Metrics reloaded: Recommendations for image analysis validation

LENA MAIER-HEIN[*][†], German Cancer Research Center (DKFZ), Germany, Heidelberg University, Germany, and National Center for Tumor Diseases (NCT), Germany
ANNIKA REINKE[*], German Cancer Research Center (DKFZ), Germany and Heidelberg University, Germany
PATRICK GODAU, German Cancer Research Center (DKFZ), Germany and Heidelberg University, Germany
MINU D. TIZABI, German Cancer Research Center (DKFZ), Germany
EVANGELIA CHRISTODOULOU, German Cancer Research Center (DKFZ), Germany
BEN GLOCKER, Imperial College London, UK
FABIAN ISENSEE, German Cancer Research Center (DKFZ), Germany
JENS KLEESIEK, University Medicine Essen, Germany
MICHAL KOZUBEK, Masaryk University, Czech Republic
MAURICIO REYES, University of Bern, Switzerland
MICHAEL A. RIEGLER, Simula Metropolitan Center for Digital Engineering, Norway and UiT The Arctic University of Norway, Norway
MANUEL WIESENFARTH, German Cancer Research Center (DKFZ), Germany
MICHAEL BAUMGARTNER, German Cancer Research Center (DKFZ), Germany
MATTHIAS EISENMANN, German Cancer Research Center (DKFZ), Germany
DOREEN HECKMANN-NÖTZEL, German Cancer Research Center (DKFZ), Germany and National Center for Tumor Diseases (NCT), Germany
A. EMRE KAVUR, German Cancer Research Center (DKFZ), Germany
TIM RÄDSCH, German Cancer Research Center (DKFZ), Germany
LAURA ACION, CONICET – Universidad de Buenos Aires, Argentina and University of Iowa, USA
MICHELA ANTONELLI, King's College London, UK and University College London, UK
TAL ARBEL, McGill University, Canada
SPYRIDON BAKAS, University of Pennsylvania, USA and Perelman School of Medicine at the University of Pennsylvania, USA
PETER BANKHEAD, University of Edinburgh, UK
ARRIEL BENIS, Holon Institute of Technology, Israel
M. JORGE CARDOSO, King's College London, UK and University College London, UK
VERONIKA CHEPLYGINA, IT University of Copenhagen, Denmark
BETH CIMINI, Broad Institute of MIT and Harvard, USA
GARY S. COLLINS, University of Oxford, UK
KEYVAN FARAHANI, National Cancer Institute, USA
LUCIANA FERRER, CONICET-UBA, Argentina
ADRIAN GALDRAN, Universitat Pompeu Fabra, Spain and University of Adelaide, Australia
BRAM VAN GINNEKEN, Fraunhofer MEVIS, Germany and Radboud University Medical Center, The Netherlands
ROBERT HAASE, DFG Cluster of Excellence „Physics of Life", Germany and Center for Systems Biology, Germany
DANIEL A. HASHIMOTO, Case Western Reserve University School of Medicine, USA
MICHAEL M. HOFFMAN, University Health Network, Canada, University of Toronto, Canada, and Vector Institute, Canada
MEREL HUISMAN, Radboud University Medical Center, The Netherlands
PIERRE JANNIN, Université de Rennes 1, Inserm, France
CHARLES E. KAHN, University of Pennsylvania, USA
DAGMAR KAINMUELLER, Max-Delbrück Center for Molecular Medicine, Germany
BERNHARD KAINZ, Imperial College London, UK
ALEXANDROS KARARGYRIS, IHU Strasbourg, France
ALAN KARTHIKESALINGAM, Google Health Deepmind, UK
HANNES KENNGOTT, Heidelberg University Hospital, Germany
FLORIAN KOFLER, Helmholtz AI, Germany

ANNETTE KOPP-SCHNEIDER, German Cancer Research Center (DKFZ), Germany
ANNA KRESHUK, European Molecular Biology Laboratory (EMBL), Germany
TAHSIN KURC, Stony Brook University, USA
BENNETT A. LANDMAN, Vanderbilt University, USA
GEERT LITJENS, Radboud University Medical Center, The Netherlands
AMIN MADANI, University Health Network, Canada
KLAUS MAIER-HEIN, German Cancer Research Center (DKFZ), Germany and Heidelberg University Hospital, Germany
ANNE L. MARTEL, Sunnybrook Research Institute, Canada and University of Toronto, Canada
PETER MATTSON, Google, USA
ERIK MEIJERING, University of New South Wales, Australia
BJOERN MENZE, University of Zurich, Switzerland
DAVID MOHER, Ottawa Hospital Research Institute, Canada and University of Ottawa, Canada
KAREL G.M. MOONS, UMC Utrecht, University Utrecht, The Netherlands
HENNING MÜLLER, University of Applied Sciences Western Switzerland (HES-SO), Switzerland and University of Geneva, Switzerland
BRENNAN NICHYPORUK, MILA (Quebec Artificial Intelligence Institute), Canada
FELIX NICKEL, Heidelberg University Hospital, Germany
JENS PETERSEN, German Cancer Research Center (DKFZ), Germany
NASIR RAJPOOT, University of Warwick, UK
NICOLA RIEKE, NVIDIA GmbH, Germany
JULIO SAEZ-RODRIGUEZ, Heidelberg University, Germany, Heidelberg University Hospital, Germany, and BioQuant, Germany
CLARISA SÁNCHEZ GUTIÉRREZ, University of Amsterdam, The Netherlands
SHRAVYA SHETTY, Google, USA
MAARTEN VAN SMEDEN, University Medical Center Utrecht, The Netherlands
CAROLE H. SUDRE, University College London, UK and King's College London, UK
RONALD M. SUMMERS, National Institutes of Health, USA
ABDEL A. TAHA, Scigility International GmbH, Austria
SOTIRIOS A. TSAFTARIS, The University of Edinburgh, Scotland
BEN VAN CALSTER, Katholieke Universiteit (KU) Leuven, Belgium and Leiden University Medical Center, The Netherlands
GAËL VAROQUAUX, INRIA Saclay-Île de France, France
PAUL F. JÄGER, German Cancer Research Center (DKFZ), Germany

Disclaimer

This document reflects intermediate results of a multi-stage Delphi process on metrics conducted by an international consortium of more than 70 image analysis experts. By making the current status of the Delphi process publicly accessible, we seek to obtain constructive feedback by the community in order to optimize the best practice recommendations in the final version of the paper.

[*] Shared first authors.
[†] The complete list of affiliations can be found in App. K.

Abstract:

Increasing evidence shows that 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. To overcome this, a large international expert consortium created *Metrics Reloaded*, a comprehensive framework guiding researchers towards choosing metrics in a problem-aware manner. Following the convergence of ML methodology across application domains, *Metrics Reloaded* fosters the convergence of validation methodology. The framework was developed in a multi-stage Delphi process and is based on the novel concept of a *problem fingerprint* – a structured representation of the given problem that captures all aspects that are relevant for metric selection from the domain interest to the properties of the target structure(s), data set and algorithm output. *Metrics Reloaded* targets image analysis problems that can be interpreted as a classification task at image, object or pixel level, namely *image-level classification*, *object detection*, *semantic segmentation*, and *instance segmentation* tasks. Users are guided through the process of selecting and applying appropriate validation metrics while being made aware of potential pitfalls. To improve the user experience, we implemented the framework in the *Metrics Reloaded* online tool, which also provides a common point of access to explore weaknesses and strengths of the most common validation metrics. An instantiation of the framework for various biological and medical image analysis use cases demonstrates its broad applicability across domains.

Keywords: Challenges, Competitions, Validation, Evaluation, Metrics, Good Scientific Practice, Image Processing, Image Analysis, Computer Vision, Classification, Segmentation, Instance Segmentation, Semantic Segmentation, Detection, Localization, Medical Imaging, Biological Imaging

INTRODUCTION

Machine learning (ML)-based automated image processing is gaining increasing traction in biological and medical imaging research and practice. So far, research has predominantly focused on the development of new image processing algorithms. The critical issue of reliable and objective performance assessment of these algorithms, however, remains largely unexplored. While the suitability of new medical treatments is typically directly assessed via well-interpretable and clinically meaningful measures, such as survival and complication rates, algorithm performance in image processing is commonly assessed with so-called validation metrics[1] that should serve as proxies for the domain interest. In consequence, the impact of validation metrics cannot be overstated; first, *they measure the scientific progress in the field*, such that all future developments directly depend on them. Second, they are the basis for deciding on the practical (e.g. clinical) suitability of an algorithm and are thus *a key component for translation into biomedical practice*. In fact, validation that is not conducted according to relevant metrics could be one major reason for why many Artificial Intelligence (AI) developments in medical imaging fail to reach (clinical) practice.

Despite their importance, an increasing body of work shows that the metrics used in common practice often do not adequately reflect the underlying biomedical problems, diminishing the validity of the investigated algorithms [31, 36, 41, 52, 54, 58, 62, 88]. This especially holds true for so-called challenges, internationally respected competitions that have over the last few years become the de facto standard for comparative performance assessment of Machine Learning (ML) algorithms and other methods. These challenges are often published in prestigious journals [22, 73, 87] and receive tremendous attention from both the scientific community and industry. Among a number of shortcomings in design and quality control that were recently unveiled by a multicenter initiative [58] performing the first comprehensive evaluation of biomedical image analysis challenges, the choice of inappropriate metrics stood out as a core problem. Compared to other areas of AI research, choosing the right metric is particularly challenging in image processing because the suitability of a metric depends on various factors. As a foundation for the present work, we identified three core categories related to pitfalls in metric selection (see Fig. 1a):

Inappropriate phrasing of the problem: The chosen metrics do not always reflect the biomedical need. For example, the popular segmentation metric Dice Similarity Coefficient (DSC) is frequently chosen as a metric for object detection problems [19, 45] although it does not reflect the critical need of detecting as many structure instances as possible (Fig. 1a, top left).

Poor metric selection: Certain characteristics of a given biomedical problem render particular metrics inadequate. Mathematical metric properties are often neglected, for example when using the DSC in the presence of particularly small structures (Fig. 1a, top right).

Poor metric application: Even if a metric is well-suited for a given problem in principle, pitfalls can occur when applying that metric to a specific data set. For example, a common flaw pertains to ignoring hierarchical data structure, as in data from multiple hospitals or a variable number of images per patient (Fig. 1a, bottom), when aggregating metric values.

[1] Not to be confused with distance metrics in the pure mathematical sense.

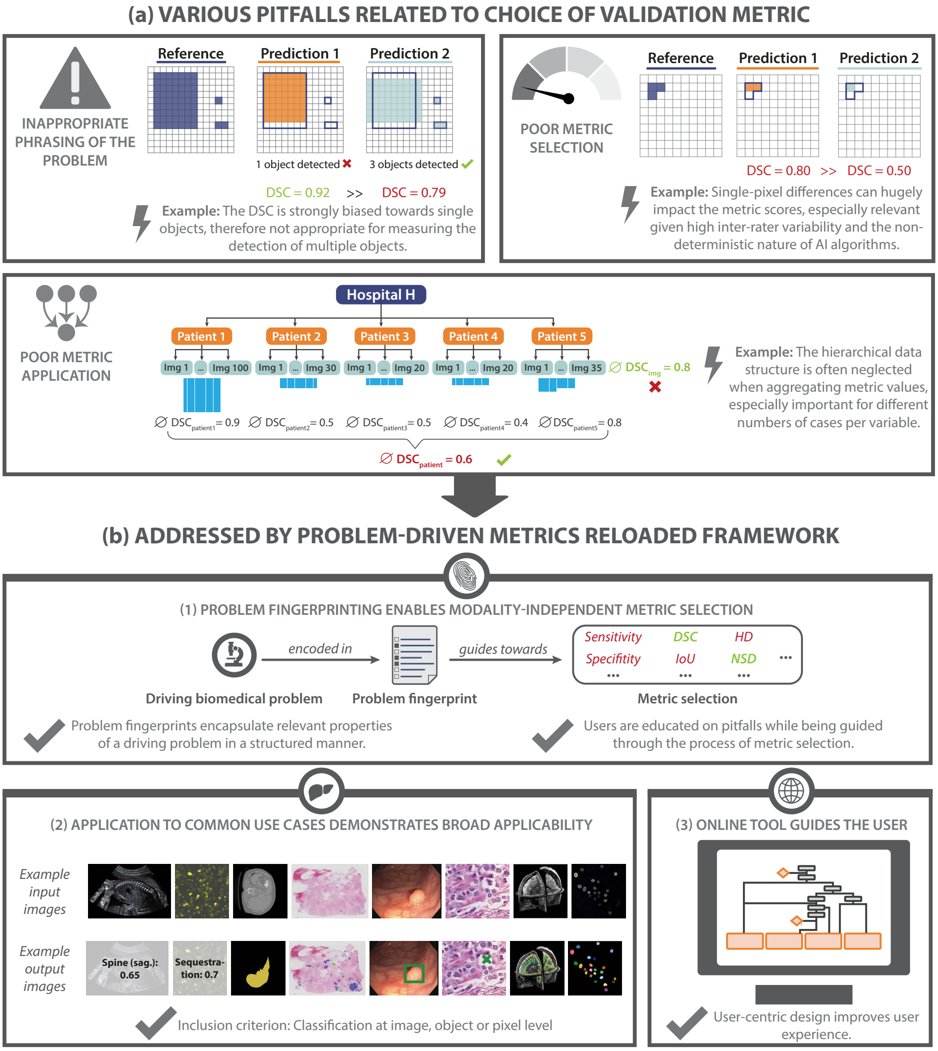

(a) VARIOUS PITFALLS RELATED TO CHOICE OF VALIDATION METRIC
INAPPROPRIATE PHRASING OF THE PROBLEM
Reference
Prediction 1
Prediction 2
1 object detected ✗
3 objects detected ✓
DSC = 0.92 >> DSC = 0.79
Example: The DSC is strongly biased towards single objects, therefore not appropriate for measuring the 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.

Fig. 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). ∅ refers to the average *DSC* values. **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.

