Metrics Reloaded

Annika Reinke[1]* **Lena Maier-Hein**[1]* **Metrics Reloaded Consortium**†

[1]Div. Intelligent Medical Systems, HI Helmholtz Imaging, German Cancer Research Center (DKFZ), Germany
{a.reinke, l.maier-hein}@dkfz-heidelberg.de

Abstract

Flaws in machine learning (ML) algorithm validation are an underestimated global problem. Particularly in automatic biomedical image analysis, chosen performance metrics often do not reflect the domain interest, thus failing to adequately measure scientific progress and hindering translation of ML techniques into practice. A large international expert consortium now created *Metrics Reloaded*, a comprehensive framework guiding researchers towards problem-aware metric selection. The framework is based on the novel concept of a problem fingerprint – a structured representation of the given problem that captures all aspects relevant for metric selection, from the domain interest to properties of the target structure(s), data set and algorithm output. It supports image-level classification, object detection, semantic and instance segmentation tasks. Users are guided through the process of selecting and applying appropriate validation metrics while being made aware of pitfalls. To improve the user experience, we implemented the framework in an online tool, which also provides a common point of access to explore metric weaknesses and strengths. An instantiation of the framework for various biomedical image analysis use cases demonstrates its broad applicability across domains.

1 Introduction

Machine learning (ML)-based automated image processing is gaining increasing traction in biomedical imaging. The critical issue of reliable and objective performance assessment of algorithms, however, remains largely unexplored. Algorithm performance is commonly assessed through validation metrics that should serve as proxies for the domain interest. They measure scientific progress in the field and are the basis for deciding on the practical suitability of algorithms, thus a key component for translation into biomedical practice. Validation not conducted according to relevant metrics could be a major reason for why many ML developments in biomedical imaging fail to reach practice.

Increasing evidence shows that the metrics used in common practice often do not adequately reflect the underlying biomedical problems, diminishing the validity of the investigated algorithms [2, 3, 4, 5, 7]. Among a number of shortcomings recently unveiled by a multi-center initiative [5] performing the first comprehensive evaluation of biomedical image analysis competitions, for example, the choice of inappropriate metrics stood out as a core problem. Pitfalls in metric choice commonly relate to either an inappropriate phrasing of the problem (e.g. disregarding the type of task), poor metric selection (e.g. disregarding mathematical properties), or poor metric application (e.g. disregarding hierarchical data structure). To dismantle such – often historically grown – poor practices, we established the multidisciplinary *Metrics Reloaded* consortium, comprising 75 international experts from various relevant fields. Its mission is to foster reliable algorithm validation through problem-aware metric choice with the long-term goal of (1) enabling the reliable tracking of scientific progress and (2)

[*] Shared first authors.

[†] **Full author list:** A. Reinke, L. Maier-Hein, P. Godau, E. Christodoulou, B. Glocker, 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.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

bridging the current chasm between AI research and practical translation. The purpose of this submission is to present the *Metrics Reloaded* framework to the image analysis community and discuss these recommendations that go beyond the state of the art.

2 Results

We first identified metric-related pitfalls in common practice and found these to generalize across different imaging domains and modalities. We hence followed a multidisciplinary cross-domain approach critically questioning common practice in different communities. The resulting *Metrics Reloaded* recommendation framework (Fig. 1) was developed using a multi-stage Delphi process [1] comprising numerous international workshops, questionnaires, expert group meetings and crowd-sourced feedback processes. Its foundation is the new concept of *problem fingerprinting*. Abstracting from a specific domain, problem fingerprinting is the generation of a structured representation of a given biomedical problem that captures all properties relevant for metric selection. These encompass domain interest-related properties (e.g. particular importance of structure boundary), target structure-related properties (e.g. shape complexity or relative structure size), data set-related properties (e.g. class imbalance), and algorithm output-related properties, (e.g. possibility of algorithm output not containing any target structure). Based on this fingerprint, the user is then transparently guided through the process of selecting an appropriate set of metrics while being made aware of potential pitfalls related to the underlying biomedical problem. The framework supports problems in which categorical output for a given *n*-dimensional input image (possibly enhanced with context information) is sought at pixel, object or image level, that is, image-level classification, object detection, semantic or instance segmentation tasks. The following key design decisions shaped our framework:

