Review of: "A machine learning platform to estimate anti-SARS-CoV-2 activities"

Babak Sokouti[1]

1 Tabriz University of Medical Sciences

Potential competing interests: The author(s) declared that no potential competing interests exist.

According to the fact that COVID-19 disease is still spreading worldwide, threatening the human life and as of 16 July 2021, WHO (world health organisation) reported 188,655,968 confirmed cases and 4,067,517 deaths (i.e., https://covid19.who.int/). A recently published research article demonstrated a machine learning infrastructure through a DrugCentral web portal, http://drugcentral.org/Redial, to predict the anti-viral activities of small molecules considering eleven assays as described by Govinda et al. [1]. Besides, various companies, publishing organizations including Springer Nature and others signed a joint statement committing to share publicly the data and research-based findings related to COVID-19 (i.e., https://wellcome.org/press-release/sharing-research-data-and-findings-relevant-novel-coronavirus-ncov-outbreak). However, at first glance, it seems that the DrugCentral web portal is not publicly available for some countries despite the successful pinging the IP address of the webserver (i.e., 192.241.184.47) being stopped by the error message “This site can’t be reached” or “The requested resource was not found on this server”.

Reviewing the whole manuscript in details resulted in some typos and considerations as below:

1- In the Methods section and Machine learning classifiers sub-section, the number of machine learning models available from the scikit-learn python package [2] has been mentioned 24 however, after carefully counting them, they are 22 also mentioned in the Results section and Model development sub-section. However, it may not influence the total results as only 15 of them employed the output class probabilities.

2- There may be some unrelated citations such as, e.g., 16. Oprea, T. I. & Waller, C. L. in Reviews in Computational Chemistry Vol. 11, 127–182 (John Wiley and Sons, 2007) cited in the Methods section and Machine learning classifiers sub-section for the sentence “All of these algorithms are implemented in the scikit-learn python package [16]”.

3- One of the generally used techniques for data division is stratified sampling, however, to perform an effective stratification the means of the strata should have the largest variances to have them homogeneous [3]. The authors used the stratified sampling to divide the samples in to 70% training, 15% validation, and 15% testing sets. However, the abovementioned critical issues have not been discussed and hence, it is not clear if they have taken them in to consideration.

4- Most poor performance results may suffer from the data division procedure that needed to be more optimized during the pre-treatment stage.

References

1. Kc GB, Bocci G, Verma S, Hassan MM, Holmes J, Yang JJ, et al. A machine learning platform to estimate anti-SARS-CoV-2 activities. Nature Machine Intelligence. 2021;3(6):527-35. doi: 10.1038/s42256-021-00335-w.

2. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12(null):2825–30.

3. Glasgow G. Stratified Sampling Types. In: Kempf-Leonard K, editor. Encyclopedia of Social Measurement. New York: Elsevier; 2005. p. 683-8.