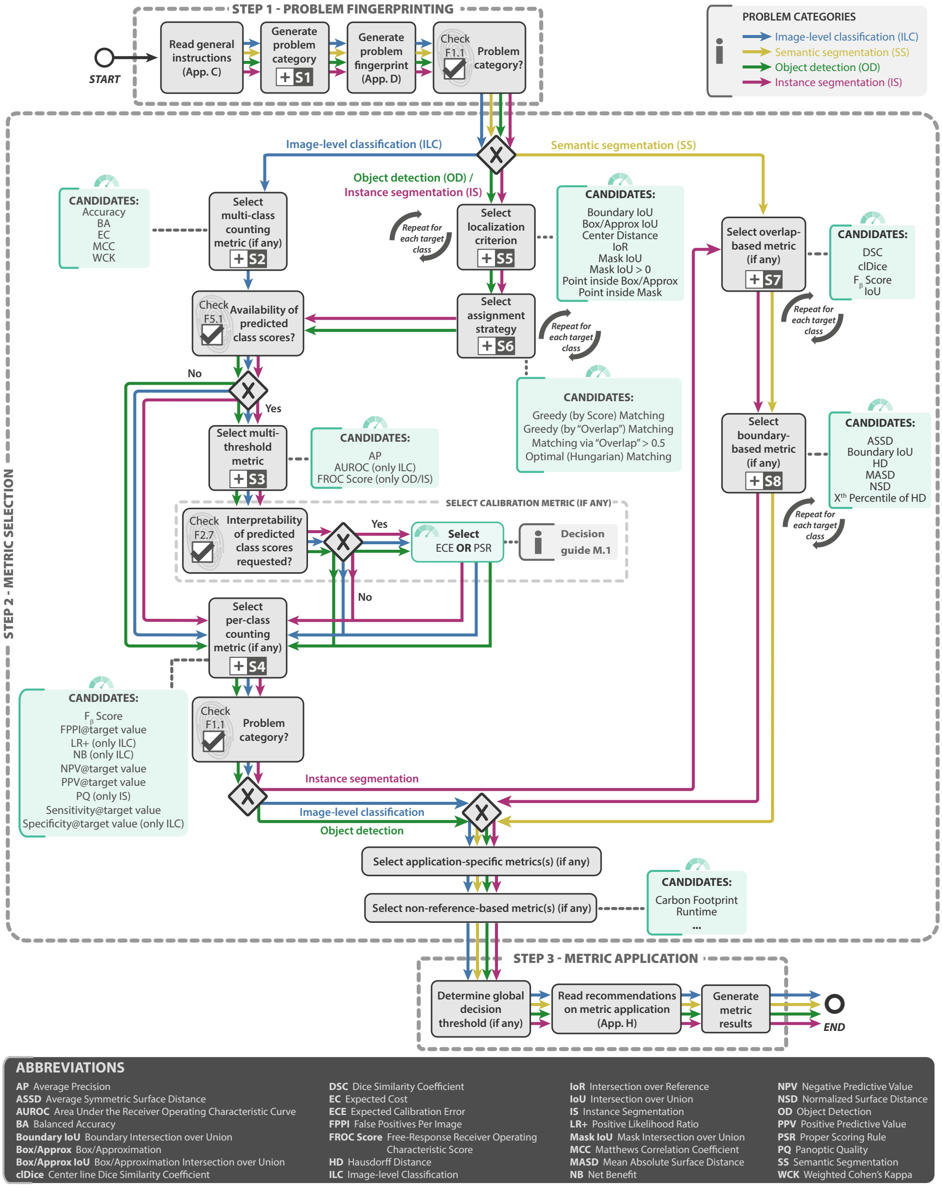
These problems are magnified by the fact that common practice often grows historically, and poor standards may be propagated from generation to generation of scientists. To dismantle such historically grown poor practices and leverage distributed knowledge from various subfields of image processing, we established the multidisciplinary *Metrics Reloaded*[2] consortium, comprising international experts from the fields of medical image analysis, biological image analysis, medical guideline development, general ML, statistics and epidemiology, and representing a large number of relevant biomedical imaging initiatives and societies.

The mission of Metrics Reloaded is to foster reliable algorithm validation through problem-aware choice of metrics with the long-term goal of (1) enabling the reliable tracking of scientific progress and (2) bridging the current chasm between AI research and translation into biomedical imaging practice.

Based on a kickoff workshop held in December 2020, the *Metrics Reloaded* framework (Fig. 1b and Fig. 2) was developed using a multi-stage Delphi process [15] for consensus building. Its primary purpose is to enable users to make educated decisions on which metrics to choose for a driving biomedical problem. The foundation of the metric selection process is the new concept of *problem fingerprinting* (Fig. 3). Abstracting from a specific domain, problem fingerprinting is the generation of a structured representation of the given biomedical problem that captures all properties relevant for the purpose of metric selection. As depicted in Fig. 3, the properties captured by the fingerprint comprise *domain interest-related* properties, such as the particular importance of structure boundary, volume or center, *target structure-related* properties, such as the shape complexity or the size of structures relative to the image grid size, *data set-related* properties, such as class imbalance, as well as *algorithm output-related* properties, such as the theoretical possibility of the algorithm output not containing any target structure. Based on the problem fingerprint, the user is then, in a transparent and understandable manner, guided through the process of selecting an appropriate set of metrics while being made aware of potential pitfalls related to the specific characteristics of the underlying biomedical problem. The *Metrics Reloaded* framework currently considers 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, as illustrated in Fig. 4. It thus considers problems that can be assigned to one of the following four *problem categories*: *image-level classification*, *semantic segmentation*, *object detection* or *instance segmentation*. Designed to be modality-independent, *Metrics Reloaded* can be suited for application in various image analysis domains even beyond the field of biomedicine.

In the following, we will present the key contributions of our work in detail, namely (1) the *Metrics Reloaded* framework for problem-aware metric selection along with the key findings and design decisions that guided its development (Fig. 2), (2) the application of the framework to common biomedical use cases, showcasing its broad applicability (selection shown in Fig. 5) and (3) the open online tool that has been implemented to improve the user experience with our framework.

[2] We thank former IMSY lab member Nina Sauter for the suggestion of the name, inspired by the *Matrix* film series.


STEP 1 - PROBLEM FINGERPRINTING
START
Read general instructions (App. C)
Generate problem category S1
Generate problem fingerprint (App. D)
Check F1.1 Problem category?
PROBLEM CATEGORIES
Image-level classification (ILC)
Semantic segmentation (SS)
Object detection (OD)
Instance segmentation (IS)
STEP 2 - METRIC SELECTION
Image-level classification (ILC)
Semantic segmentation (SS)
Object detection (OD) / Instance segmentation (IS)
CANDIDATES: Accuracy BA EC MCC WCK
Select multi-class counting metric (if any) S2
Repeat for each target class
Select localization criterion S5
CANDIDATES: Boundary IoU Box/Approx IoU Center Distance IoR Mask IoU Mask IoU > 0 Point inside Box/Approx Point inside Mask
Select overlap-based metric (if any) S7
CANDIDATES: DSC clDice F_β Score IoU
Repeat for each target class
Check F5.1 Availability of predicted class scores?
Select assignment strategy S6
Repeat for each target class
CANDIDATES: Greedy (by Score) Matching Greedy (by "Overlap") Matching Matching via "Overlap" > 0.5 Optimal (Hungarian) Matching
No
Yes
Select multi-threshold metric S3
CANDIDATES: AP AUROC (only ILC) FROC Score (only OD/IS)
Select boundary-based metric (if any) S8
CANDIDATES: ASSD Boundary IoU HD MASD NSD X^th Percentile of HD
Repeat for each target class
SELECT CALIBRATION METRIC (IF ANY)
Check F2.7 Interpretability of predicted class scores requested?
Yes
Select ECE OR PSR
Decision guide M.1
No
Select per-class counting metric (if any) S4
CANDIDATES: F_β Score FPPI@target value LR+ (only ILC) NB (only ILC) NPV@target value PPV@target value PQ (only IS) Sensitivity@target value Specificity@target value (only ILC)
Check F1.1 Problem category?
Instance segmentation
Image-level classification
Object detection
Select application-specific metrics(s) (if any)
Select non-reference-based metric(s) (if any)
CANDIDATES: Carbon Footprint Runtime ...
STEP 3 - METRIC APPLICATION
Determine global decision threshold (if any)
Read recommendations on metric application (App. H)
Generate metric results
END

ABBREVIATIONS

AP Average Precision
ASSD Average Symmetric Surface Distance
AUROC Area Under the Receiver Operating Characteristic Curve
BA Balanced Accuracy
Boundary IoU Boundary Intersection over Union
Box/Approx Box/Approximation
Box/Approx IoU Box/Approximation Intersection over Union
clDice Center line Dice Similarity Coefficient
DSC Dice Similarity Coefficient
EC Expected Cost
ECE Expected Calibration Error
FPPI False Positives Per Image
FROC Score Free-Response Receiver Operating Characteristic Score
HD Hausdorff Distance
ILC Image-level Classification
IoR Intersection over Reference
IoU Intersection over Union
IS Instance Segmentation
LR+ Positive Likelihood Ratio
Mask IoU Mask Intersection over Union
MCC Matthews Correlation Coefficient
MASD Mean Absolute Surface Distance
NB Net Benefit
NPV Negative Predictive Value
NSD Normalized Surface Distance
OD Object Detection
PPV Positive Predictive Value
PSR Proper Scoring Rule
PQ Panoptic Quality
SS Semantic Segmentation
WCK Weighted Cohen's Kappa

Fig. 2. ***Metrics Reloaded* recommendation framework from a user's perspective.** In **step 1 - problem fingerprinting**, the given biomedical image analysis problem is mapped to the appropriate image *problem category*, namely *image-level classification (ILC)*, *semantic segmentation (SS)*, *object detection (OD)*, or *instance segmentation (IS)* (Fig. 4). The problem category as well as further characteristics of the given biomedical problem that are relevant for metric selection are then captured in a *problem fingerprint* (Fig. 3). In **step 2 - metric selection**, the user follows the respective coloured path of the chosen problem category (ILC →, SS →, OD → or IS →) to select a suitable pool of metrics from the *Metrics Reloaded* pools shown in green. When a branching of the tree occurs, the fingerprint items determine which exact path to take. Finally, in **step 3 - metric application**, the user is supported in applying the metrics to a given data set. During the traversal of the decision tree, the user passes *subprocesses*, indicated by the ⊞-symbol, which are provided in dedicated figures in App. E and represent relevant steps in the metric selection process. Ambiguities related to metric selection are resolved via *decision guides* (App. F) that help users make an educated decision when multiple options are possible. An overview of the symbols used in the process diagram is provided in Fig. 6.

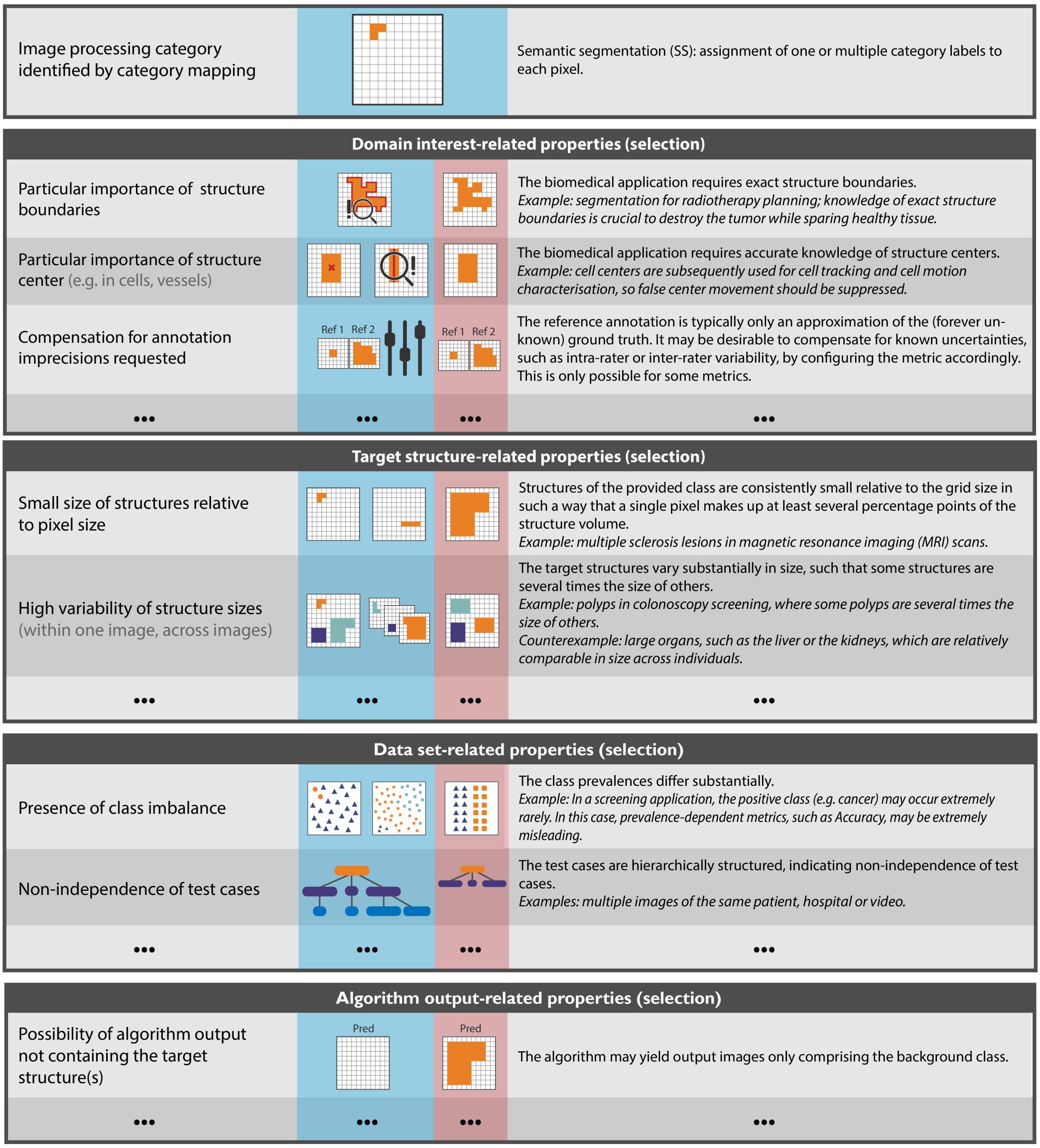


Image processing category identified by category mapping	Semantic segmentation (SS): assignment of one or multiple category labels to each pixel.