Encapsulating domain knowledge: Our approach encapsulates domain knowledge by capturing the properties relevant for metric selection in a problem fingerprint that ultimately abstracts from the specific image modality and domain of a given problem.

Exploiting synergies across classification scales: We address all tasks that can be considered classification tasks at pixel, object or image level in one common framework that also covers the selection of the problem category itself, a common pitfall.

Exploiting complementary metric strengths: To account for complementary metric strengths and weaknesses [6], the framework generally recommends the use of multiple metrics. This can include the metrics not commonly used in the biomedical imaging communities.

Abstracting from methodology: Metrics are chosen based solely on the driving biomedical problem, not based on algorithm design choices.

Involving and educating users: Rather than providing a black box recommendation, the framework guides users through the entire process while raising awareness on potential pitfalls. Decision guides assist in cases of difficult tradeoffs while respecting individual preferences.

To validate the framework, we instantiated it for several common use cases across various biomedical image processing domains and modalities. We found shared properties of problems from different domains - captured by the fingerprint - to result in almost identical recommendations, confirming the framework's feasibility and broad applicability. Finally, to improve user experience and encourage widespread adoption, we implemented the framework in an online tool which will soon be publicly available.

3 Conclusion

The novel *Metrics Reloaded* framework enables problem-aware selection of validation metrics in biomedical image analysis, thus addressing a major roadblock in practical ML translation.

It should be noted that our framework only supports classification tasks at pixel, object or image level. It only addresses reference-based validation metrics; non-reference-based metrics such as algorithm runtime or carbon footprint are not yet included. As all recommendations are based on the generation of a problem fingerprint, their reliability depends on users providing correct information and thoroughly answering the questions posed. Incorrect information by users in relation to the underlying research problem may result in inadequate metric selection and ultimately yield negative consequences for patients.

