View Reviews

Paper ID

237

Paper Title

Same Score, Different Failures: Understanding MAI Performance Beyond Simple Averages

Reviewer #1

Questions

3. Please categorize the relevancy of the paper (you may choose more than one). Note the different assessment criteria for the different paper categories as outlined in the Reviewer Guidelines: https://conferences.miccai.org/2025/en/REVIEWER-GUIDELINES.html

MIC

4. How would you describe the paper?

Methodological contribution

5. Please describe the main contribution of the paper.

The authors propose and have developed a toolbox (to be made openly available) for the simple yet meaningful statistical (re-)analysis of medical image segmentation challenges. Key methodological components are
i) pairwise statistical significance testing between competitor methods, assessing whether observed differences in performance are statistically significant or not,
ii) a semi-automated choice of dice score vs. normalized surface distance (NSD) as the most appropriate metric for a given task, and
iii) stratified performance analysis by demographic attributes.

A reanalysis of the TotalSegmentator challenge results, as well as results on a proprietary dataset are provided.

6. Please list the major strengths of the paper: you should highlight a novel formulation, an original way to use data, demonstration of clinical feasibility, a novel application, a particularly strong evaluation, or anything else that is a strong aspect of this work. Please provide details, for instance, if a method is novel, explain what aspect is novel and why this is interesting.

All three key methodologies represent valid and important improvements to standard benchmark evaluations: ranking stability and demographic biases should be assessed as a default, and the choice of an appropriate metric is of course crucial.

While these may appear like methodologically trivial tasks, the authors make a very valid point: this is additional (and in fact not completely trivial) work that challenge organizers often simply do not have the time to conduct. As a result, many challenges are methodologically poorly evaluated. The provision of a streamlined toolbox that provides a meaningful analysis with minimal effort for the challenge organizers is thus a potentially very valuable contribution to the community.

7. Please list the major weaknesses of the paper. Please provide details: for instance, if you state that a formulation, way of using data, demonstration of clinical feasibility, or application is not novel, then you must provide specific references to prior work.

The manuscript was clearly submitted in a rush. I would consider it an unfinished draft. There is a leftover paragraph with comments from the drafting stage (end of page 5), the description of Fig.1 does not match its contents, "Title Suppressed Due to Excessive Length", the demographic bias analysis is presented in such little detail as to be essentially not understandable, and there are typos / grammatical errors all over the place.

Many of the methodological choices are either undescribed, unclear, or unmotivated. This is especially crucial given that this is supposed to be a contribution describing methodological best practices.

8. Please rate the clarity and organization of this paper.

Poor

9. Please comment on the reproducibility of the paper. Please be aware that providing code and data is a plus, but not a requirement for acceptance.

The authors claimed to release the source code and/or dataset upon acceptance of the submission.

10. Optional: If you have any additional comments to share with the authors, please provide them here. Please also refer to our Reviewer's guide on what makes a good review and pay specific attention to the different assessment criteria for the different paper categories: https://conferences.miccai.org/2025/en/REVIEWER-GUIDELINES.html

Which statistical test is used to assess significance? This is not specified.

Bonferroni correction is very conservative. Would it not be more appropriate to use e.g. a Holm-Bonferroni correction or another alternative?

It is written in three different places that in Fig. 1, the color would indicate

"Statistical Significance: Method (Left) is Better than Method (Bottom)". However, the matrix is symmetrical, which seems impossible to reconcile. (Method A is highly significantly better than method B while the inverse is also true?)

For the demographic bias assessment, many things remain unclear. Are all possible intersectional sub-sub-subgroups assessed, or just up to a certain interaction level? Is any correction for multiple testing performed here, as well?

The title should be adjusted to reflect the fact that the manuscript is specifically about challenge / benchmark performance evaluation and model ranking.

Finally, as a general remark: statistical significance is important but on its own not a valid indicator of the importance of a difference. To judge whether performance differences between methods or demographic groups are meaningful, this always needs to be combined with a measure of effect size.

11. Rate the paper on a scale of 1-6, 6 being the strongest (6-4: accept; 3-1: reject). Please use the entire range of the distribution. Spreading the score helps create a distribution for decision-making. (Visible to authors.)

2. Reject — should be rejected, independent of rebuttal

12. Please justify your recommendation. What were the major factors that led you to your overall score for this paper?

While in principle a potentially valid contribution, I consider the submission of such unfinished work disrespectful of reviewer time. (Too) many methodological details are missing, unclear, or questionable.

14. In view of your answers above and your overall experience, how would you rate your confidence in your review?

Very confident (4)

Reviewer #2

Questions

3. Please categorize the relevancy of the paper (you may choose more than one). Note the different assessment criteria for the different paper categories as outlined in the Reviewer Guidelines: https://conferences.miccai.org/2025/en/REVIEWER-GUIDELINES.html

MIC

4. How would you describe the paper?

Methodological contribution

5. Please describe the main contribution of the paper.

In this paper the authors touch on several important shortcomings of current method and challenge evaluations, like statistical significance testing,, choice of metric and demographic parity. They propose a toolkit to produce these insights.

