Responses to the review of paper "Testing the Consistency of Performance Scores Reported for Binary Classification Problems" submitted to Applied Soft Computing

György Kovács and Attila Fazekas

June 3, 2024

1 Cover letter

The authors are grateful for the valuable comments of the Handling Editor and all Reviewers contributing to the improvement of the paper.

A short summary of the main changes in the paper:

- Additional references from recent years have been added as requested by the Handling Editor and Reviewer # 1 ([2, 3, 1]).
- Theoretical details have been elaborated (space complexity estimations and subsections on power analysis).
- The limitations of the proposed method have been elaborated in more detail.
- The "Applications" section (Section 5) has been extended.

In the rest of the document we provide detailed responses to the comments of all reviewers. The changes in the revised version of the paper are indicated by blue color.

2 Responses to the comments of the Handling Editor

HE: The paper need a major revisions:

Comment 1 the backgruond must be updated to the last 2 years - 2023 and 2024

Response We added the relevant references [2, 3, 1].

Comment 2 the solution needs a visualization

Response We have added Figure 1 to illustrate the intended use of the proposed methods.

Comment 3 there is no theoretical analysis

Response We believe that the derivation of the proposed methods in the body of the text is rigorous. The only additional theoretical analysis we can think of is estimating the capability of the proposed tests to recognize inconsistencies, namely, the *power analysis*, similarly to statistical hypothesis testing. To improve the paper, we have added multiple subsections and paragraphs where we discuss power analysis both in general and in particular applications.

Comment 4 what is the time/computational complexities?

Response The time complexities were shared in the original version of the paper. In the revised paper we added estimations of space complexities.

Comment 5 experimental section is in the initial state of research - it should be extended to more tests, more statistical analysis and details.

Response The experimental section has been extended, we added more details, results and power analyses as requested.

Comment 6 Also, the solution is not described in term of the current research in this matter.

Response According to our best knowledge, there is limited research providing numerical tools for the meta-analysis of machine learning research, and we cited and discussed the few existing results we are aware of in the Introduction (Section 1). Nevertheless, with the addition of new references, we have enriched the context in both the Introduction (Section 1) and the Conclusions (Section 6). We are fully open to incorporate any specific additional suggestions from the Editor.

3 Responses to the comments of Reviewer # 1

Reviewer #1: I have reviewed the manuscript entitled "Testing the Consistency of Performance Scores Reported for Binary Classification Problems" submitted to Applied Soft Computing (ASOC) journal. I was impressed by the mathematical motivation, ingredients, and reproducibility of the proposed method to a great extent. However, there are some minor issues as follows:

Comment 1 Contribution to recent papers from ASOC is weak. Please cite relevant studies from 2019 to the present (at least three).

Response We added the relevant references [2, 3, 1].

Comment 2 Can you please add another dataset to showcase the superiority of the proposed method?

Response We understand the request of the Reviewer, however:

- The proposed techniques are intended to aid in the meta-analysis of research in fields involving binary classification. Conducting proper meta-analysis in a field requires a deep understanding and expertise in the best practices, datasets, results and potential flaws specific to that fields and worths standalone papers.
- We had previous expertise in the three fields that we used for illustration (retinal imaging, synthetic minority oversampling and skin lesion classification), and the known flaws in these fields motivated the development of the proposed methods. Acquiring the necessary expertise in another field to conduct a meta-analysis is a meticulous task, involving the review of dozens of papers and may not be feasible within the time constraints of a review process. However, there is no reason to assume that other fields where binary classification is used are free from similar flaws.
- The successful application of the proposed methods in three unrelated fields demonstrates their effectiveness.

For these reasons, and primarily because covering another field would require efforts equivalent to writing a separate paper, we respectfully decline to cover another field. *Nevertheless, to improve the paper, and addressing the request of the Handling Editor, we significantly extended the Applications section (Section 5), sharing more technical and statistical details, and carrying out the power analyses for the problems.*

Comment 3 Adding a flowchart to graphically illustrate the developed method is highly recommended (optional).

Response The flowchart of the intended use of the method was added as Figure 1 in the revised paper.

Comment 4 It is of prime importance to elaborate on the limitations and scope of this paper.

Response We added paragraphs discussing various aspects of limitations in section 1 (Introduction), sections 3 and 4 (elaborating the proposed method) as well as in sections 5 (Applications) and Section 6 (Conclusions).

Comment 4 Please double-check the whole manuscript to address potential typos.

Response Spell- and grammar-check were applied.

4 Responses to the comments of Reviewer # 3

There seems to be a copy and paste issue in the review, as the first block contains the same summary 4 times in different wording, which we highlighted by

different shades.