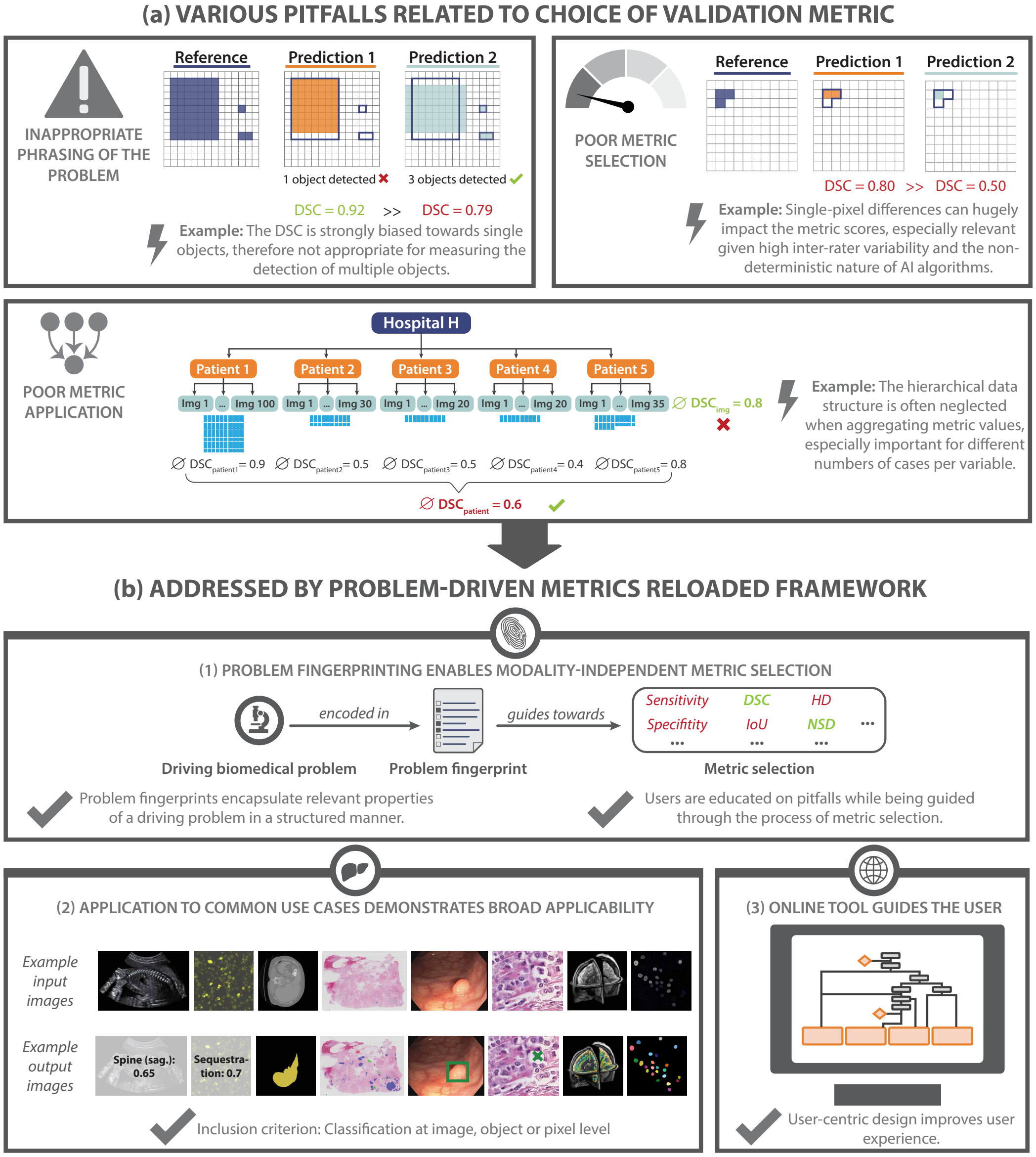
(a) VARIOUS PITFALLS RELATED TO CHOICE OF VALIDATION METRIC
INAPPROPRIATE PHRASING OF THE PROBLEM
Reference
Prediction 1
Prediction 2
1 object detected ✗
3 objects detected ✓
DSC = 0.92 >> DSC = 0.79
Example: The DSC is strongly biased towards single objects, therefore not appropriate for measuring the detection of multiple objects.
POOR METRIC SELECTION
Reference
Prediction 1
Prediction 2
DSC = 0.80 >> DSC = 0.50
Example: Single-pixel differences can hugely impact the metric scores, especially relevant given high inter-rater variability and the non-deterministic nature of AI algorithms.
POOR METRIC APPLICATION
Hospital H
Patient 1
Patient 2
Patient 3
Patient 4
Patient 5
Img 1 ... Img 100
Img 1 ... Img 30
Img 1 ... Img 20
Img 1 ... Img 20
Img 1 ... Img 35
∅ DSC_img = 0.8 ✗
∅ DSC_patient1 = 0.9
∅ DSC_patient2 = 0.5
∅ DSC_patient3 = 0.5
∅ DSC_patient4 = 0.4
∅ DSC_patient5 = 0.8
∅ DSC_patient = 0.6 ✓
Example: The hierarchical data structure is often neglected when aggregating metric values, especially important for different numbers of cases per variable.
(b) ADDRESSED BY PROBLEM-DRIVEN METRICS RELOADED FRAMEWORK
(1) PROBLEM FINGERPRINTING ENABLES MODALITY-INDEPENDENT METRIC SELECTION
encoded in
guides towards
Sensitivity DSC HD
Specifity IoU NSD ...
Driving biomedical problem
Problem fingerprint
Metric selection
Problem fingerprints encapsulate relevant properties of a driving problem in a structured manner.
Users are educated on pitfalls while being guided through the process of metric selection.
(2) APPLICATION TO COMMON USE CASES DEMONSTRATES BROAD APPLICABILITY
Example input images
Example output images
Spine (sag.): 0.65
Sequestration: 0.7
Inclusion criterion: Classification at image, object or pixel level
(3) ONLINE TOOL GUIDES THE USER
User-centric design improves user experience.