Domain interest-related properties (selection)

Particular importance of structure boundaries	The biomedical application requires exact structure boundaries. *Example: segmentation for radiotherapy planning; knowledge of exact structure boundaries is crucial to destroy the tumor while sparing healthy tissue.*
Particular importance of structure center (e.g. in cells, vessels)	The biomedical application requires accurate knowledge of structure centers. *Example: cell centers are subsequently used for cell tracking and cell motion characterisation, so false center movement should be suppressed.*
Compensation for annotation imprecisions requested	The reference annotation is typically only an approximation of the (forever unknown) ground truth. It may be desirable to compensate for known uncertainties, such as intra-rater or inter-rater variability, by configuring the metric accordingly. This is only possible for some metrics.
•••	•••

Target structure-related properties (selection)

Small size of structures relative to pixel size	Structures of the provided class are consistently small relative to the grid size in such a way that a single pixel makes up at least several percentage points of the structure volume. *Example: multiple sclerosis lesions in magnetic resonance imaging (MRI) scans.*
High variability of structure sizes (within one image, across images)	The target structures vary substantially in size, such that some structures are several times the size of others. *Example: polyps in colonoscopy screening, where some polyps are several times the size of others.* *Counterexample: large organs, such as the liver or the kidneys, which are relatively comparable in size across individuals.*
•••	•••

Data set-related properties (selection)

Presence of class imbalance	The class prevalences differ substantially. *Example: In a screening application, the positive class (e.g. cancer) may occur extremely rarely. In this case, prevalence-dependent metrics, such as Accuracy, may be extremely misleading.*
Non-independence of test cases	The test cases are hierarchically structured, indicating non-independence of test cases. *Examples: multiple images of the same patient, hospital or video.*
•••	•••

Algorithm output-related properties (selection)

Possibility of algorithm output not containing the target structure(s)	The algorithm may yield output images only comprising the background class.
•••	•••

Fig. 3. **Relevant properties of a driving biomedical image analysis problem are captured by the *problem fingerprint*** (selection for semantic segmentation shown here). The fingerprint comprises a set of items, each of which represents a specific property of the problem, is either binary or categorical and must be instantiated by the user. Besides the problem category, the fingerprint comprises *domain interest-related*, *target structure-related*, *data set-related* and *algorithm output-related* properties. A comprehensive version of the fingerprints for all problem categories can be found in Fig. 10 (image-level classification), Figs. 12/13 (semantic segmentation), Figs. 14-16 (object detection) and Figs. 17-19 (instance segmentation).

RESULTS

Metrics Reloaded is the result of a multi-stage Delphi process, comprising five international workshops, six questionnaires, numerous expert group meetings and social media-based crowdsourced feedback processes, all conducted over the past two years.

Image analysis pitfalls generalize across domains

As a foundation of the *Metrics Reloaded* recommendation framework, we identified common and rare pitfalls of metrics in the field of biomedical image analysis using a community-powered process, detailed in the sister publication of this work [72]. Notably, many pitfalls generalize not only across the four problem categories that our framework addresses but also across domains (Fig. 4). This is because the source of the pitfall, such as class imbalance, uncertainties in the reference or poor image resolution, can occur irrespective of a specific modality or application. *Following the convergence of AI methodology across domains and problem categories, we therefore argue for the analogous convergence of validation methodology.*

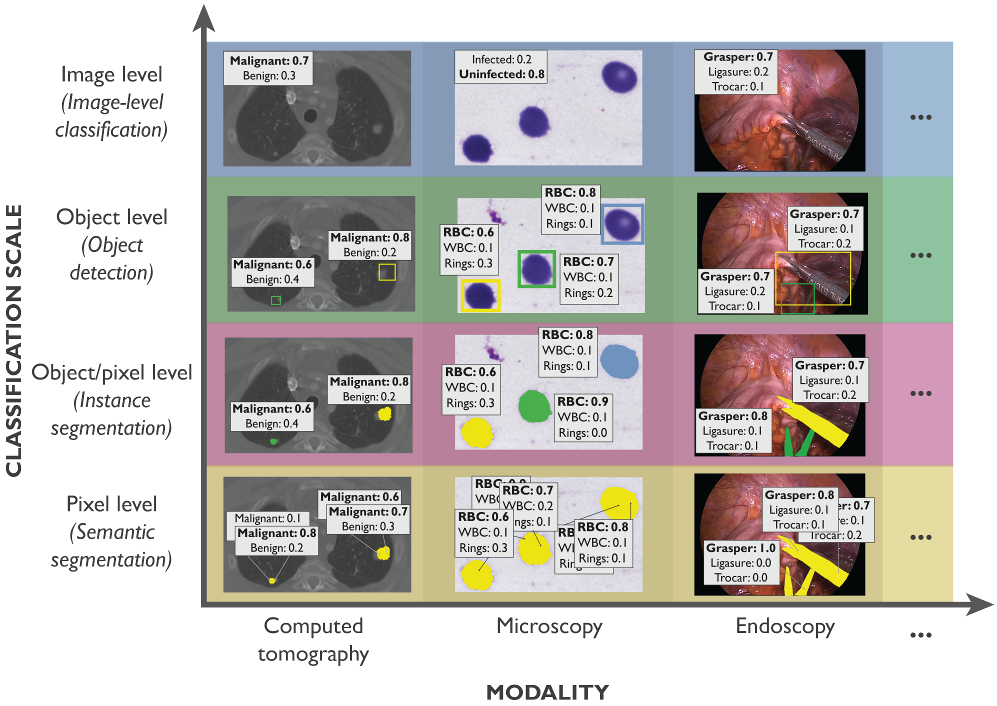

CLASSIFICATION SCALE
Image level (Image-level classification)
Object level (Object detection)
Object/pixel level (Instance segmentation)
Pixel level (Semantic segmentation)
Malignant: 0.7 Benign: 0.3
Infected: 0.2 Uninfected: 0.8
Grasper: 0.7 Ligasure: 0.2 Trocar: 0.1
Malignant: 0.8 Benign: 0.2
Malignant: 0.6 Benign: 0.4
RBC: 0.8 WBC: 0.1 Rings: 0.1
RBC: 0.6 WBC: 0.1 Rings: 0.3
RBC: 0.7 WBC: 0.1 Rings: 0.2
Grasper: 0.7 Ligasure: 0.1 Trocar: 0.2
Grasper: 0.7 Ligasure: 0.2 Trocar: 0.1
Malignant: 0.8 Benign: 0.2
Malignant: 0.6 Benign: 0.4
RBC: 0.8 WBC: 0.1 Rings: 0.1
RBC: 0.6 WBC: 0.1 Rings: 0.3
RBC: 0.9 WBC: 0.1 Rings: 0.0
Grasper: 0.7 Ligasure: 0.1 Trocar: 0.2
Grasper: 0.8 Ligasure: 0.1 Trocar: 0.1
Malignant: 0.6
Malignant: 0.7 Benign: 0.3
Malignant: 0.1
Malignant: 0.8 Benign: 0.2
RBC: 0.7 WBC: 0.2
RBC: 0.6 WBC: 0.1 Rings: 0.3
RBC: 0.8 WBC: 0.1 Rings: 0.1
Grasper: 0.8 Ligasure: 0.1 Trocar: 0.1
Trocar: 0.2
Grasper: 1.0 Ligasure: 0.0 Trocar: 0.0
...
Computed tomography
Microscopy
Endoscopy
MODALITY

Fig. 4. ***Metrics Reloaded* fosters the convergence of validation methodology across modalities, application domains and classification scales.** The framework considers problems in which categorical output is sought at image, object and/or pixel level, resulting (from top to bottom) in *image-level classification*, *object detection*, *instance segmentation* or *semantic segmentation* problems. These problem categories are relevant across modalities (here Computed Tomography (CT), microscopy and endoscopy) and application domains. From left to right: annotation of (left) benign and malignant lesions in CT images, (middle) different cell types in microscopy images, and (right) medical instruments in laparoscopic surgery images.

Historically grown practices are not always justified

To better understand common practice, we prospectively captured the designs of challenges organized by the IEEE Society of the International Symposium of Biomedical Imaging (ISBI), the Medical Image Computing and Computer Assisted Interventions (MICCAI) Society and the Medical Imaging with Deep Learning (MIDL) consortia. The organizers of the respective competitions were asked to provide a rationale for why they picked the chosen metrics for their competition. An analysis of a total of 138 competitions conducted between 2018 and 2022 revealed that metrics are frequently (in 24% of the competitions) based on common practice in the community. We found, however, that common practice is often not well-justified, and poor practices may even be propagated from one generation to the next.

One remarkable example is the widespread adaptation of an incorrect naming and inconsistent mathematical formulation of a metric proposed for cell instance segmentation. The term "Mean Average Precision (mAP)" usually refers to one of the most common metrics in object-level classification [57, 72]. Here, Precision denotes the Positive Predictive Value (PPV), which is "averaged" over varying thresholds on the predicted class scores of an object detection algorithm. The "mean" Average Precision (AP) is then obtained by taking the mean over classes [33, 72]. Despite the popularity of mAP, a widely known challenge on cell segmentation [48] introduced a new "Mean Average Precision" in 2018. Here, all the terms of the metric (mean, average and precision) refer to entirely different concepts. For instance, the common definition of Precision from literature $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ has been altered to $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$, where TP, FP, and FN refer to the cardinalities of the confusion matrix (i.e. the true/false positives/negatives). The latter formula actually denotes the Intersection over Union (IoU) metric. Despite these problems, the terminology has been adapted by subsequent influential work [49, 76, 82] indicating a wide-spread usage within the community.

To break such historically grown poor practices, we followed a multidisciplinary cross-domain approach that enabled us to critically question common practice in different communities and integrate distributed knowledge in one common framework.

Cross-domain approach enables integration of distributed knowledge

To leverage distributed knowledge on metric pitfalls, strengths and weaknesses, and thus arrive at a "best of all worlds" solution, we formed an international multidisciplinary consortium of 73 experts from various biomedical image analysis-related fields. Furthermore, we crowdsourced metric pitfalls and feedback on our approach in a social media campaign. This process led to a total of 126 researchers contributing to this work (including 53 mentioned in the acknowledgements). Consideration of the different knowledge and perspectives on metrics led to the following key design decisions for *Metrics Reloaded.*

Encapsulating domain knowledge: The questions asked to select a suitable metric are mostly similar regardless of image modality or application: Are the classes balanced? Is there a specific preference for the positive or negative class? What is the accuracy of the reference annotation? Is the structure boundary or volume of relevance for the target application?, and so on. Importantly, *while answering these questions requires domain expertise, the consequences in terms of metric selection can be regarded as domain-independent.* Our approach is thus to abstract from the specific image modality and domain of a given problem by capturing the properties relevant for metric selection in a *problem fingerprint* (Fig. 3).

Exploiting synergies across classification scales: Likewise, similar considerations must be made when choosing metrics for classification, detection and segmentation tasks. The reason is that they can all be regarded as classification tasks at different scales (Fig. 4). The similarities between the categories, however, can also lead to problems when the wrong category is chosen (see Fig. 1 (a) top left)). We therefore (i) address all four problem categories in one common framework (Fig. 2) and (ii) cover the selection of the problem category itself in our framework (Fig. 8).

Exploiting complementary metric strengths: A single metric is typically not able to cover the complex requirements of the driving biomedical problem [71]. To account for the complementary strengths and weakness of metrics, we generally recommend the usage of multiple complementary metrics to validate image analysis problems. A Delphi process yielded the pool of metrics shown in Tab. 2 that our *Metrics Reloaded* recommendations for common reference-based metrics are based on. Notably, while this includes broadly applied metrics, such as the F_1 Score, Area under the Receiver Operating Characteristic Curve (AUROC), or AP, it also led to the recommendation of metrics that have not received much attention to date. A prominent example is the Net Benefit (NB) [90] metric designed for determining whether basing decisions on a method would do more good than harm. A diagnostic test, for example, may lead to early identification and treatment of a disease, but typically the process will also cause some patients without disease being subjected to unnecessary further interventions. NB allows to consider such trade-offs by putting benefits and harms on the same scale so that they can be compared directly. Another example is the metric Expected Cost (EC) [56], which can be seen as a generalization of Accuracy with many desirable added features but is not well-known in the biomedical image analysis communities. As part of our crowdsourcing-based feedback on *Metrics Reloaded*, a researcher from the field of speech recognition, in which EC is applied very commonly, triggered the integration of the metric in our framework.

Abstracting from methodology: Metrics should be chosen based solely on the driving biomedical problem and not be affected by algorithm design choices. For example, the error functions applied in common neural network architectures do not justify the use of corresponding metrics (e.g. validating with DSC to match the Dice loss used for training a neural network). Instead, the domain interest should guide the choice of metric, which, in turn, can guide the choice of the loss term.

Involving and educating users: Choosing adequate validation metrics is a complex process. Rather than providing a black box recommendation, *Metrics Reloaded* therefore guides the user through the process of metric selection while raising awareness on pitfalls that may occur. In cases in which the tradeoffs between different choices must be considered, *decision guides* (App. F) guide the decision making process while respecting individual preferences.

Problem fingerprints encapsulate relevant domain knowledge

To encapsulate relevant domain knowledge in a common format and then enable a modality-agnostic metric recommendation approach that generalizes over domains, we developed the concept of *problem fingerprinting*, illustrated in Fig. 3. As a foundation, we crowdsourced all properties of a driving biomedical problem that are potentially relevant for metric selection. This process resulted in a list of binary and categorical variables (*fingerprint items*) that must be instantiated by a user to trigger the *Metrics Reloaded* recommendation process. Common issues often relate to selecting metrics from the wrong problem category, as illustrated in Fig. 1 (a, top left). To avoid such issues, the problem fingerprinting begins with the step of mapping a given problem with all its intrinsic and data set-related properties to the corresponding problem category via the *category mapping* shown in Fig. 8. The problem category is a fingerprint item itself.

In the following, we will refer to all fingerprint items with the notation *FX.Y*, where Y is a numerical identifier and the index *X* represents one of the following families:

1 - **General properties** solely refer to the problem category encoded in *F1.1: Problem category.*

2 - **Domain interest-related properties** reflect user preferences and are highly dependent on the target application. A semantic image segmentation that serves as the foundation for radiotherapy planning, for example, would require exact contours (*F2.1 Particular importance of structure boundaries* = TRUE). On the other hand, for a cell segmentation problem that serves as prerequisite for cell tracking, the object centers may be be much more important (*F2.1 Particular importance of structure center* = TRUE). Both problems could be tackled with identical network architectures, but the validation metrics should be different.

