Metrics Reloaded – A new recommendation framework for biomedical image analysis validation

Annika Reinke [1,2] A.REINKE@DKFZ-HEIDELBERG.DE
[1] *Div. Intelligent Medical Systems, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[2] *Helmholtz Imaging, German Cancer Research Center (DKFZ), Heidelberg, Germany*
Metrics Reloaded Delphi Consortium *

Abstract

Meaningful performance assessment of biomedical image analysis algorithms depends on objective and appropriate performance metrics. There are major shortcomings in the current state of the art. Yet, so far limited attention has been paid to practical pitfalls associated when using particular metrics for image analysis tasks. Therefore, a number of international initiatives have collaborated to offer researchers with guidance and tools for selecting performance metrics in a problem-aware manner. In our proposed framework, the characteristics of the given biomedical problem are first captured in a problem fingerprint, which identifies properties related to domain interests, the target structure(s), the input datasets, and algorithm output. A problem category-specific mapping is applied in the second step to match fingerprints to metrics that reflect domain requirements. Based on input from experts from more than 60 institutions worldwide, we believe our metric recommendation framework to be useful to the MIDL community and to enhance the quality of biomedical image analysis algorithm validation.

Keywords: Validation, Metrics, Good Scientific Practice, Classification, Semantic Segmentation, Object Detection, Instance Segmentation, Medical Imaging.

1. Framework for metric selection

Validating a biomedical image analysis algorithm is crucial to enhance medical decision making and healthcare delivery. Metrics are the performance measures that typically compare a reference annotation with the prediction generated by an image analysis algorithm. As an indicator of how well an algorithm reproduces a reference, they are thus essential to evaluating the performance of image analysis algorithms in an objective manner. Despite previous work on the clinical relevance of metrics (Vaassen et al., 2020), researchers lack guidelines for selecting the right metric for a given biomedical image analysis problem (Maier-Hein et al., 2018). Our earlier work highlighted several limitations of metrics in biomedical image analysis (Reinke et al., 2021). This MIDL presentation summarizes our proposed metric recommendation framework, compiled by an expert panel of 68 researchers. Our framework enables users to make educated decisions about which validation metrics

* **Full author list:** A. Reinke, L. Maier-Hein, E. Christodoulou, B. Glocker, P. Godau, F. Isensee, J. Kleesiek, M. Kozubek, M. Reyes, M. Riegler, M. Wiesenfarth, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, A.E. Kavur, T. Rädsch, M.D. Tizabi, L. Acion, M. Antonelli, T. Arbel, S. Bakas, P. Bankhead, A. Benis, M.J. Cardoso, V. Cheplygina, B. Cimini, G. Collins, K. Farahani, B. van Ginneken, F. Hamprecht, D. Hashimoto, M. Hoffman, M. Huisman, P. Jannin, C.E. Kahn, A. Karargyris, A. Karthikesalingam, H. Kenngott, A. Kopp-Schneider, A. Kreshuk, T. Kurc, B.A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A.L. Martel, P. Mattson, E. Meijering, B. Menze, D. Moher, K.G.M. Moons, H. Müller, B. Nichyporuk, F. Nickel, J. Petersen, N. Rajpoot, N. Rieke, J. Saez-Rodriguez, C. Sánchez Guitiérrez, S. Shetty, M. van Smeden, C.H. Sudre, R. Summers, A.A. Taha, S.A. Tsaftaris, B. Van Calster, G. Varoquaux, P. Jäger.

to use for a given biomedical problem and addresses the community request for guidance overcoming the presented limitations. Two main steps are involved in choosing metrics for a given biomedical question (see Figure 1):

