation and added references as advised.
2- We conducted another major study that includes more cancer types (Colon, Liver, Lung, Kidney, Thyroid, Prostate and Breast) profiled using the Illumina Infinium 450k platform.
3- The studies were re-implemented/carried out using only the training dataset (70%) for both stages of the proposed method.
4- For convenience and reproducibility, the developed software in its current stage, along with the Keras Tensorflow trained model have been uploaded to this public GitHub repository: https://github.com/smnoureldini/MetaMethyLib.
5- The manifest data required to download the samples used in the studies, along with their platform version, are added to the supplementary files section, and are referenced in the manuscript body.

Kindly find below an itemized point-by-point response for each comment raised by you and the two reviewers.

Editor's comments

1) Please correct the description about DNA methylation, e.g., CpG islands. The following recent review paper may be informative: Yousefi et al (2022) DNA methylation-based predictors of health: applications and statistical considerations. Nat Rev Genet 2022 Mar 18. doi: 10.1038/s41576-022-00465-w. PMID: 35304597
Response:
Thank you very much for bringing this to our attention. We corrected the DNA methylation definition and included the above reference, along with another reference. Changes can be checked on page 2, line 15 to line 21. The two added references are numbered [7][8].

2) This manuscript used an out-of-date DNA methylation platform (27K) in the TCGA, while the majority TCGA samples have been profiled by the 450K array. The proposed method needs to be applied to the 450K DNA methylation data to demonstrate whether it can be scaled up to much larger number of CpG sites. In addition to breast, ovarian and stomach cancer, the authors need to include additional major cancer sites, e.g., lung cancer and prostate cancer, to demonstrate the pan-cancer classification performance in terms of discriminating tumors from different sites, as well as discriminating tumor from normal tissue from the same site.

Response:

Thank you for the insightful feedback. The 27k DNA Methylation platform dataset and the selected cancer types were initially used to test and present the proposed system and check whether the classification would be reliable or not, which was the initial focus of this study. As per your recommendation and to better demonstrate the proposed system and its ability to handle larger number of features and discriminate between the different cancer types, we conducted a new study that encompassed more primary cancer diseases, that were sampled using the superior Illumina Infinum 450K platform. The chosen cancer diseases were based on the availability of enough number of samples (malignant and benign) on TCGA GDC portal and our resources. For instance, the 450K platform did not have enough samples for the ovary cancer, so we had to skip it in 450K platform study.

3) Please provide the date, portal and version information of the downloaded TCGA data to facilitate reproducibility.

Response:

Thank you for the advice. The data was downloaded using the written software, with the help of GDC Client tool. The exact manifest files required to download all the samples used in the studies are attached as supplementary/additional files in the manuscript. The description includes the access date as well as the DNA methylation platform. To facilitate reproducibility, we also uploaded the trained Keras tensorflow pan-classification model to the same GitHub repository with the software.

4) The feature selection step needs to be based on a training set, rather than the entire dataset, to avoid double use of the data.

Response:

Thank you for bringing this up to our attention. Apparently, we were inadvertently peeking into our data during the feature selection stage. We reconducted all studies, and implemented the new ones, using only the training group (70%) to do the feature selection and to train the DNN. We have explicitly added this to the manuscript body as well, which can be found on page 13, line 27 to line 30.

5) Please provide the computer codes/software at a public website, e.g., Github.

Response:

Thanks for the reminder. The software is still in development, but to facilitate the reproducibility and the reviewing process, as required, we have uploaded a copy of the software, along with the trained keras tensorflow model on the public GitHub repository: https://github.com/smnoureldini/MetaMethyLib.

6) Bottom of Page 15: the definition of TPR/specificity and FPR/sensitivity is incorrect.
Response:
Thanks a lot for catching this. It was unintentional and was mixed up during the writing of the manuscript. We corrected the mistake. (Correction is on page 20, line 30 to 32).

Reviewer 1 comments

The authors developed a novel deep learning method using DNA methylation data for pan-classification of cancer in the TCGA data. This work showed some nice figures and interesting results indicating that DNA methylation can be used for cancer classifications

Response: Thank you for taking the time to thoroughly review our manuscript. We really appreciate it, and we are thankful for your comments and insightful feedback, that helped enhance the quality of our manuscript.

Major Comments

1) The paper is written and presented pretty clearly. However, the methodology part for the second stage of supervised deep leaning classification can be further elaborated as it is not clearly explained how the hidden layers were trained and how the model is built related to the output layer.
Response:
Thanks for the praise and for pointing this out. We included detailed information in the "Supervised Deep Learning Classification" section, about the DNN. The layers use ReLU activation functions, and we use Adam as an optimizer function. The details about the size of layers can be found in Figure 4. We used dense layers (each neuron receives input from all neurons of its previous layer) for the whole model and added information about it in the classification section. Grid search based on the confusion matrix was used to tune the hyper-parameters.