3 - **Target structure-related properties** represent inherent properties of target structure(s) (if any), such as the size, size variability and the shape.

4 - **Data set-related properties** capture properties that are inherent to the provided training/test data. They relate primarily to class prevalences, uncertainties of the reference annotations as well as a (potentially) hierarchical data structure.

5 - **Algorithm output-related properties** encode properties of the output, such as the availability of predicted class scores.

Note that not all properties are relevant for all problem categories. For example, the shape and size of target structures is highly relevant for segmentation problems but irrelevant for image classification problems. The complete problem category-specific fingerprints are provided in Figs. 10/11 (image-level classification), Figs. 12/13 (semantic segmentation), Figs. 14-16 (object detection) and Figs. 17-19 (instance segmentation), respectively. Fingerprints are referred to in various parts of the manuscript including the decision trees (Fig. 2 and App. E), the decision guides (App. F) and in the cross-category recommendations (Tab. 1).

Metrics Reloaded addresses all three types of metric pitfalls

Metrics Reloaded was designed to address all three types of metric pitfalls identified in [72] and illustrated in Fig. 1a. More specifically, each of the three steps shown in Fig. 2 addresses one type of pitfall:

Step 1 - Fingerprinting. A user should begin by reading the general instructions of the recommendation framework, provided in App. C. Next, the driving biomedical problem should be converted to a problem fingerprint, as detailed in the previous paragraph. This step is not only a prerequisite for applying the framework across application domains and classification scales, but it also specifically addresses the *inappropriate phrasing of the problem* via the integrated category mapping. Once the user's domain knowledge has been encapsulated in the problem fingerprint, the actual metric selection is conducted according to a domain- and modality-agnostic process.

Step 2 - Metric Selection. The metric recommendation is then performed with a Business Process Model and Notation (BPMN)-inspired flowchart (see Fig. 6), in which conditional operations are based on one or multiple fingerprint properties (Fig. 2). It can be subdivided into three substeps, each addressing the complementary strengths and weaknesses of common metrics. First, common *reference-based metrics*, which are based on the comparison of the algorithm output to a reference annotation, are selected. Next, the pool of standard metrics can be complemented with custom metrics to address application-specific complementary properties. Finally, non-reference-based metrics that assess speed, memory consumption or carbon footprint, for example, can be added to the metric pool(s).

Depending on the problem category, different paths through the mapping are taken; however, their huge overlap demonstrates substantial synergies. All paths comprise several subprocesses S (indicated by the ⊞-symbol), each of which is a placeholder for a decision tree representing one specific step of the selection process. Traversal of a subprocess typically leads to the addition of a metric to the metric pool. In multi-class prediction problems, dedicated metric pools for each class must be generated as relevant properties may differ from class to class. A 3D semantic segmentation problem, for example, could require the simultaneous segmentation of both tubular and non-tubular structures (e.g. liver and its vessels). These require different metrics for validation. In ambiguous cases, i.e. when the user can choose between two options in one step of the decision tree, a corresponding *decision guide* details the tradeoffs that need to be considered (App. F). For example, the IoU and the DSC are mathematically closely related. The concrete choice typically boils down to a simple user/community preference.

The following paragraphs present a summary of the four different colored paths through Step 2 - Metric Selection of the recommendation tree (Fig. 2), with a focus on reference-based metrics.

Image-level Classification. Image-level classification is conceptionally the most straightforward problem category, as the task is simply to assign one or multiple labels to the entire image. The validation metrics are designed to measure two key properties: *discrimination* and *calibration*. *Discrimination* refers to the ability of a classifier to discriminate two or more classes from each other. This can be achieved with *counting metrics* that operate on the cardinalities of a fixed confusion matrix (i.e. the true/false positives/negatives in the binary case). Prominent examples are Sensitivity, Specificity or F_1 Score for binary settings and Matthews Correlation Coefficient (MCC) for multi-class settings. Setting a (potentially arbitrary) cutoff on the predicted class scores to obtain a confusion matrix can be regarded as problematic [72]. *Multi-threshold metrics*, such as AUROC, are therefore based on varying the cutoff, which enables the explicit analysis of the tradeoff between competing properties such as Sensitivity and Specificity. This, in turn, can lead to an in-depth understanding of the inherent capabilities of a method. Another complementary important property of classification algorithms is the *calibration* performance. A method is well-calibrated if the predicted class scores match the probability of class membership. Overconfident or underconfident classifiers can be especially problematic in prediction tasks where a clinical decision may be made based on the risk of the patient of developing a certain condition. Based on these considerations *Metrics Reloaded* provides recommendations for both discrimination and calibration capabilities of algorithms. We recommend the following process for classification problems (blue path in Fig. 2):

1: Select multi-class metric (if any): Multi-class metrics have the unique advantage that they capture the performance of an algorithm for all classes in a single value. With the ability of taking into account all entries of the multi-class confusion matrix, they provide a holistic measure of performance without the need for customized class-aggregation schemes. We recommend using a multi-class metric if a cutoff on predicted class scores is requested. The concrete choice of metric depends on the distribution of classes and whether there is an unequal interest in class confusions. Details can be found in subprocess S2 for selecting multi-class metrics (Fig. 20).

2: Select multi-threshold metric (if any): As detailed class-specific analyses are not possible with multi-class metrics, we recommend an additional per-class validation based on a multi-threshold metric. Each class of interest is separately assessed, preferably in a "One-vs-the-Rest" fashion (see App. H). While we recommend AUROC as the default multi-threshold metric for classification, AP can be suitable when class imbalances should be compensated. Details can be found in subprocess S3 for selecting multi-threshold metrics (Fig. 21).

3: Select calibration metric (if any): If predicted class probabilities are provided and should be human-interpretable, an additional metric that measures the calibration capabilities of the method should be added to the metric pool. Depending on user preferences, we recommend the Expected Calibration Error (ECE) or Proper Scoring Rules (PSR), as detailed in the respective decision guide (App. F DM.1).

4: Select per-class counting metric (if any): If a cutoff on the predicted class score is desired, we further recommend the selection of a per-class counting metric. The choice depends primarily on the cutoff strategy (see App. D) and the distribution of classes. Details can be found in subprocess S4 for selecting per-class counting metrics (Fig. 22).

Semantic segmentation. In semantic segmentation, classification occurs at pixel level. However, it is not advisable to simply apply the standard classification metrics to the entire collection of pixels in a data set for two reasons. Firstly, pixels of the same image are highly correlated. Hence, to respect the hierarchical data structure, metric values should first be computed per image and then be aggregated over the set of images. Note in this context that the commonly used DSC is mathematically identical to the popular F_1 Score applied at pixel level. Secondly, in segmentation problems, the user typically has an inherent interest in structure boundaries, shapes or volumes of structures (F2.1, F2.2, F2.3). The family of *boundary-based metrics* (subset of *distance-based metrics*) therefore requires the extraction of structure boundaries from the binary segmentation masks as a foundation for segmentation assessment. Based on these considerations and given all the complementary strengths and weaknesses of common segmentation metrics [72], we recommend the following process for segmentation problems (orange path in Fig. 2):

1: Select overlap-based metric (if any): In segmentation problems, counting metrics such as the DSC or IoU measure the overlap between the reference annotation and the prediction of the algorithm. As they can be considered the de facto standard for assessing segmentation quality and are well-interpretable, we recommend using them by default unless the target structures are consistently small (relative to the grid size) *and* the reference may be noisy. Depending on the specific properties of the problems, we recommend the DSC or IoU (default recommendation), the F_β Score (preferred when there is a preference for either FP or FN) or the Center line Dice Similarity Coefficient (clDice) (for tubular structures). Details can be found in subprocess S7 for selecting overlap-based metrics (Fig. 25).

2: Select boundary-based metric (if any): Key weaknesses of overlap-based metrics include shape unawareness and limitations when dealing with small structures or high size variability [72]. Our general recommendation is therefore to complement an overlap-based metric with a boundary-based metric. Depending on the specific properties of the driving problem, we recommend the Normalized Surface Distance (NSD) or Boundary IoU (good for addressing inter-rater variability), the Average Symmetric Surface Distance (ASSD) or Mean Average Surface Distance (MASD) (good for distance-based penalization of outliers with contour focus), or the Hausdorff Distance (HD) or its variants (good for distance-based penalization of outliers with outlier focus), as detailed in subprocess S8 for selecting boundary-based metrics (Fig. 26). Note that distance-based metrics are not appropriate under several circumstances.

In scenarios in which multiple structures of the same type can be seen within the same image (e.g. in Multiple Sclerosis (MS) lesion segmentation), for example, a potential pitfall is related to comparing a given structure boundary to the boundary of the wrong instance in the reference (Fig. 9). Similar issues arise in the case of completely missed instances. In such scenarios, we recommend reconsideration to phrase the problem as an instance segmentation problem. If semantic segmentation remains the chosen category, we advise against the use of distance-based metrics, as these are not designed for cases where mix-up of boundaries across different instances can occur.

When complementing these common metrics by further application-specific metrics, various properties may be taken into account. To address the particular importance of structure volume (F2.2), for example, volume-based metrics [69, 69, 91] may be added to the pool. Similarly, the compliance with prior knowledge, such as hierarchical label structure (F3.4), can be measured with further application-specific metrics.

Object detection. Object detection problems differ from segmentation problems in several key features with respect to metric selection. Firstly, they are capable of distinguishing different instances of the same class and thus require the step of locating objects and assigning them to the corresponding reference object. Secondly, the granularity of localization is comparatively rough, which is why no boundary-based metrics are required (otherwise the problem would be phrased as an instance segmentation problem). Finally, and crucially important from a mathematical perspective is the absence of True Negative (TN) in object detection problems, which renders many popular classification metrics (e.g. Accuracy, Specificity, AUROC) invalid. In fact, suitable counting metrics can only be based on three of the four entries of the confusion matrix. Based on these considerations and taking into account all the complementary strengths and weaknesses of existing metrics [72], we propose the following steps for object detection problems (green path in Fig. 2):

1: Select localization criterion: An essential part of the validation is to decide whether a prediction matches a reference object. To this end, (1) the location of both the reference objects and the predicted objects must be adequately represented (e.g. by masks, bounding boxes or center points) and (2) a metric for deciding on a match (e.g. Mask IoU) must be chosen. As detailed in subprocess S5 for selecting the localization criterion (Fig. 23), our recommendation considers both the granularity of the provided reference and the required granularity of the localization.

2: Select assignment strategy: As the localization does not necessarily lead to unambiguous matchings, an assignment strategy needs to be chosen to potentially resolve ambiguities that occurred during localization. As detailed in subprocess S6 for selecting the assignment strategy (Fig. 24), the recommended assignment strategy depends on the availability of continuous class scores, the possibility of overlapping predictions as well as on whether double assignments should be punished.

3: Select classification metric(s) (if any): Once objects have been located and assigned to reference objects, generation of a confusion matrix (without TN) is possible. The final step therefore simply comprises choosing suitable classification metrics for validation. Several subfields of biomedical image analysis have converged to choosing solely a counting metric, such as the F_β Score, as primary metric. We follow this recommendation when no continuous class scores are available for the detected objects. Otherwise, we disagree with the practice of basing performance assessment solely on a single (potentially suboptimal) cutoff on the continuous class scores. Instead, we propose selecting a multi-threshold metric (subprocess

S3, Fig. 21) and potentially complementing it with a counting metric (subprocess S4, Fig. 22) to present a more holistic picture of performance. As multi-threshold metric, we recommend AP or Free-Response Receiver Operating Characteristic (FROC) Score, depending on whether an easy interpretation (FROC Score) or a standardized metric (AP) is preferred. The choice of the per-class counting metric depends primarily on the cutoff strategy (F2.6)

Note that the previous description implicitly assumed single-class problems, but generalization to multi-class problems is straightforward by applying the validation per class. It is further worth mentioning that metric *application* is not straightforward in object detection problems as the number of objects in an image may be extremely small (even zero) compared to the number of pixels in an image. Special considerations with respect to aggregation must therefore be made, as detailed in App. H.

Instance segmentation. Instance segmentation delivers the tasks of object detection and semantic segmentation at the same time. Thus, the pitfalls and recommendations for instance segmentation problems are closely related to those for segmentation and object detection [72]. This is directly reflected in our metric selection process (purple path in Fig. 2):

1: Select object detection metric(s): To overcome problems related to instance unawareness (Fig. 1a (top)), we recommend selection of a set of detection metrics to explicitly measure detection performance. To this end, we recommend almost the exact process as for object detection with two exceptions. Firstly, given the fine granularity of both the output and the reference annotation, our recommendation for the localization strategy differs, as detailed in subprocess S5 (Fig. 23). Secondly, as depicted in S4, Fig. 22, we recommend the Panoptic Quality (PQ) [51] as an alternative to the F_β Score. This metric is especially suited for instance segmentation, as it combines the assessment of overall detection performance and segmentation quality of successfully matched (True Positive (TP)) instances in a single score.

2: Select segmentation metric(s) (if any): In a second step, metrics to explicitly assess the segmentation quality for the TP instances may be selected. Here, we follow the exact same process as in semantic segmentation (subprocesses S7, Fig. 25 and S8, Fig. 26). The primary difference is that the segmentation metrics are applied per instance.

While we have found our generic recommendation for instance segmentation to match the majority of biomedical problems, the standard reference-based metrics are not well-suited for some applications. Specifically, reference-based metrics struggle in images with structures of extreme density and complex shapes, where overlap fails as a criterion for one-to-one correspondences between predictions and reference. In such cases, specialized metrics targeted to the specific problem may be required.

