**Reviewer’s report**

**Title:** Diagnosis of thyroid nodules for ultrasonographic characteristics indicative of malignancy using random forest

**Version:** 0  **Date:** 27 Mar 2020

**Reviewer:** Xingjian Chen

**Reviewer's report:**

This paper utilized a logistic regression model to identify the potential thyroid malignancy, and the LASSO method is adopted to simultaneously select the US characteristics, based on the selected US characteristics, the authors calculate the probability for each of thyroid nodules via random forest (RF) and develop a scoring system to classify thyroid nodules. While the paper is well-motivated, the work would require a substantial amount of revision (from both methodological and writing perspectives) before acceptance. Here is a list of comments and suggestions for the author:

1. For the Selection of US Characteristics, the author assumed that "there was a sparsity for the above considered logistic regression. In this case, it is rather difficult to adopt the traditional variable selection method to select US characteristics because of a relatively large number of US characteristics involved." But it seems that there are only 16 US characteristics and these medical features are not really sparse, therefore the explanation for using LASSO is not convincing.

2. As I know, MCC is not commonly used for assessing the prediction accuracy of a classifier, especially for that your dataset is imbalanced, so the evaluation metrics should be considered, or at least add more related latest references that used the same metrics which may be more convincing.

3. The performance improvement is not obvious. And what is most important, the comparison methods and proposed method including feature selection methods did not talk about how to select the parameters for each model, sometimes this will lead to a huge difference to the final performance.

4. The comparison methods should include more state-of-the-art methods such as NN-based, SVM-based algorithms and also the same for the feature selection algorithms.

5. The accuracy may not be a good metric for the imbalanced dataset, it seems that AUC or F1 is better, as well as ROC.

6. The typo of some places, especially for some equations and some descriptions such as "The reported metrics were the averages for 100 repetitions. Examination of Fig. 3 showed that the LR was slightly better than the LR with..."
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:

Acceptable

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