2) Does this method incorporate the correlations between CpG sites? CpG sites can be highly correlated depending on its locations and relations to each other, such as CpG island. Is this considered in this method and how is the correlation taken into account? In addition, CpG sites on different chromosomes can also be correlated to each other. In the first stage, feature selection was performed on each chromosome in parallel, does it consider the possible correlations across the chromosomes to further reduce the CpG sites number to improve the second stage modeling?
Response:
Thanks for your comment. We did not include or incorporate the correlations between the CpG sites in our research. The focus of our research was to use completely unsupervised metaheuristic technique due to its ability to discover relevant hidden patterns and information in the data, by taking advantage of the data's mathematical similarities and elemental structure. In future studies, we can implement semi-supervised feature selection and test the correlation between the different chromosomes.

Minor Comments

1) The proportion of training group to be 30% and testing group to be 70% seems arbitrary. As machine learning methods usually guarantee at least 50% of the total

sample size for training purpose, please elaborate the reasons using those proportions in this work. How does the result change if the proportion of testing set and training set is different.

Response:

We thank the reviewer for the great suggestion. We re-conducted the whole studies using randomly divided groups of 70% for testing/feature classification and 30% for training. The extracted features were slightly less (Page 13 line 45 to Page 14 line 30). The results of the classification stage were the same, however the ROC area under the curve increased slightly. Also, the first stage (the feature selection) took more time per generation (e.g. 17m for breast cancer) compared to previously (e.g. 7m for breast cancer).

2) The computational speed for this method is not presented in this work, please add this information. In other words, how fast is the method? Generally random forest (RF) has great computational advantage and the proposed method shows improvement in recall compared to the RF, how does the computational speed compares to other methods in table 5?

   Response: Thanks for the good question. The authors of the RF method (Modhukur et al.) did not provide any speed metrics to be able to compare with. Since the proposed system was not meant to run in real-time (real-time machine learning), we did not include a speed/time comparison. To give an insight of the speed of the proposed system: for the 450K DNA methylation profile, the proposed system took 27m, 22m, 22m, 19m, 17m, 15m and 13m to finish one GA generation for breast, thyroid, kidney, prostate, colon, lung, and liver, respectively. As for the second stage (DNN classification), it took 44s per epoch.

3) The method is applied to three cancer types, in the TCGA database, there are up to 33 types of cancers and most of them have DNA methylation data available. Does the method perform well/computationally compatible with applications of more than 3 cancer types?
4) The input dataset for the deep learning framework is 27k for first stage feature selection, i.e., unsupervised clustering. Recent technologies can measure more than 450k or 850k methylation CpG sites. Can this be incorporated into the model for a wider application.

   Response:

   Thanks a lot for the great suggestion. To better highlight the proposed system and measure its ability to handle larger number of features, and discriminate between the different cancer types, we performed a new study that encompassed more primary cancer diseases (liver, kidney, lung, prostate, thyroid, colon, and breast), that were sampled using the superior Illumina Infinum 450K platform. The chosen cancer diseases were based on the availability of enough number of samples (malignant and benign) on TCGA GDC portal and our resources. For instance, the 450K platform did not have enough samples for the ovary cancer, so we had to skip it in 450K platform study. We updated the different sections of the manuscript to reflect the new studies and their results (Abstract: page 1 line 25 to 29, line 37 to 39, Methodology and Tools: page 4 line 20 to 24, line 36 to 41, page 5 line 1 to 16, Supervised Deep Learning Classification: page 11 line 31 to 33, Feature Selection: page 13 line 27 to 30, line 42 to 46, page 14 whole page, page 15 line 1 to 24, DNN Pan Classification: page 15 line 27 to 46, page 16 line 1 to 11, page 18 whole page, page 19 line 14 to 38, page 20 line 11 to 27, page 21 line 1 to 24, Conclusion: page 21 line 41 to 46).

Reviewer 2 comments:

1) Some descriptions are wrong, which makes me feel the authors are not very familiar with DNA methylation data. For example, the author stated that "In humans, DNA methylation occurs in genomic regions known as CpG islands, where a Guanine nucleotide follows Cytosine." This is a wrong statement. The authors may check what is "CpG islands" stand for.

   Response:

   Thanks a lot for pointing this out. We corrected the DNA methylation definition and included the above reference, along with another reference. Changes can be checked on page 2, line 15 to line 21. The two added references are numbered [7][8].