Step 3 - Metric Application. Once a suitable metric pool has been generated, the chosen metrics must be applied to the given data set. While this appears straightforward, numerous pitfalls can occur [72]. We recommend beginning with the setting of the global decision threshold in case metrics based on a fixed cutoff on the predicted class scores (F2.6) have been selected, which is generally the case. In order to avoid overestimation of algorithm performance, this threshold needs to be set globally for all classes and metrics, as detailed in App. C. Once raw metric values have been computed for all metrics, metric values are aggregated, potentially combined (for rankings) and reported. Cross-category recommendations for this process are summarized in Tab. 1. As our recommendations with respect to aggregation go beyond the current state of the art, we provide a dedicated section on this topic (App. H)

Table 1. **Cross-category recommendations on metric application.**

Recommendation	**Further Information**
Recommendations on metric aggregation	
Perform a per-class validation	App. H
Select a strategy for missing data handling if *F5.3 Possibility of invalid algorithm output* holds	App. H
Respect the hierarchical data structure when aggregating metric values if *F4.5 non-independence of test data holds*	App. H
Stratify by size if *F3.2 high variability of structures sizes* holds (only OD and IS)	App. H
Perform weighted class aggregation if *F2.5.1 unequal interest across classes* holds	App. H
Address the potential correlation between classes when aggregating, especially when *F3.4 Possibility of multiple labels per unit* holds	App. H
Leverage metadata (if any) to reveal potential algorithmic bias	Discussion
Recommendations on metric reporting	
Include a visualization of the raw metric values	[93]
Choose the number of decimal places such that they reflect both relevance and the uncertainties of the reference	
Report on the quality of the reference (e.g. intra-rater and inter-rater variability)	[55]
Consider multiple test set runs to address the variability of results resulting from non-determinism	[50] and [83]
When combining metrics for algorithm ranking, be aware of their relation	[72, 81, 84]
If a ranking of competing algorithms is desired, choose a set of primary metrics (PM) that best reflect the underlying biomedical goal. Complement the PM by secondary metrics (SM) such that all relevant performance measures are covered by PM and SM.	App. C
Make uncertainties in rankings (if any) explicit	[93]
Depending on the specific application, apply existing reporting guidelines, such as BIAS [59] (for challenges), TRIPOD [67] (for validation of prediction models, including recommendations that are applicable to the validation of medical image analysis algorithms—TRIPOD-AI guidelines currently in preparation [29]) and CLAIM [66] (checklist for AI algorithms in medical imaging).	[1]

Metrics Reloaded is broadly applicable in biomedical image analysis

To validate the *Metrics Reloaded* framework, we used it to generate recommendations for common use cases in biomedical image processing (see App. I). The traversal through the decision tree of our framework is detailed for eight selected use cases corresponding to the four different problem categories (Fig. 5):

Image-level classification: frame-based sperm motility classification based on microscopy time-lapse video containing human spermatozoa and disease classification in dermoscopic images (Fig. 42).

Semantic segmentation: lung cancer segmentation in microscopy images and liver segmentation in CT images (Fig. 44).

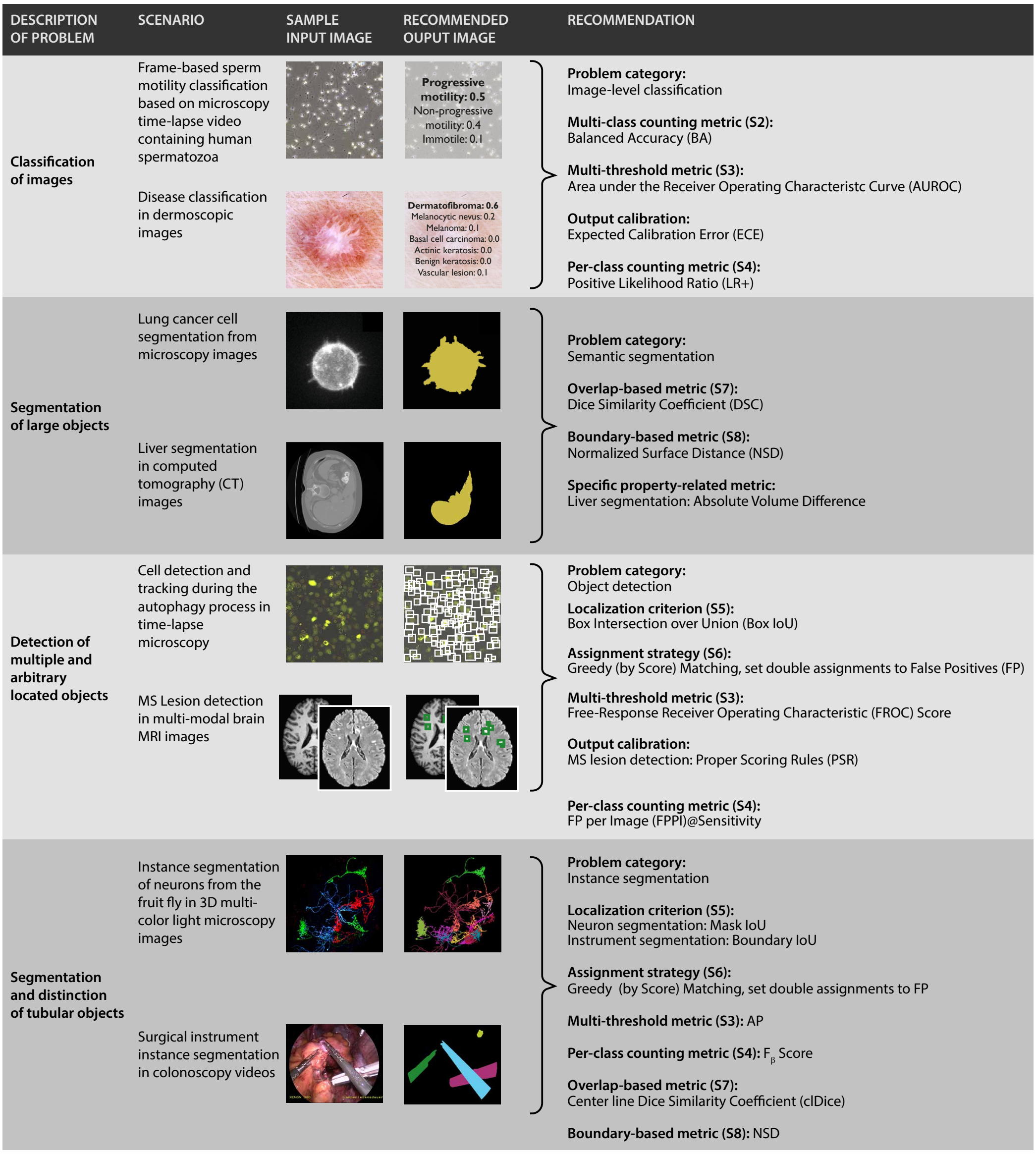
Object detection: cell detection and tracking during the autophagy process in time-lapse microscopy and MS lesion detection in multi-modal brain Magnetic Resonance Imaging (MRI) images (Fig. 46).

Instance segmentation: instance segmentation of neurons from the fruit fly in 3D multi-color light microscopy images and surgical instrument instance segmentation in colonoscopy videos (Fig. 48).

The resulting metric recommendations shown in Fig. 5 demonstrate that a common framework across subfields is sensible. In the showcased examples, shared properties of problems from different domains result in almost identical recommendations. In the semantic segmentation use cases, for example, the specific image modality is irrelevant for metric selection. What matters is the fact that a single object with a large size relative to the grid size should be segmented - properties that are captured by the proposed fingerprint. In App. I, we present recommendations for several other biomedical use cases.

The *Metrics Reloaded* online tool improves user experience

Selecting appropriate validation metrics while considering all potential pitfalls that may occur is a highly complex process, as demonstrated by the large number of figures in this paper. Some of the complexity, however, also results from the fact that the figures needed to capture all eventualities at once. For example, many of the figures could be simplified substantially for problems based on only two classes. To leverage this potential and to improve the user experience with our framework in general, we have therefore developed the *Metrics Reloaded* online tool, which captures all peculiarities of our framework in a user-centric manner and will soon be publicly available in a web-based version. The online tool can serve as a trustworthy common access point for image analysis validation.


Progressive motility: 0.5
Non-progressive motility: 0.4
Immotile: 0.1
Dermatofibroma: 0.6
Melanocytic nevus: 0.2
Melanoma: 0.1
Basal cell carcinoma: 0.0
Actinic keratosis: 0.0
Benign keratosis: 0.0
Vascular lesion: 0.1

DESCRIPTION OF PROBLEM	SCENARIO	SAMPLE INPUT IMAGE	RECOMMENDED OUPUT IMAGE	RECOMMENDATION
Classification of images	Frame-based sperm motility classification based on microscopy time-lapse video containing human spermatozoa Disease classification in dermoscopic images			**Problem category:** Image-level classification **Multi-class counting metric (S2):** Balanced Accuracy (BA) **Multi-threshold metric (S3):** Area under the Receiver Operating Characteristc Curve (AUROC) **Output calibration:** Expected Calibration Error (ECE) **Per-class counting metric (S4):** Positive Likelihood Ratio (LR+)
Segmentation of large objects	Lung cancer cell segmentation from microscopy images Liver segmentation in computed tomography (CT) images			**Problem category:** Semantic segmentation **Overlap-based metric (S7):** Dice Similarity Coefficient (DSC) **Boundary-based metric (S8):** Normalized Surface Distance (NSD) **Specific property-related metric:** Liver segmentation: Absolute Volume Difference
Detection of multiple and arbitrary located objects	Cell detection and tracking during the autophagy process in time-lapse microscopy MS Lesion detection in multi-modal brain MRI images			**Problem category:** Object detection **Localization criterion (S5):** Box Intersection over Union (Box IoU) **Assignment strategy (S6):** Greedy (by Score) Matching, set double assignments to False Positives (FP) **Multi-threshold metric (S3):** Free-Response Receiver Operating Characteristic (FROC) Score **Output calibration:** MS lesion detection: Proper Scoring Rules (PSR) **Per-class counting metric (S4):** FP per Image (FPPI)@Sensitivity
Segmentation and distinction of tubular objects	Instance segmentation of neurons from the fruit fly in 3D multi-color light microscopy images Surgical instrument instance segmentation in colonoscopy videos			**Problem category:** Instance segmentation **Localization criterion (S5):** Neuron segmentation: Mask IoU Instrument segmentation: Boundary IoU **Assignment strategy (S6):** Greedy (by Score) Matching, set double assignments to FP **Multi-threshold metric (S3):** AP **Per-class counting metric (S4):** F_β Score **Overlap-based metric (S7):** Center line Dice Similarity Coefficient (clDice) **Boundary-based metric (S8):** NSD

Fig. 5. **Instantiation of the framework with recommendations for concrete biomedical questions.** From top to bottom: **(1)** Image classification for the examples of sperm motility classification [40] and disease classification in dermoscopic images [27]. **(2)** Semantic segmentation of large objects for the examples of lung cancer cell segmentation from microscopy [20] and liver segmentation in Computed Tomography (CT) images [2, 80]. **(3)** Detection of multiple and arbitrarily located objects for the examples of cell detection and tracking during the autophagy process [68, 96] and MS lesion detection in multi-modal brain MRI images [30, 53]. **(4)** Instance segmentation of tubular objects for the examples of instance segmentation of neurons from the fruit fly [61, 65, 86] and surgical instrument instance segmentation [60]. The corresponding traversals through the decision trees are shown in App. I.

CURRENT STATUS

The *Metrics Reloaded* consortium is currently finalizing the paper. We are still integrating some of the valuable community feedback that we have received via the questionnaire issued as part of the previous version of this paper. The future version will comprise:

(1) An elaboration on the cross-category recommendations summarized in Tab. 1
(2) A comprehensive Methods section detailing the Delphi process and other aspects of our work
(3) A comprehensive discussion of *Metrics Reloaded* including open research questions and future work
(4) An introduction of the *Metrics Reloaded* online tool (including a link to the tool)

In parallel, we are working on the next version of the corresponding pitfalls paper [72], which will be complemented by further figures, resulting - among others - from community contributions.

REFERENCES

[1] Douglas G Altman, Iveta Simera, John Hoey, David Moher, and Ken Schulz. 2008. EQUATOR: reporting guidelines for health research. *Open Medicine* 2, 2 (2008), e49.

[2] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. 2021. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735* (2021).

[3] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. 2019. Bach: Grand challenge on breast cancer histology images. *Medical image analysis* 56 (2019), 122–139.

[4] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics* 38, 2 (2011), 915–931.

[5] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, et al. 2015. Data From LIDC-IDRI [Data set]. *The Cancer Imaging Archive* (2015).

[6] John Attia. 2003. Moving beyond sensitivity and specificity: using likelihood ratios to help interpret diagnostic tests. *Australian prescriber* 26, 5 (2003), 111–113.

[7] Marc Aubreville, Nikolas Stathonikos, Christof A Bertram, Robert Klopleisch, Natalie ter Hoeve, Francesco Ciompi, Frauke Wilm, Christian Marzahl, Taryn A Donovan, Andreas Maier, et al. 2022. Mitosis domain generalization in histopathology images–The MIDOG challenge. *arXiv preprint arXiv:2204.03742* (2022).

[8] Andriy I Bandos, Howard E Rockette, Tao Song, and David Gur. 2009. Area under the free-response ROC curve (FROC) and a related summary index. *Biometrics* 65, 1 (2009), 247–256.

[9] Christian F Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P Fletcher, Sandra Smith, Lisa M Koch, Bernhard Kainz, and Daniel Rueckert. 2017. SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE transactions on medical imaging* 36, 11 (2017), 2204–2215.

[10] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama* 318, 22 (2017), 2199–2210.

[11] Miroslav Beneš and Barbara Zitová. 2015. Performance evaluation of image segmentation algorithms on microscopic image data. *Journal of microscopy* 257, 1 (2015), 65–85.

[12] Jorge Bernal, Aymeric Histace, Marc Masana, Quentin Angermann, Cristina Sánchez-Montes, Cristina Rodríguez de Miguel, Maroua Hammami, Ana García-Rodríguez, Henry Córdova, Olivier Romain, et al. 2019. GTCreator: a flexible annotation tool for image-based datasets. *International journal of computer assisted radiology and surgery* 14, 2 (2019), 191–201.

[13] Sebastian Bickelhaupt, Paul Ferdinand Jaeger, Frederik Bernd Laun, Wolfgang Lederer, Heidi Daniel, Tristan Anselm Kuder, Lorenz Wuesthof, Daniel Paech, David Bonekamp, Alexander Radbruch, et al. 2018. Radiomics based on adapted diffusion kurtosis imaging helps to clarify most mammographic findings suspicious for cancer. *Radiology* 287, 3 (2018), 761–770.

[14] Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*. Vol. 4. Springer.

[15] Bernice B Brown. 1968. *Delphi process: a methodology used for the elicitation of opinions of experts.* Technical Report. Rand Corp Santa Monica CA.

[16] Samuel Budd, Prachi Patkee, Ana Baburamani, Mary Rutherford, Emma C Robinson, and Bernhard Kainz. 2020. Surface agnostic metrics for cortical volume segmentation and regression. In *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-Oncology*. Springer, 3–12.