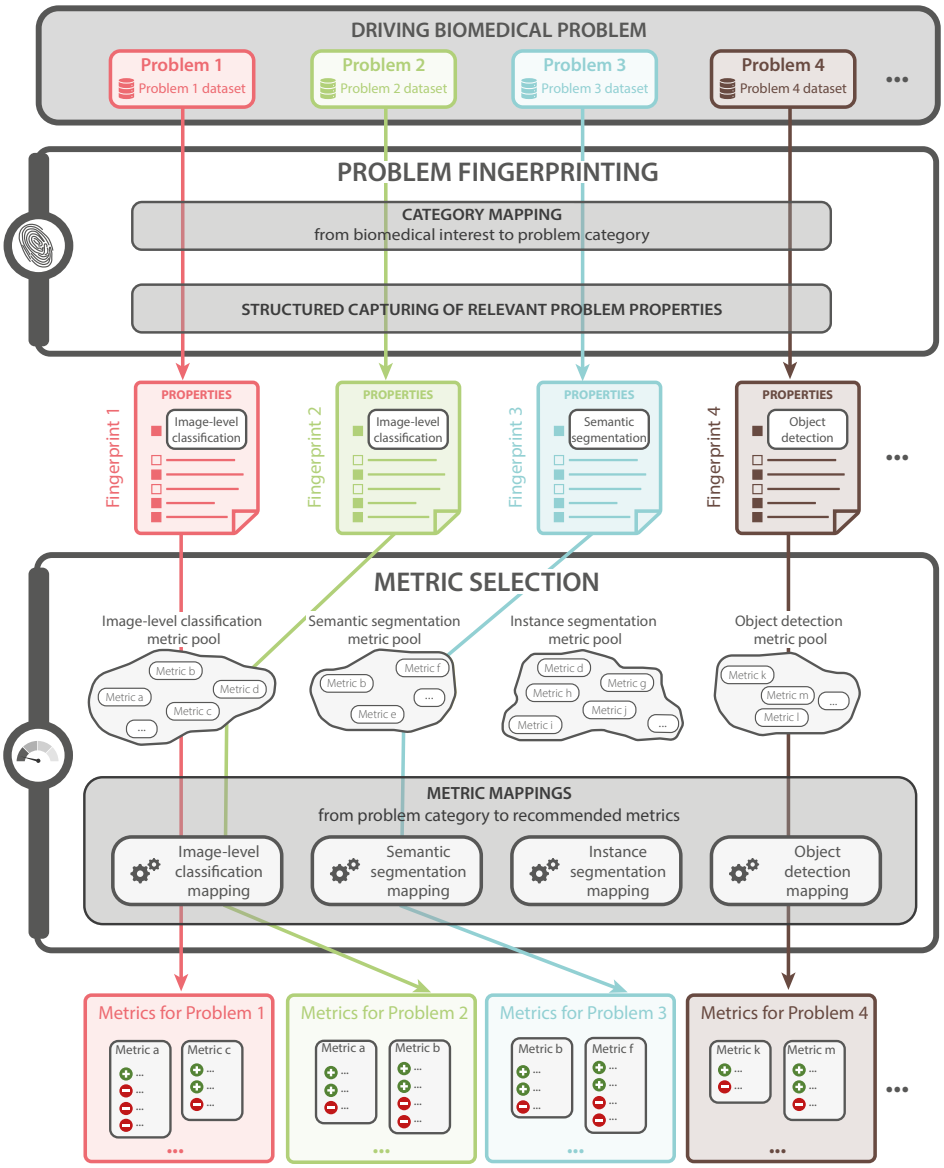

Figure 1: **Framework for metric selection.** Choosing metrics for a biomedical problem is achieved in two main steps. First, the **problem fingerprint** is generated, which captures the characteristics of the driving biomedical problem relevant for metric selection. During this step, the category mapping assigns the given problem to the appropriate image processing category, namely image-level classification, semantic segmentation, object detection and instance segmentation. Finally, the **metric selection** step generates a set of suitable metrics from a category-specific pool of metrics in a domain interest-aware manner.

Problem fingerprint: The user is guided towards the generation of a problem fingerprint, which refers to a structured representation of a given biomedical problem that captures its context and characteristics. In addition to general properties, such as the dimensionality of the input image, it contains information about the biomedical domain interest, the target structure(s), the input datasets, and the algorithm output. During this step, the research problem is assigned to the suitable image analysis category (currently, we cover image-level classification, semantic segmentation, instance segmentation and object detection). This step serves the purpose of grouping problems based on their validation similarity while abstracting from the individual algorithms that were used to solve the problems.

Metric selection: Once the problem fingerprint is identified, the researcher is guided through selecting an appropriate set of metrics while considering potential risks that may be associated with the specific characteristics of the underlying biomedical problem. This step will be facilitated by a web toolkit (currently under construction) that makes the complete metric recommendation framework accessible and usable for researchers including non-experts.

2. Conclusion

The choice of the most suitable metric(s) for a particular image processing problem is not trivial. In our MIDL presentation, we outline a framework for choosing the most appropriate metric(s) with respect to the biomedical context. An upcoming publication will provide details regarding the metric recommendation framework as well as regarding the Metrics Reloaded Delphi Consortium that compiled the guidance process.

Acknowledgments

This work was initiated by Helmholtz Imaging (HI) and supported by the NIH Clinical Center Intramural Research Program, the NIH National Cancer Institute (NCI: U01CA242871) and the NIH National Institute of Neurological Disorders and Stroke (NINDS: R01NS042645).

References

Lena Maier-Hein et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature communications*, 9(1):1–13, 2018.

Annika Reinke et al. Common limitations of image processing metrics: A picture story. *arXiv preprint arXiv:2104.05642*, 2021.

Femke Vaassen et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology*, 13: 1–6, 2020.