Figure 1: **Contributions of the *Metrics Reloaded* framework. a)** Motivation: Common problems related to metrics typically arise from (top left) inappropriate phrasing of the problem (here: object detection confused with semantic segmentation), (top right) poor metric selection (here: neglecting the small size of structures) and (bottom) poor metric application (here: inappropriate aggregation scheme). **b)** *Metrics Reloaded* addresses these pitfalls. (1) To enable the selection of metrics that match the domain interest, the framework is based on the new concept of *problem fingerprinting*; the generation of a structured representation of the given biomedical problem that captures all properties that are relevant for metric selection. Based on the problem fingerprint, *Metrics Reloaded* guides the user through the process of metric selection and application while raising awareness of relevant pitfalls. (2) An instantiation of the framework for common biomedical use cases demonstrates its broad applicability. (3) A publicly available online tool facilitates application of the framework.

4 Acknowledgements

This work was initiated by the Helmholtz Association of German Research Centers in the scope of the Helmholtz Imaging Incubator (HI), the MICCAI special interest group on biomedical image analysis challenges and the benchmarking working group of the MONAI initiative. It was further supported in part by the Intramural Research Program of the National Institutes of Health Clinical Center as well as by the National Cancer Institute (NCI) and the National Institute of Neurological Disorders and Stroke (NINDS) of the National Institutes of Health (NIH), under award numbers NCI:U01CA242871 and NINDS:R01NS042645. The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH. S.A. Tsaftaris acknowledges the support of Canon Medical and the Royal Academy of Engineering and the Research Chairs and Senior Research Fellowships scheme (grant RCSRF1819\825).

We thank Nina Sauter for proposing the name for the project.

References

[1] B. B. Brown. Delphi process: a methodology used for the elicitation of opinions of experts. Technical report, Rand Corp Santa Monica CA, 1968.

[2] P. Correia and F. Pereira. Video object relevance metrics for overall segmentation quality evaluation. *EURASIP Journal on Advances in Signal Processing*, 2006:1–11, 2006.

[3] M. J. Gooding, A. J. Smith, M. Tariq, P. Aljabar, D. Peressutti, J. van der Stoep, B. Reymen, D. Emans, D. Hattu, J. van Loon, et al. Comparative evaluation of autocontouring in clinical practice: a practical method using the turing test. *Medical physics*, 45(11):5105–5115, 2018.

[4] F. Kofler, I. Ezhov, F. Isensee, C. Berger, M. Korner, J. Paetzold, H. Li, S. Shit, R. McKinley, S. Bakas, et al. Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient. *arXiv preprint arXiv:2103.06205v1*, 2021.

[5] L. Maier-Hein, M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz, T. Arbel, H. Bogunovic, A. P. Bradley, A. Carass, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature communications*, 9(1):1–13, 2018.

[6] A. Reinke, M. Eisenmann, M. D. Tizabi, C. H. Sudre, T. Rädsch, M. Antonelli, T. Arbel, S. Bakas, M. J. Cardoso, V. Cheplygina, et al. Common limitations of image processing metrics: A picture story. *arXiv preprint arXiv:2104.05642*, 2021.

[7] F. Vaassen, C. Hazelaar, A. Vaniqui, M. Gooding, B. van der Heyden, R. Canters, and W. van Elmpt. 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.