[17] Juan C Caicedo, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, et al. 2019. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nature methods* 16, 12 (2019), 1247–1253.

[18] Juan C Caicedo, Jonathan Roth, Allen Goodman, Tim Becker, Kyle W Karhohs, Matthieu Broisin, Csaba Molnar, Claire McQuin, Shantanu Singh, Fabian J Theis, et al. 2019. Evaluation of deep learning strategies for nucleus segmentation in fluorescence images. *Cytometry Part A* 95, 9 (2019), 952–965.

[19] Aaron Carass, Snehashis Roy, Adrian Gherman, Jacob C Reinhold, Andrew Jesson, Tal Arbel, Oskar Maier, Heinz Handels, Mohsen Ghafoorian, Bram Platel, et al. 2020. Evaluating white matter lesion segmentations with refined Sørensen-Dice analysis. *Scientific reports* 10, 1 (2020), 1–19.

[20] Carlos Castilla, Martin Maška, Dmitry V Sorokin, Erik Meijering, and Carlos Ortiz-de Solórzano. 2018. 3-D quantification of filopodia in motile cancer cells. *IEEE transactions on medical imaging* 38, 3 (2018), 862–872.

[21] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. 2021. Boundary IoU: Improving Object-Centric Image Segmentation Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15334–15342.

[22] Nicolas Chenouard, Ihor Smal, Fabrice De Chaumont, Martin Maška, Ivo F Sbalzarini, Yuanhao Gong, Janick Cardinale, Craig Carthel, Stefano Coraluppi, Mark Winter, et al. 2014. Objective comparison of particle tracking methods. *Nature methods* 11, 3 (2014), 281–289.

[23] Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. 2021. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining* 14, 1 (2021), 1–22.

[24] Nancy Chinchor. 1992. MUC-4 Evaluation Metrics. In *Proceedings of the 4th Conference on Message Understanding* (McLean, Virginia) *(MUC4 '92)*. Association for Computational Linguistics, USA, 22–29. https://doi.org/10.3115/1072064.1072067

[25] Neil T Clancy, Geoffrey Jones, Lena Maier-Hein, Daniel S Elson, and Danail Stoyanov. 2020. Surgical spectral imaging. *Medical image analysis* 63 (2020), 101699.

[26] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. 2013. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging* 26, 6 (2013), 1045–1057.

[27] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368* (2019).

[28] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[29] Gary S Collins and Karel GM Moons. 2019. Reporting of artificial intelligence prediction models. *The Lancet* 393, 10181 (2019), 1577–1579.

[30] Olivier Commowick, Audrey Istace, Michael Kain, Baptiste Laurent, Florent Leray, Mathieu Simon, Sorina Camarasu Pop, Pascal Girard, Roxana Ameli, Jean-Christophe Ferré, et al. 2018. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific reports* 8, 1 (2018), 1–17.

[31] Paulo Correia and Fernando Pereira. 2006. Video object relevance metrics for overall segmentation quality evaluation. *EURASIP Journal on Advances in Signal Processing* 2006 (2006), 1–11.

[32] Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology* 26, 3 (1945), 297–302.

[33] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88, 2 (2010), 303–338.

[34] Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association* 102, 477 (2007), 359–378.

[35] Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Statist. Assoc.* 102, 477 (March 2007), 359–378. https://doi.org/10.1198/016214506000001437

[36] Mark J Gooding, Annamarie J Smith, Maira Tariq, Paul Aljabar, Devis Peressutti, Judith van der Stoep, Bart Reymen, Daisy Emans, Djoya Hattu, Judith van Loon, et al. 2018. Comparative evaluation of autocontouring in clinical practice: a practical method using the Turing test. *Medical physics* 45, 11 (2018), 5105–5115.

[37] Margherita Grandini, Enrico Bagli, and Giorgio Visani. 2020. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756* (2020).

[38] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On Calibration of Modern Neural Networks. *ICML* (2017), 10.

[39] James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 1 (1982), 29–36.

[40] Trine B Haugen, Steven A Hicks, Jorunn M Andersen, Oliwia Witczak, Hugo L Hammer, Rune Borgli, Pål Halvorsen, and Michael Riegler. 2019. Visem: A multimodal video dataset of human spermatozoa. In *Proceedings of the 10th ACM Multimedia Systems Conference*. 261–266.

[41] Katrin Honauer, Lena Maier-Hein, and Daniel Kondermann. 2015. The hci stereo metrics: Geometry-aware performance analysis of stereo algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*. 2120–2128.

[42] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. 1993. Comparing images using the Hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence* 15, 9 (1993), 850–863.

[43] BSEN ISO 9000. 2000. Quality management systems: Fundamentals and vocabulary. *London: British Standards Institution* (2000).

[44] Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist* 11, 2 (1912), 37–50.

[45] Paul Ferdinand Jäger. 2020. Challenges and Opportunities of End-to-End Learning in Medical Image Classification. (2020).

[46] Pierre Jannin, Christophe Grova, and Calvin R Maurer. 2006. Model for defining and reporting reference-based validation protocols in medical image processing. *International Journal of Computer Assisted Radiology and Surgery* 1, 2 (2006), 63–73.

[47] Liang Jin, Jiancheng Yang, Kaiming Kuang, Bingbing Ni, Yiyi Gao, Yingli Sun, Pan Gao, Weiling Ma, Mingyu Tan, Hui Kang, et al. 2020. Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet. *EBioMedicine* 62 (2020), 103106.

[48] Kaggle. [n.d.]. 2018 Data Science Bowl. https://www.kaggle.com/competitions/data-science-bowl-2018/overview/about. Accessed: 2022-09-07.

[49] Kaggle. 2021. Satorius Cell Instance Segmentation 2021. https://www.kaggle.com/c/sartorius-cell-instance-segmentation. [Online; accessed 25-April-2022].

[50] Daanish Ali Khan, Linhong Li, Ninghao Sha, Zhuoran Liu, Abelino Jimenez, Bhiksha Raj, and Rita Singh. 2019. Non-Determinism in Neural Networks for Adversarial Robustness. *arXiv preprint arXiv:1905.10906* (2019).

[51] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. 2019. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9404–9413.

[52] Florian Kofler, Ivan Ezhov, Fabian Isensee, Christoph Berger, Maximilian Korner, Johannes Paetzold, Hongwei Li, Suprosanna Shit, Richard McKinley, Spyridon Bakas, et al. 2021. 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).

[53] Florian Kofler, Suprosanna Shit, Ivan Ezhov, Lucas Fidon, Rami Al-Maskari, Hongwei Li, Harsharan Bhatia, Timo Loehr, Marie Piraud, Ali Erturk, et al. 2022. blob loss: instance imbalance aware loss functions for semantic segmentation. *arXiv preprint arXiv:2205.08209* (2022).

[54] Ender Konukoglu, Ben Glocker, Dong Hye Ye, Antonio Criminisi, and Kilian M Pohl. 2012. Discriminative segmentation-based evaluation through shape dissimilarity. *IEEE transactions on medical imaging* 31, 12 (2012), 2278–2289.

[55] Jan Kottner, Laurent Audigé, Stig Brorson, Allan Donner, Byron J Gajewski, Asbjørn Hróbjartsson, Chris Roberts, Mohamed Shoukri, and David L Streiner. 2011. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International journal of nursing studies* 48, 6 (2011), 661–671.

[56] David A van Leeuwen and Niko Brümmer. 2007. An introduction to application-independent evaluation of speaker recognition systems. In *Speaker classification I*. Springer, 330–353.

[57] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[58] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P Bradley, Aaron Carass, et al. 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature communications* 9, 1 (2018), 1–13.

[59] Lena Maier-Hein, Annika Reinke, Michal Kozubek, Anne L Martel, Tal Arbel, Matthias Eisenmann, Allan Hanbury, Pierre Jannin, Henning Müller, Sinan Onogur, et al. 2020. BIAS: Transparent reporting of biomedical image analysis challenges. *Medical image analysis* 66 (2020), 101796.

[60] Lena Maier-Hein, Martin Wagner, Tobias Ross, Annika Reinke, Sebastian Bodenstedt, Peter M Full, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, et al. 2021. Heidelberg colorectal data set for surgical data

science in the sensor operating room. *Scientific data* 8, 1 (2021), 1–11.

[61] Lisa Mais, Peter Hirsch, and Dagmar Kainmueller. 2020. Patchperpix for instance segmentation. In *European Conference on Computer Vision*. Springer, 288–304.

[62] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. 2014. How to evaluate foreground maps?. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 248–255.

[63] Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405, 2 (1975), 442–451.

[64] Pavel Matula, Martin Maška, Dmitry V Sorokin, Petr Matula, Carlos Ortiz-de Solórzano, and Michal Kozubek. 2015. Cell tracking accuracy measurement based on comparison of acyclic oriented graphs. *PloS one* 10, 12 (2015), e0144959.

[65] G. Meissner, A. Nern, Z. Dorman, DePasquale G.M., K. Forster, T. Gibney, Hausenfluck J.H., Y. He, N. Iyer, J. Jeter, et al. 2022. A searchable image resource of Drosophila GAL4-driver expression patterns with single neuron resolution. *BioRxiv* (2022), 2020.05.29.080473.

[66] John Mongan, Linda Moy, and Charles E Kahn Jr. 2020. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiology. Artificial Intelligence* 2, 2 (2020).

[67] Karel GM Moons, Douglas G Altman, Johannes B Reitsma, John PA Ioannidis, Petra Macaskill, Ewout W Steyerberg, Andrew J Vickers, David F Ransohoff, and Gary S Collins. 2015. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine* 162, 1 (2015), W1–W73.

[68] Yukiko Nagao, Mika Sakamoto, Takumi Chinen, Yasushi Okada, and Daisuke Takao. 2020. Robust classification of cell cycle phase and biological feature extraction by image-based deep learning. *Molecular biology of the cell* 31, 13 (2020), 1346–1354.

[69] Ying-Hwey Nai, Bernice W Teo, Nadya L Tan, Sophie O'Doherty, Mary C Stephenson, Yee Liang Thian, Edmund Chiong, and Anthonin Reilhac. 2021. Comparison of metrics for the evaluation of medical segmentations using prostate MRI dataset. *Computers in Biology and Medicine* 134 (2021), 104497.

[70] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernadino Romera-Paredes, et al. 2021. Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. *Journal of Medical Internet Research* 23, 7 (2021), e26151.

[71] Annika Reinke, Matthias Eisenmann, Sinan Onogur, Marko Stankovic, Patrick Scholz, Peter M Full, Hrvoje Bogunovic, Bennett A Landman, Oskar Maier, Bjoern Menze, et al. 2018. How to exploit weaknesses in biomedical challenge design and organization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 388–395.

[72] Annika Reinke, Matthias Eisenmann, Minu D Tizabi, Carole H Sudre, Tim Rädsch, Michela Antonelli, Tal Arbel, Spyridon Bakas, M Jorge Cardoso, Veronika Cheplygina, et al. 2021. Common limitations of image processing metrics: A picture story. *arXiv preprint arXiv:2104.05642* (2021).

[73] Daniel Sage, Hagai Kirshner, Thomas Pengo, Nico Stuurman, Junhong Min, Suliana Manley, and Michael Unser. 2015. Quantitative evaluation of software packages for single-molecule localization microscopy. *Nature methods* 12, 8 (2015), 717–724.

[74] Frank W Samuelson and Nicholas Petrick. 2006. Comparing image detection algorithms using resampling. In *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006.* IEEE, 1312–1315.

[75] Cristina Sánchez-Montes, Francisco Javier Sánchez, Jorge Bernal, Henry Córdova, María López-Cerón, Miriam Cuatrecasas, Cristina Rodríguez De Miguel, Ana García-Rodríguez, Rodrigo Garcés-Durán, María Pellisé, et al. 2019. Computer-aided prediction of polyp histology on white light colonoscopy using surface pattern analysis. *Endoscopy* 51, 03 (2019), 261–265.

[76] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. 2018. Cell detection with star-convex polygons. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 265–273.

[77] Teddy Seidenfeld. 1985. Calibration, coherence, and scoring rules. *Philosophy of Science* 52, 2 (1985), 274–294.

[78] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. 2017. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Medical image analysis* 42 (2017), 1–13.

[79] Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien PW Pluim, Ulrich Bauer, and Bjoern H Menze. 2021. clDice-a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16560–16569.

[80] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063* (2019).

[81] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* 45 (2009), 427–437.

[82] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. 2021. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods* 18, 1 (2021), 100–106.

[83] Cecilia Summers and Michael J Dinneen. 2021. Nondeterminism and instability in neural network optimization. In *International Conference on Machine Learning*. PMLR, 9913–9922.

[84] Abdel Aziz Taha and Allan Hanbury. 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC medical imaging* 15, 1 (2015), 1–28.

[85] Alaa Tharwat. 2020. Classification assessment methods. *Applied Computing and Informatics* (2020).

[86] Laszlo Tirian and Barry J Dickson. 2017. The VT GAL4, LexA, and split-GAL4 driver line collections for targeted expression in the Drosophila nervous system. *BioRxiv* (2017), 198648.

[87] Vladimír Ulman, Martin Maška, Klas EG Magnusson, Olaf Ronneberger, Carsten Haubold, Nathalie Harder, Pavel Matula, Petr Matula, David Svoboda, Miroslav Radojevic, et al. 2017. An objective comparison of cell-tracking algorithms. *Nature methods* 14, 12 (2017), 1141–1152.

[88] Femke Vaassen, Colien Hazelaar, Ana Vaniqui, Mark Gooding, Brent van der Heyden, Richard Canters, and Wouter van Elmpt. 2020. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology* 13 (2020), 1–6.

[89] Bram Van Ginneken, Samuel G Armato III, Bartjan de Hoop, Saskia van Amelsvoort-van de Vorst, Thomas Duindam, Meindert Niemeijer, Keelin Murphy, Arnold Schilham, Alessandra Retico, Maria Evelina Fantacci, et al. 2010. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study. *Medical image analysis* 14, 6 (2010), 707–722.

[90] Andrew J Vickers, Ben Van Calster, and Ewout W Steyerberg. 2016. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *bmj* 352 (2016).

[91] Peter L Warburton, Jenna L Wang, and Paul G Mezey. 2008. On the balance of simplification and reality in molecular modeling of the electron density. *Journal of Chemical Theory and Computation* 4, 10 (2008), 1627–1636.