2) What's the reason for using breast cancer, ovary cancer, and stomach cancer as examples? In TCGA, we have many other cancer data, such as prostate cancer

8) It seems the authors used 27K array data, which is old. I am very curious why not use more standard 450K array data in TCGA.

   Response:

   Thanks for the good question and suggestions. We used The 27k DNA Methylation platform dataset and the selected cancer types to initially test and present the proposed system and check whether the classification process will be reliable or not, which was the initial focus of this study. To better demonstrate the proposed system and its ability to handle larger number of features, and discriminate between the different cancer types, we have added a new study that encompassed more primary cancer diseases as suggested (colon, kidney, lung, prostate, thyroid, liver, and breast). The samples in the new study were profiled using the superior Illumina Infinum 450K platform. The chosen cancer diseases were based on the availability of enough number of samples (malignant and benign) on TCGA GDC portal and our resources. For instance, the 450K platform did not have enough samples for the ovary cancer, so we had to skip it in 450K platform study. We updated the different sections of the manuscript to reflect the new studies and their results (Abstract: page 1 line 25 to 29, line 37 to 39, Methodology and Tools: page 4 line 20 to 24, line 36 to 41, page 5 line 1 to 16, Supervised Deep Learning Classification: page 11 line 31 to 33, Feature Selection: page 13 line 27 to 30, line 42 to 46, page 14 whole page, page 15 line 1 to 24, DNN Pan Classification: page 15 line 27 to 46, page 16 line 1 to 11, page 18 whole page, page 19 line 14 to 38, page 20 line 11 to 27, page 21 line 1 to 24, Conclusion: page 21 line 41 to 46).

3) It is very interesting to use the genetic algorithm to select factures. However, I am unable to understand this part very clearly even though I learned the genetic algorithm a long time ago. The authors may consider adding some explanations for this part. For example, what's the reason for using the inner layer genetic algorithm to evaluate the fitness of the selected features? I am unable to understand why a genetic algorithm is needed here.

   Response:

   Thanks a lot for the comment. The focus of this study was to use completely unsupervised metaheuristic technique to perform the feature selection, due to its ability to discover relevant hidden patterns and information in the data by taking advantage of the data's mathematical similarities and elemental structure. The outer GA is mainly used to provide the inner GA with the randomly (at first) selected features. The inner GA uses these features to cluster the data. Based on the designed GA operators (explained in the unsupervised Metaheuristic Feature Selection)

section, the separability of the clusters will be evaluated by the inner fitness function, where better separation of clusters signifies better feature selection and noise elimination. Each chromosome in the outer GA will calculate its fitness (the relevance of the selected features) based on the feedback from the inner GA. We added a recent reference (page 5 line 30) to an article that reviews genetic algorithms and metaheuristic techniques.

4) The data seems double used. It seems the feature selection step uses all data, which may yield over-optimistic results. The authors may consider dividing the data into model building and external validation parts.
Response:
Thanks for the great suggestion. Apparently, we were unintentionally peeking through the data during the feature selection stage. We reconducted all studies and implemented the newer ones using only the training group (70%) to do the feature selection and to train the DNN. We have explicitly added this to the manuscript body as well, which can be found on page 13, line 27 to line 30.

5) The feature selection step only removes about 25% of CpG sites. I am curious how this step helps. In other words, what's the result if we use all CpG sites for the prediction model building step.
Response:
Thanks a lot for pointing this out. We are sorry for not making this part clear, we rephrased the section to indicate that we used the common features (extracted from all selected features of different cancer types) in the classification part. The common features for the 27k DNA methylation profile and the 450k DNA methylation profile were 3,391 and 4,273, respectively. We amended this section in the manuscript (Page 15, line 15 to line 24).

6) I am curious if any covariates have been adjusted. It seems the authors using CpG sites only. DNA methylation is impacted by many covariates, such as age, smoking, cell type composition, etc.
Response:
Thanks for the good question. For the time being, even though it is proven that DNA methylation is impacted by external/internal covariates (we included this in our manuscript on page 2 line 15 to line 21), our main focus was to use the DNA methylation data (beta values) independent of any covariates, to help classify the different cancer diseases. Hence why we used an unsupervised metaheuristic technique to do the feature selection.

7) The authors spent extensive time developing software. I am curious how readers can access this software.
Response:
Thanks a lot for the praise and for pointing this out. The software is still under development, and more modules are being added to it. For convenience, we uploaded the current version of software to help, along with the Keras tensorflow trained model, to this public GitHub repository:
https://github.com/smnoureldini/MetaMethyLib

Thanks again to all reviewers and for the editor, for their constructive criticism and the valuable feedback that helped make our manuscript better. We sincerely appreciate the time and effort you have put so far in the reviewing process.