Reviewer’s report

Title: Predicting factors for survival of breast cancer patients using machine learning techniques

Version: 0 Date: 14 Oct 2018

Reviewer: Didi Surian

Reviewer’s report:

In this paper, the authors present a work in predicting factors that influencing the survival rate of breast cancer patients using machine learning methods. The authors compared several machine learning methods and showed that random forest gave a slightly better performance, i.e. 82.7% accuracy. Although the authors described in detail their work, there are many things in the paper that the authors should address:

1. Introduction, para 2. The authors said "They are reported as most efficient and accurate prediction algorithms". This statement is too strong and could not be justified.

2. As many previous related studies have been presented in Table 1, some of them share a similarity with this work, i.e. aim and methods (Random Forest, Decision Tree). Please state at the beginning the difference, uniqueness, contributions of this work compared with other studies.

3. The number of samples in Table 4 is 6,461, which is different from the total number of samples, i.e. 8,066. Please why there is a difference and how to choose these 6,461 samples.

4. In Step 2 (Variable Selection), the authors wrote "The threshold was set to the mean of the absolute values of the first order differentiated errors in the model with 24 variables: ". It is not clear whether this sentence is a complete sentence.

5. Table 6. In the caption the authors mentioned "Model accuracy of six algorithms. Random Forest yielded slightly better accuracy using 'all data'.". Does Table 6 only show the results for 'all data'? If yes, please describe the reasons why only the results from 'all data' were showed here, i.e. was it because Random Forest gave the best results for 'all data'? If not, please make it more clear, which results were for which data.

6. Figure 4. The method names are inconsistent with the ones defined in the text (especially in Model Evaluation).

7. Figure 5. Please explain why the training data gave the largest error rate, following testing and validation data. Please also make it more clear, what is 'training', 'testing', and 'validation'.

8. Figure 7 - 33. It is helpful for the readers to at least give an explanation how to interpret the results in each figure rather than just presenting them. Please have a short explanation/description regarding this in the paper.
9. There are many typos in the paper, one example is, 'the United Kingdom' not 'United Kingdom'.

10. The reviewer understands that the authors present all details in the paper and really appreciate the effort. However, it is hard to follow and there are multiple same descriptions or information in the paper. The authors should fix the flow and structure of the paper.

11. Although this is not the final version, all figures were produced poorly in quality so they are almost not readable.

12. It is not clear why the authors mean with 'a methodological approach' and 'using machine learning techniques' in the title. This is because at the end the authors focused more on 'Random Forest'.

**Are the methods appropriate and well described?**
If not, please specify what is required in your comments to the authors.

Yes

**Does the work include the necessary controls?**
If not, please specify which controls are required in your comments to the authors.

Yes

**Are the conclusions drawn adequately supported by the data shown?**
If not, please explain in your comments to the authors.

Yes

**Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?**
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

**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.