[92] Matthijs J Warrens. 2012. Some paradoxical results for the quadratically weighted kappa. *Psychometrika* 77, 2 (2012), 315–323.

[93] Manuel Wiesenfarth, Annika Reinke, Bennett A Landman, Matthias Eisenmann, Laura Aguilera Saiz, M Jorge Cardoso, Lena Maier-Hein, and Annette Kopp-Schneider. 2021. Methods and open-source toolkit for analyzing and visualizing challenge results. *Scientific reports* 11, 1 (2021), 1–15.

[94] Varduhi Yeghiazaryan and Irina Voiculescu. 2015. An overview of current evaluation methods used in medical image segmentation. *Department of Computer Science, University of Oxford* (2015).

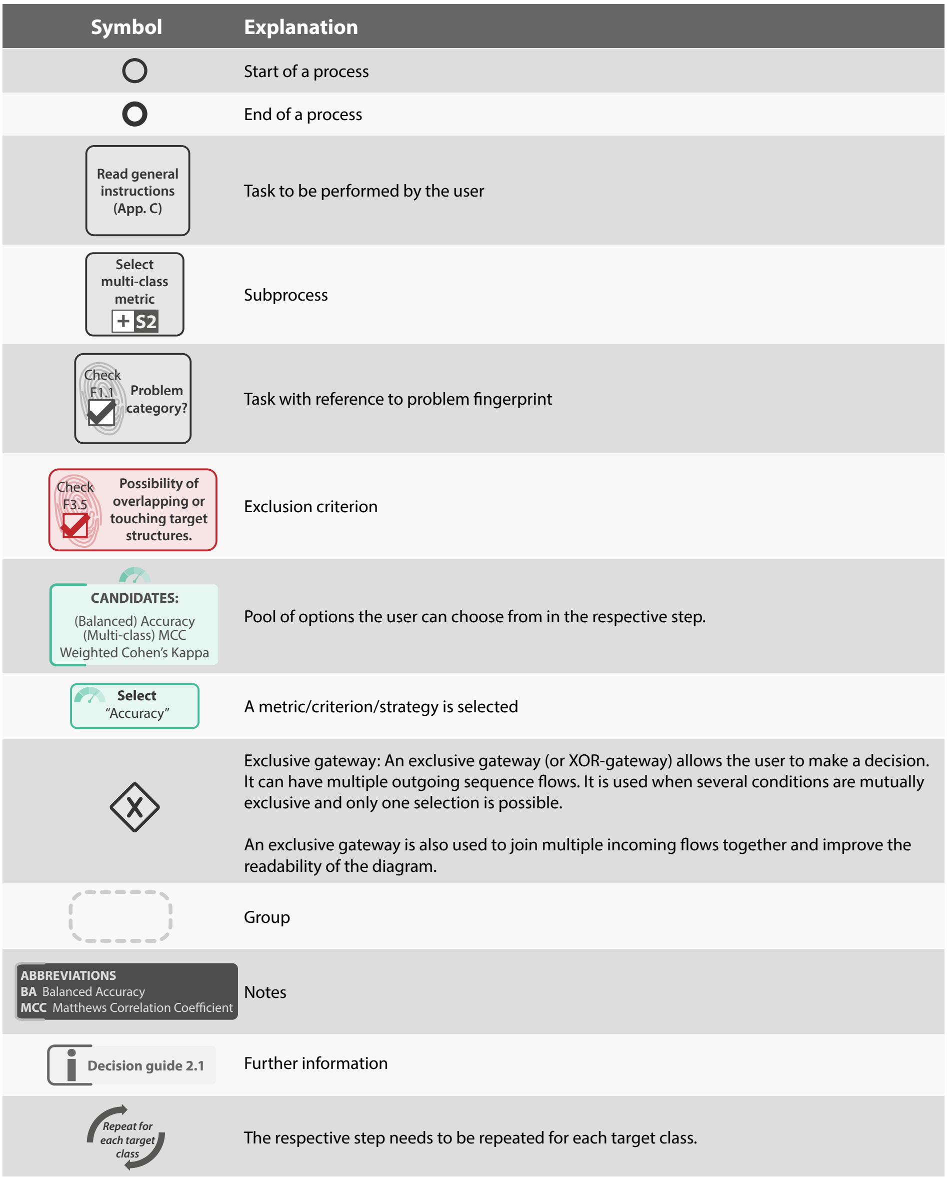
[95] Varduhi Yeghiazaryan and Irina D Voiculescu. 2018. Family of boundary overlap metrics for the evaluation of medical image segmentation. *Journal of Medical Imaging* 5, 1 (2018), 015006.

[96] Ying Zhang, Yubin Xie, Wenzhong Liu, Wankun Deng, Di Peng, Chenwei Wang, Haodong Xu, Chen Ruan, Yongjie Deng, Yaping Guo, et al. 2020. DeepPhagy: a deep learning framework for quantitatively measuring autophagy activity in Saccharomyces cerevisiae. *Autophagy* 16, 4 (2020), 626–640.

[97] Qiuming Zhu. 2020. On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. *Pattern Recognit. Lett.* 136 (2020), 71–80.

APPENDIX

A Symbol References



Symbol	Explanation
○	Start of a process
○	End of a process
Read general instructions (App. C)	Task to be performed by the user
Select multi-class metric + S2	Subprocess
Check F1.1 ☑ Problem category?	Task with reference to problem fingerprint
Check F3.5 ☑ Possibility of overlapping or touching target structures.	Exclusion criterion
CANDIDATES: (Balanced) Accuracy (Multi-class) MCC Weighted Cohen's Kappa	Pool of options the user can choose from in the respective step.
Select "Accuracy"	A metric/criterion/strategy is selected
X	Exclusive gateway: An exclusive gateway (or XOR-gateway) allows the user to make a decision. It can have multiple outgoing sequence flows. It is used when several conditions are mutually exclusive and only one selection is possible. An exclusive gateway is also used to join multiple incoming flows together and improve the readability of the diagram.
⬚	Group
ABBREVIATIONS BA Balanced Accuracy MCC Matthews Correlation Coefficient	Notes
i Decision guide 2.1	Further information
Repeat for each target class	The respective step needs to be repeated for each target class.

Fig. 6. **Overview of symbols used in the process diagrams.** The notation used in the process diagrams originates from Business Process Model and Notation (BPMN).

B Acronyms

AI Artificial Intelligence

AP Average Precision

ASSD Average Symmetric Surface Distance

AUROC Area under the Receiver Operating Characteristic Curve

BA Balanced Accuracy

BPMN Business Process Model and Notation

BS Brier Score

clDice Center line Dice Similarity Coefficient

CT Computed Tomography

DSC Dice Similarity Coefficient

EC Expected Cost

ECE Expected Calibration Error

FN False Negative

FP False Positive

FPPI False Positives per Image

FROC Free-Response Receiver Operating Characteristic

HD Hausdorff Distance

HD95 Hausdorff Distance 95% Percentile

IoU Intersection over Union

IoR Intersection over Reference

LR+ Positive Likelihood Ratio

LS Logarithmic Score

mAP Mean Average Precision

MASD Mean Average Surface Distance

MCC Matthews Correlation Coefficient

MICCAI Medical Image Computing and Computer Assisted Interventions

ML Machine Learning

MRI Magnetic Resonance Imaging

MS Multiple Sclerosis

NB Net Benefit

NPV Negative Predictive Value

NSD Normalized Surface Distance

PPV Positive Predictive Value

PQ Panoptic Quality

PHI Protected Health Information

PSR Proper Scoring Rules

ROC Receiver Operating Characteristic

ROI Region of Interest

TN True Negative

TNR True Negative Rate

TP True Positive

TPR True Positive Rate

WCK Weighted Cohen's Kappa

WSI Whole Slide Imaging

Xth Percentile HD Xth Percentile Hausdorff Distance

C General Instructions

Metrics Reloaded guides the user through the process of selecting and applying validation metrics in a problem-aware manner. The following guiding principles build the foundation of our framework.

Inclusion criteria. The *Metrics Reloaded* framework currently considers problems in which categorical output for a given *n*-dimensional input image is sought. Hence, it covers a broad range of imaging modalities from classical 2D/3D modalities, such as fluorescence, Computed Tomography (CT) or X-ray, to novel modalities such as spectral imaging modalities that yield high-dimensional output per pixel [25]. Classification can occur at pixel, object or image level, resulting in the four problem problem categories covered by the framework and depicted in Fig. 4:

Image-level classification refers to the assignment of one or multiple category labels to the entire image or fixed regions/predefined locations within an image.

Semantic segmentation refers to the assignment of one or multiple category labels to each pixel. For many segmentation problems, object boundaries are generated in addition to the pixel-wise classification images, which enables the computation of distance-based metrics, such as the Normalized Surface Distance (NSD).

Object detection refers to the localization and categorization of an unknown number of structures.

Instance segmentation refers to the localization and delineation of each distinct structure of a particular class. It can be regarded as delivering the tasks of object detection and semantic segmentation at the same time. In contrast to object detection, instance segmentation also involves the accurate marking of the structure boundary. In contrast to semantic segmentation, it distinguishes different structures of the same class.

Notably, the four different categories are mathematically closely related, as illustrated in Fig. 4. Application examples for all categories can be found in Fig. 5. Importantly, *Metrics Reloaded* does not require an entire image to be provided as input for the validation. For example, the classification of a Region of Interest (ROI) within a medical image may be required. In this example, the framework would proceed with the ROI as input as if it was an entire image. Furthermore, the shape of the image/input does not need to be rectangular. Finally, context information may be provided along with the input. For example, medical images may be processed along with clinical data to arrive at a diagnosis; video frames may be processed along with preceeding video snippets. What matters is that the *output* of the algorithm, for which a reference annotation is provided, is simply an *n*-dimensional image.

Phrasing of the biomedical task. The recommendation framework has been designed in a way to support the metric selection and application process for one specific driving biomedical question. In practice, multiple questions are often addressed with one given data set. For example, a clinician may have the ultimate interest of diagnosing brain cancer in a patient based on a given Magnetic Resonance Imaging (MRI) data set. While this would be phrased as an image-level classification task, an interesting *surrogate task* could be that of segmentation to assess the quality of tumor delineation. In the case of multiple different driving biomedical questions, a recommendation is generated separately for each question.

Matching reference annotations. The metric selection process begins with the step of mapping a given problem with all its intrinsic and data set-related properties to the corresponding problem category via the *category mapping* shown in Fig. 8. Our framework assumes that the reference annotations of the given dataset meet the requirements of the problem category that has been identified. Expected formats are provided in App. G.

Model-agnostic metric recommendation. Metrics should be chosen based solely on the driving biomedical problem and not be affected by algorithm design choices. For example, the error functions applied in common neural network architectures do not justify the use of corresponding metrics (e.g. validating with Dice Similarity Coefficient (DSC) to match the Dice loss used for training a neural network). Instead, the domain interest should guide the choice of metric, which, in turn, can guide the choice of the loss term.

Dealing with multiple classes. Multi-class metrics, such as Accuracy or Matthews Correlation Coefficient (MCC), have the unique advantage that they capture the performance of an algorithm for all classes in a single score without the need for customized class-aggregation schemes. On the other hand, they do now allow for detailed class-specific analyses. *Metrics Reloaded* therefore generally recommends performing a per-class validation for all target classes (in addition to potential multi-class validation). In segmentation problems, target classes are typically all classes except for a potential background class. As problem properties may differ from class to class (e.g. the size or size variability of target structures in segmentation problems), the problem fingerprint needs to be generated separately for each class. In consequence, several subprocesses (denoted by the ⊞-symbol in the framework overview shown in Fig. 2) need to be traversed separately for each target class in case the fingerprint properties relevant for the respective subprocess differ. Although not common in current validation practice, this may - in theory - lead to different validation metrics for different classes. We speak of class-specific metric pools in this case, which are generated in addition to the multi-class metric pool.

Primary and secondary metrics. In general, the biomedical interest cannot be captured with a single metric. The framework has therefore been designed to recommend multiple complementary metrics for a given task. We assume two primary use cases of metrics in this paper. In **comparative benchmarking studies** (e.g. competitive challenges), multiple algorithms or algorithm variants are compared on identical data sets. This requires the ranking of the competing algorithms according to performance. Typically, multiple complementary validation metrics are applied in this use case, resulting in either multiple rankings or a merged ranking that takes all or several metric values into account, as detailed in App. H. We refer to the metrics that contribute to the (primary) ranking(s) as *primary metrics. Secondary metrics* can additionally be applied for comprehensive reporting, for example because they reflect complementary properties of interest (e.g. compute time, carbon footprint), or for providing performance measures that are comparable across publications. The computer vision community, for example, typically reports the Intersection over Union (IoU) rather than the DSC. While our recommendation focuses on the recommendation of primary metrics, users are invited to complement them by further secondary metrics according to their specific needs. The second use case of metrics addressed by our framework are **validation studies centered around a single algorithm** that focus on comprehensive diagnostics rather than comparative assessment. In this case, it is often desired to report as many complementary metrics as possible in order to comprehensively analyse the properties of an algorithm. If a user of our framework is interested in this second type of study, the notion of primary and secondary metrics can be ignored.

Global decision rule. The proposed framework offers different strategies for validation of multi-class problems including application of multi-class metrics, aggregation of per-class scores (iterated e.g. in a “one versus rest” fashion), or selection of different metrics per class. In this context, it is important to note that a system in practice will typically convert the raw algorithm output into meaningful information that can be interpreted by a domain expert in a straightforward manner. For example, a diagnosis system will typically be designed to output one of multiple options rather than the continuous class scores of the algorithm. From a validation perspective, this implies that global cutoff strategy needs to be defined to avoid pitfalls such as the one shown in Fig. 7. Most importantly, even when a per-class validation is performed, this global strategy needs to be applied to avoid overestimation of algorithm performance. To address this important aspect, our recommendation shown in Fig. 2 comprises the step "Determine global decision threshold" before the metrics are applied to a given data set.

Notation. The notation for our recommendations have been based on BPMN[3]. The individual components used in the recommendation diagrams are explained in Fig. 6. Please note that we do not strictly follow BPMN to improve clarity of presentation.