6. Please list the major strengths of the paper: you should highlight a novel formulation, an original way to use data, demonstration of clinical feasibility, a novel application, a particularly strong evaluation, or anything else that is a strong aspect of this work. Please provide details, for instance, if a method is novel, explain what aspect is novel and why this is interesting.

As evidenced by earlier papers, the highlighted shortcomings lead to skewed method evaluations and hinder a fair evaluation of methods in challenges. This leads, among other things, to methods gaining popularity and usage by the community, only due to skewed evaluations.

7. Please list the major weaknesses of the paper. Please provide details: for instance, if you state that a formulation, way of using data, demonstration of clinical feasibility, or application is not novel, then you must provide specific references to prior work.

I have three main issues with this paper:
1. The paper promises to present a toolkit for better evaluation of medical AI challenges. However, what it does is just to show some results from significance tests, different metrics and Demographic Parity DIfference.
2. It doesn't really introduce anything new, except for maybe putting multiple existing works in a common toolkit. It is unclear to me, where the actual benefit comes from. Metric selection is more complex than choosing DSC or NSD based on the organ to be segmented, which is usually not the task in most challenges. This is exactly where the metrics reloaded framework comes in, which proposes appropriate metrics for the specific task. Ranking analyses are also covered by a toolkit, which generates a full report, provided case-wise results. Insights like NSD is better for some structures then DSC is nothing new, the fact that rankings can change based on the metric also not, especially if these metrics are not evaluated on the same structures.
3. This paper doesn't seem to be ready for submission. There are still instructions left at the bottom of page 5. In addition, the paper lacks content. In the abstract, the authors claim that they analyze the results of several recent challenges, in the final paper, they only show some results on TotalSegmentator and a proprietary dataset.

8. Please rate the clarity and organization of this paper.

Satisfactory

9. Please comment on the reproducibility of the paper. Please be aware that providing code and data is a plus, but not a requirement for acceptance.

The authors claimed to release the source code and/or dataset upon acceptance of the submission.

11. Rate the paper on a scale of 1-6, 6 being the strongest (6-4: accept; 3-1: reject). Please use the entire range of the distribution. Spreading the score helps create a distribution for decision-making. (Visible to authors.)

2. Reject — should be rejected, independent of rebuttal

12. Please justify your recommendation. What were the major factors that led you to your overall score for this paper?

See the list of weaknesses above. This paper lacks novelty and is not ready for publication.

14. In view of your answers above and your overall experience, how would you rate your confidence in your review?

Very confident (4)

Reviewer #3

Questions

3. Please categorize the relevancy of the paper (you may choose more than one). Note the different assessment criteria for the different paper categories as outlined in the Reviewer Guidelines: https://conferences.miccai.org/2025/en/REVIEWER-GUIDELINES.html

MIC

4. How would you describe the paper?

Methodological contribution

5. Please describe the main contribution of the paper.

Improving the accessibility and useability of medical AI challenge results processing via an open-source toolkit.

6. Please list the major strengths of the paper: you should highlight a novel formulation, an original way to use data, demonstration of clinical feasibility, a novel application, a particularly strong evaluation, or anything else that is a strong aspect of this work. Please provide details, for instance, if a method is novel, explain what aspect is novel and why this is interesting.

Potentially valuable toolkit for processing of Challenge results enabling more insight into method comparisons. Collects together relevant prior work. Illustration with reference TotalSegmentator challenge is informative and provides interesting examples of how interpretation of results can change with respect to statistical significance tests, appropriate metrics and demographic performance.

7. Please list the major weaknesses of the paper. Please provide details: for

instance, if you state that a formulation, way of using data, demonstration of clinical feasibility, or application is not novel, then you must provide specific references to prior work.

Figure 1 confuses me: for instance Diff-UNet appears to be better than SAM-Adapter and SAM-Adapter appears to be better than Diff-UNet. Either I've misunderstood this figure or the figure is incorrect and should not be symmetric about the diagonal (above diagonal should be inverse of below diagonal).
The claimed benefit of increasing useability is difficult to assess without specific illustration or access to the toolkit.

8. Please rate the clarity and organization of this paper.

Satisfactory

9. Please comment on the reproducibility of the paper. Please be aware that providing code and data is a plus, but not a requirement for acceptance.

The submission has provided an anonymized link to the source code, dataset, or any other dependencies.

10. Optional: If you have any additional comments to share with the authors, please provide them here. Please also refer to our Reviewer's guide on what makes a good review and pay specific attention to the different assessment criteria for the different paper categories: https://conferences.miccai.org/2025/en/REVIEWER-GUIDELINES.html

The final paragraph on page 5 ("Another figure: DSC- (Add NSD), by organ size...") is unclear. Is this a "to Do" comment that should have been deleted?

11. Rate the paper on a scale of 1-6, 6 being the strongest (6-4: accept; 3-1: reject). Please use the entire range of the distribution. Spreading the score helps create a distribution for decision-making. (Visible to authors.)

4. Weak Accept — could be accepted, dependent on rebuttal

12. Please justify your recommendation. What were the major factors that led you to your overall score for this paper?

Collates existing methods. Potentially very useful but claimed "accessibility and useability" novelty is difficult to assess.

14. In view of your answers above and your overall experience, how would you rate your confidence in your review?

Very confident (4)