Manuscript Reviewing Opinions to Authors

"Challenges and Limitations of Synthetic Minority Oversampling Techniques in Machine Learning"

In the short manuscript, authors claimed a viewpoint that, in machine learning, oversampling and oversampling based SMOTE techniques are with strong drawbacks. Hence, authors did not recommend using such techniques and proposed other better methods.

In my opinion, above viewpoint is interesting because dataset imbalance and solution explorations are indeed frequently met problems in machine learning modelling processes against read-world dataset. Though this was a short manuscript for editorial, authors are required to present more evidences so as to provide more solid support for such viewpoint and texts in the manuscript.

First, in the first paragraph of main text section, authors wrote such sentences:

" Oversampling, on the other hand, is the most utilized approach, as seen by the plethora of oversampling methods developed in the last two decades, to address the issue of class imbalance in the medical field, many publications have used oversampling, particularly Synthetic Minority Oversampling Technique approaches (SMOTE) to construct artificial samples from minority samples. For example, searching PubMed for the phrase "oversampling" OR "smote" yielded 2157 results produced between 2000 and 2022, whereas searching Web of Science (WOS) yielded 2185 hits (but restricted to medical subjects). This is essentially an indication of the developing trend of oversampling research in the medical literature, which dealt with or simply discussed oversampling. (Figure 1) "

Above sentences led to a series of questions in my mind:

1.1 How could authors conclude that, oversampling was the most used techniques in so simple way ? Did authors searched databases using keywords of other 5 types of techniques and compare those search results of different keywords with those of oversampling's ? I don't find such operations nor comparative results described in authors' manuscript text nor reflected in the Figure 1.

Response : thank you for raising this point , its is one of the most used not the most used we agree with you and we have edited it accordingly, also a comparison of oversampling and under-sampling techniques over multiple years as an example of a comparative per your request in figure 2, thus providing a more robust basis for our

conclusions. While we acknowledge the limitations of our method, our goal was to gain an overview of technique utilization.

1.2 Both PubMed and Web of Science are comprehensive/multi- disciplinary database. Not only medical research papers are adopted in those databases. How could authors limit searches against

"medical subjects" only ? Please describe the detailed steps of how to limit search range within"medical subjects" against above literature databases.

1.3 "Medical subjects" covers super wide range of concepts and terminology, with possibility those that inter-disciplinary contents (those that not purely medical) are also included. Using suchbroad term "medical subjects" instead of terms of precise definitions and indications, how can people believe that the returned literature number of authors' search results are correct ? (Meaning that the numbers showed by authors in the manuscript can truly represent the publication numbers of relevant theme)

Response : thank you for addressing this point, for the web of science we used the categories filter to only include medical related categories. For PubMed it is for healthcare database as it used for the retrieval of biomedical and life sciences literature so no need for adjusting it because its scope is already set (please see the reference https://pubmed.ncbi.nlm.nih.gov/about/). Additionally, we have edited the medical subjects point in the manuscript to match your request. We are open to any suggestions, recommendations, or critiques that will help enhance the quality and clarity of the research presented. Thank you in advance for your time and consideration.

Second, author cited others' works and well interpreted those drawbacks about oversampling and SMOTE techniques in paragraph 2~4. However, in the last paragraph, where authors proposed those substitutes were better than oversampling techniques, authors just wrote very simple and short summary for this section. Here, I expected that I would be able to read a well interpreted section just like how authors interpreted drawbacks of oversampling and SMOTE techniques.

Response: thank you for your valuable comment, we have added a new well-interpreted section discussing the recent substitutes in the literature. Also, we have

paraphrased the manuscript based on the editors request . please find the new edited version well

Finally, authors should bear in mind that, the core purpose of machine learning is to construct high quality predictive models. So, I don't agree that undersampling is better than oversampling just because undersampling does not used those training dataset which might be artificial or wrong. Note that, undersampling techniques are not using training dataset sufficiently. In such case, parts of features of training dataset are not learned by the machine. Consequently, when similar data are in the prediction dataset, the models trained by undersampling techniques may not be able to classify those data correctly. Again, remember that, the core purpose of machine learning is to construct high accuracy classifiers.

Response: Thank you for the great comment, we agree with you. In the newly added section, we have addressed this point and we discussed relevant new techniques.

Overall, techniques handling imbalanced dataset issues of machine learning are interesting and are with significance for discussions. However, authors are required to answer questions or address issues above to improve this manuscript.

Many thanks for your valuable insights