Reviewer #3: The manuscript introduces several assessments aimed at gauging the coherence of reported performance metrics and experimental configurations in binary classification. These tests encompass a wide array of evaluation scenarios. Notably, a specific test is devised to examine the consistency of performance scores derived from a singular evaluation set, accommodating up to 20 distinct metrics. Moreover, when scrutinizing scores aggregated through the Score of Means strategy, the examination reverts to the scenario of scores originating from a single evaluation set. In this manuscript, various evaluations are proposed to ensure the reliability of reported performance measures and experimental setups in binary classification. These evaluations encompass a multitude of scenarios to cover a comprehensive range of possibilities. One noteworthy evaluation focuses on assessing the consistency of performance scores obtained from a single evaluation set, allowing for the examination of 20 different metrics. Additionally, when analyzing scores aggregated using the Score of Means strategy, the evaluation methodology reverts to the scenario of dealing with scores derived from a single evaluation set. The manuscript presents a series of tests designed to assess the consistency of performance scores and experimental designs in binary classification. These tests are diverse in scope, covering a wide range of evaluation scenarios. Particularly, one test is tailored to evaluate the consistency of performance scores originating from a single evaluation set, accommodating as many as 20 distinct metrics. Furthermore, when analyzing scores aggregated via the Score of Means strategy, the testing procedure reverts to the framework applicable to scores derived from a single evaluation set. In this manuscript, a set of tests is introduced to evaluate the reliability of reported performance scores and experimental configurations in binary classification. These tests are comprehensive, considering various evaluation scenarios. Specifically, a test is developed to assess the consistency of performance scores obtained from a single evaluation set, with support for up to 20 different metrics. Moreover, when examining scores aggregated using the Score of Means strategy, the testing protocol aligns with the framework for scores derived from a single evaluation set.

The authors suggested addressing the following comments and suggestions when preparing the revised version:

Comment 1 Abstract: The section needs to be re-drafted to be self-contained, which means it has to clearly show the hypothesis, methodology, techniques, and tools used, and the results obtained.

Response The abstract has been rewritten and split to three paragraphs to clearly reflect the background, methods and results.

Comment 2 Keywords: Authors suggested to update the keywords by selecting more relevant terms. Keywords play an important role in the manuscript's appearance in scholars' searches, which will give it more hits and citations.

Response The keywords have been double-checked.

Comment 3 What assumptions did the authors make during the simulation phase of this research work? If there is any.

Response In there original version of the paper there were no simulations. In the revised paper we conducted the *power analysis* of the proposed method in two applications, which involves simulations. The details of the simulations are shared in the body of the text (Section 5).

Comment 4 What limitations did the authors face during this research work? If there is any.

Response In accordance with Comment 4 of Reviewer #1, we elaborated the limitations of the proposed method at multiple points in the body of the paper. For more details, see our response to Comment 4 of Reviewer #1.

****General comment from Managing Editor : Note that this reviewer requests for inclusion of further references in the paper. Such additional references should be only included by authors if authors have studied these suggested publications carefully and are fully aware of their content and their relevance for the current paper's contribution. It is up to the discretion of the authors to finally decide about inclusion or not. However, inclusion of unrelated references should be strictly avoided.*****

Comment 5 Authors suggested to go through the following references, and they MAY make use of them in updating the introduction and the related work sections:

- Ali Akbar Movassagh, Jafar A. Alzubi, Mehdi Gheisari, Mohamadtaghi Rahimi, Senthil kumar Mohan, Aaqif Afzaal Abbasi, Narjes Nabipour, "Artificial neural networks training algorithm integrating invasive weed optimization with diferential evolutionary model" Journal of Ambient Intelligence Humanized Computing, `https://doi.org/10.1007/s12652-020-02623-6`.
- Omar A. Alzubi, Jafar A. Alzubi, Mohammed Alweshah, Issa Qiqieh, Sara Al-Shami, Manikandan Ramachandran, "An Optimal Pruning Algorithm of Classifier Ensembles: Dynamic Programming Approach" Neural Computing and Applications, 2020.

Response We carefully considered the suggested papers but have decided not to cite them, in accordance with the guidelines provided by the Managing Editor, as we did not find them closely related to the topic of the submitted paper.

Comment 6 Conclusion: The conclusion should be abstracted, so authors need to consider re-drafting it.

Response The "Conclusions" section has been improved.

Comment 7 Authors need to confirm that all acronyms are defined before being used for the first time.

Response The authors double-checked and confirm that.

Comment 8 Authors need to confirm that all mathematical notations are defined when being used for the first time.

Response The authors double checked and confirm that.

Comment 9 The Authors suggested proofreading the manuscript after addressing all comments to avoid typos and grammatical and lingual mistakes.

Response Proofreading, as well as spell- and grammar-check were applied.

References

[1] Salim Rezvani and Xizhao Wang. A broad review on class imbalance learning techniques. *Applied Soft Computing*, 143:110415, 2023.

[2] Szilárd Szabó, Imre J. Holb, Vanda Éva Abriha-Molnár, Gábor Szatmári, Sudhir Kumar Singh, and Dávid Abriha. Classification assessment tool: A program to measure the uncertainty of classification models in terms of class-level metrics. *Applied Soft Computing*, 155:111468, 2024.

[3] Ayfer Ezgi Yilmaz and Haydar Demirhan. Weighted kappa measures for ordinal multi-class classification performance. *Applied Soft Computing*, 134:110020, 2023.