Terminology. Terminology may differ substantially across communities. For example, the statistics community prefers the term Positive Predictive Value (PPV) over *Precision*, as the latter can be confused with the confidence of an output. In the medical domain, the term *validation* is used for an independent assessment (untouched test set) of an algorithm, while the machine learning community commonly uses a *validation set* for hyperparameter tuning. To avoid confusion resulting from unclear terminology, we follow the general terminology of [72] and have included a glossary in the App. J.

[3] https://www.omg.org/spec/BPMN/

Pitfall: Reporting AUROCs for multiple classes conceals the future need for a suboptimal global cutoff

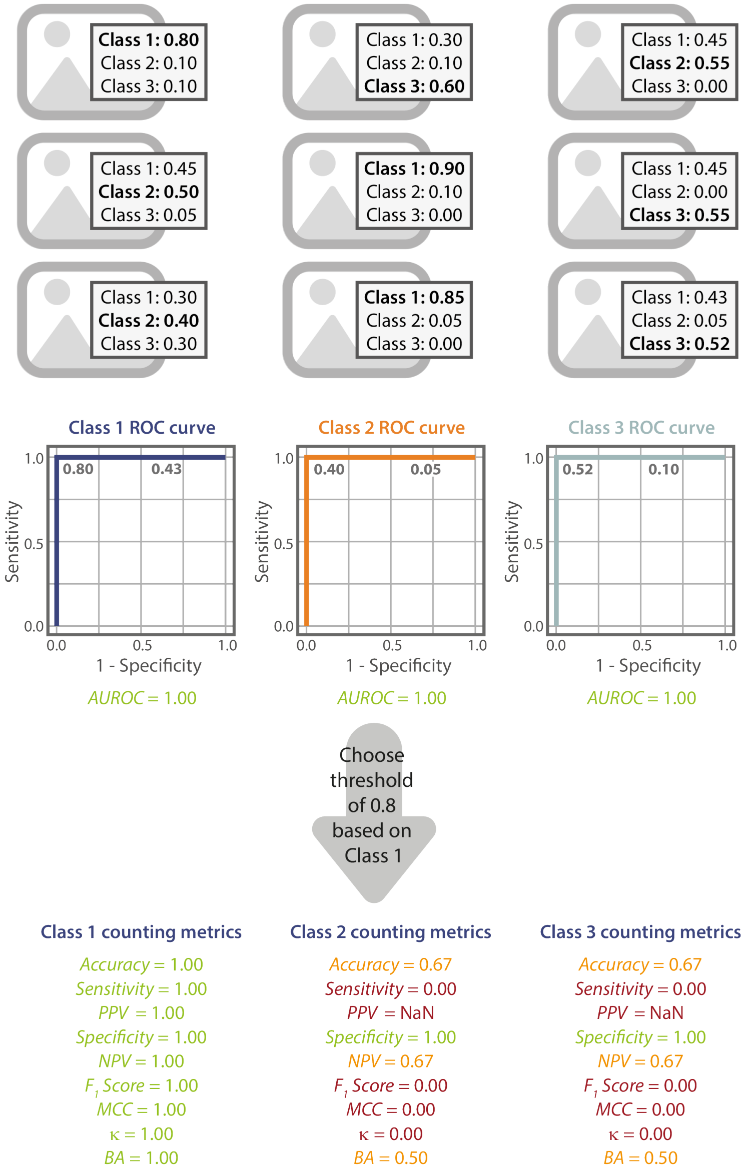


Fig. 7. Reporting the Area under the Receiver Operating Characteristic Curve (AUROC) for multiple classes may convey overly optimistic classifier performance. This is because any classifier upon application requires a global decision rule, which can not be optimized w.r.t to each class individually (see paragraph *Global decision rule*). Thus, in practice, Receiver Operating Characteristic (ROC) curves for most classes will be subject to a suboptimal cutoff. In the depicted example, an AUROC of 1.0 for each class allowing for individually optimized cutoffs. In contrast, the (realistic) application of a global cutoff does not lead to good performance for the classes it has not been optimized on. Used abbreviations: Positive Predictive Value (PPV), Negative Predictive Value (NPV), Matthews Correlation Coefficient (MCC), Cohen's Kappa κ and Balanced Accuracy (BA).

D Problem Fingerprinting

The problem fingerprints encapsulate the properties of the driving biomedical problem that are relevant for metric selection. The problem fingerprinting begins with the step of mapping a given problem with all its intrinsic and data set-related properties to the corresponding problem category via the *category mapping* shown in Fig. 8. Next, the user needs to instantiate the category-specific fingerprint provided in Figs. 10/11 (image-level classification), Figs. 12/13 (semantic segmentation), Figs. 14 - 16 (object detection) and Figs. 17 - 19 (instance segmentation).

It may not always be straightforward to instantiate a fingerprint item because of their binary/categorical nature. To address this issue, the *Metrics Reloaded* tool comprises a "Why are we asking this question?" button in each branch based on a fingerprint that may not be straightforward to instantiate. Furthermore, some fingerprint items depend on user preferences and/or deserve particular attention. These are the following:

F2.6: Cutoff on predicted class scores. Modern algorithms output (continuous) predicted class scores. To classify cases in an actual biomedical application (i.e. to make actual decisions), however, setting a cutoff value on the scores is required. Currently, the available strategies are not transparently motivated or distinguished in common practice. They are:

Target-value based Sometimes, the underlying problem provides a specific target metric value to be reached (e.g. Sensitivity of 0.95), requiring a corresponding cutoff value. In this case, we use the notation <Metric1>@<Metric2>, for example Specificity@Sensitivity, denoting the Specificity for a Sensitivity matching the target value (here 0.95).

Optimization-based If no specific target value is provided, a data-based cutoff can also be achieved by optimizing a primary metric (e.g. F_1 Score) using a dedicated data set for hyperparameter tuning.

Argmax-based An alternative widely used strategy is to simply determine the cutoff based on the "argmax" operation, which boils down to a threshold of 0.5 in binary classification problems.

Risk-based (for binary problems) In case the predicted class scores express the risk associated with a case belonging to a certain class (see F2.7), and there is a task-related risk threshold provided (e.g., only treat patients with cancer risk >10%), one can apply this threshold directly to the scores without data-driven optimization. Notably, provided risk thresholds correspond to a cost ratio of TP versus FP (e.g., not more than 10 FP per 1 TP should be treated). The Net Benefit (NB) metric considers this cost ratio as an inherent part of performance measure. This option is not suited for multi-class problems.

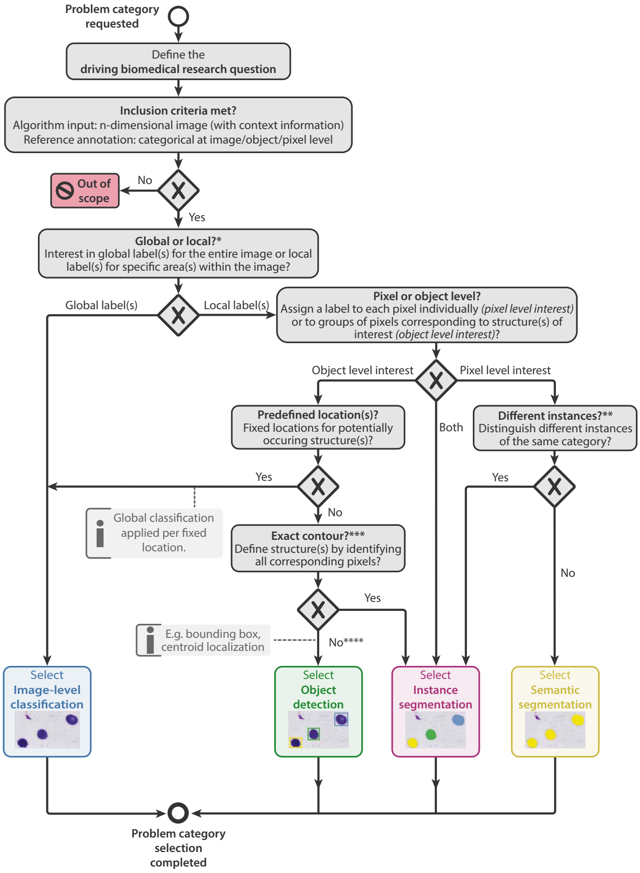
No cutoff A complementary strategy is to abstain from validating algorithms at a certain cutoff and exclusively report results on multi-threshold metrics (averaging over various cutoffs) instead.

The deciding factor for choosing whether or not to use a cutoff should be how much focus is to be put on a certain application at hand versus general and application-agnostic algorithm assessment. While some communities have converged to cutoff-based validation (e.g. cell instance segmentation [17]), recent clinical initiatives advocate for cutoff-agnostic validation arguing that cutoffs are often over-optimized on a specific data set and that associated results are not transferable across study cohorts (e.g. with differing disease prevalence) and clinical applications (e.g. with differing cost-benefit trade-offs for patients) [10, 67, 90]. We handle this controversy in current practices by making validation at certain cutoffs optional (for all tasks except semantic segmentation) and encoding user preferences in this fingerprint.

F2.7: Interpretability of predicted class scores requested. : When validating classification methods, it is often crucial for the predicted class scores themselves to be interpretable. This holds especially true for problems that involve direct human read-out. Interpretability is commonly achieved by requesting an algorithm to be well-calibrated. This is the case if p percent of all predictions reported at probability p are true [77]. In a clinical setting, for example, this would imply that out of all patients that are assigned a score of 0.9, 90% should actually belong to the target class (e.g. have a certain disease). In practice, many methods are overconfident, meaning that less than p percent of all predictions reported at probability p are true. Calibration is commonly measured by the Expected Calibration Error (ECE). An alternative to achieve interpretability are Proper Scoring Rules (PSR) [34], such as Brier Score (BS) and the Logarithmic Score (LS), which validate discrimination and calibration in a single score.

F4.2 Class prevalences reflect the population of interest. Class prevalences and their differences across data sets are highly important, although this aspect is often ignored in common validation practice. This can best be explained with the example of diagnostic tests (e.g. image-based classification of a disease). While several metrics, such as Sensitivity and Specificity, are independent of the occurrence of the target class and measure the inherent properties of the test, other metrics, such as Accuracy, measure the performance of the tests for the specific prevalence of the test set. This is not problematic if the class prevalences of the provided test set reflect the population of interest but can lead to problems otherwise. This fingerprint should hence be set to true if either the validation interest is constrained to the data set at hand (no future comparison to data sets with different class prevalences is desired) or no variation of prevalences is expected in other cohorts and upon application of the method.

+ S1


Problem category requested
Define the driving biomedical research question
Inclusion criteria met?
Algorithm input: n-dimensional image (with context information)
Reference annotation: categorical at image/object/pixel level
Out of scope
No
Yes
Global or local?*
Interest in global label(s) for the entire image or local label(s) for specific area(s) within the image?
Global label(s)
Local label(s)
Pixel or object level?
Assign a label to each pixel individually (pixel level interest) or to groups of pixels corresponding to structure(s) of interest (object level interest)?
Object level interest
Pixel level interest
Predefined location(s)?
Fixed locations for potentially occuring structure(s)?
Both
Different instances?**
Distinguish different instances of the same category?
Yes
Yes
Global classification applied per fixed location.
No
Exact contour?***
Define structure(s) by identifying all corresponding pixels?
No
Yes
E.g. bounding box, centroid localization
No****
Select Image-level classification
Select Object detection
Select Instance segmentation
Select Semantic segmentation
Problem category selection completed

* We define the scale of validation according to the confusion matrix, i.e. the scale at which predictions are matched with reference labels. **Global labels** refer to reference annotations on image level, **local labels** to reference annotations on object or pixel level.

** If multiple structures of the same type can be seen within the same image, this property should typically be set to true, to avoid issues with boundary-based metrics resulting from comparing a given structure boundary to the boundary of the wrong instance in the reference (Fig. 8).

*** If a substantial fraction of objects is small (F3.1), any outline-based definition becomes very brittle/noisy and experimenters might want to consider alternatives such as center point-based definitions (→ Object detection).

**** If the number of structures per category and image is known, the task would be considered an image-level regression problem and thus defined as **out of scope**.

Fig. 8. **Subprocess S1 for selecting a problem category.** The *Category Mapping* maps a given research problem to the appropriate problem category with the goal of **grouping problems by similarity of validation.** The leaf nodes represent the categories: image-level classification, object detection, instance segmentation or semantic segmentation. F3.1 refers to fingerprint 3.1 (see Fig. 13). An overview of the symbols used in the process diagram is provided in Fig. 6.

Pitfall: Comparing wrong boundaries

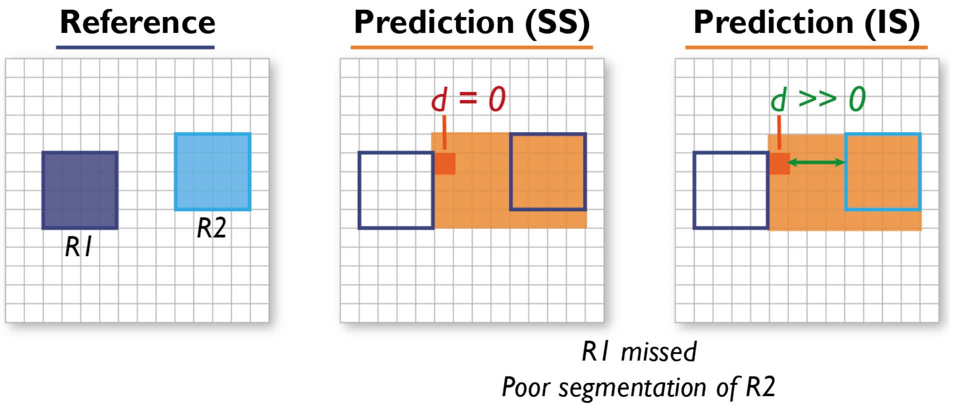


Fig. 9. **Boundary-based metrics in semantic/instance segmentation problems.** If multiple structures of the same type can be seen within the same image (here: reference objects *R1* and *R2*), it is generally advisable to formulate the problem as instance segmentation (IS; right) rather than semantic segmentation (SS; left) problem. This way, issues with boundary-based metrics resulting from comparing a given structure boundary to the boundary of the wrong instance in the reference, can be avoided. In the provided example, the distance of the red boundary pixel to the reference, as measured by a boundary-based metric in SS problems, would be zero, because different instances of the same structure cannot be distinguished. This problem is overcome by formulating the problem as an IS problem. In this case, (only) the boundary of the matched instance (here: R2) is considered for distance computation..

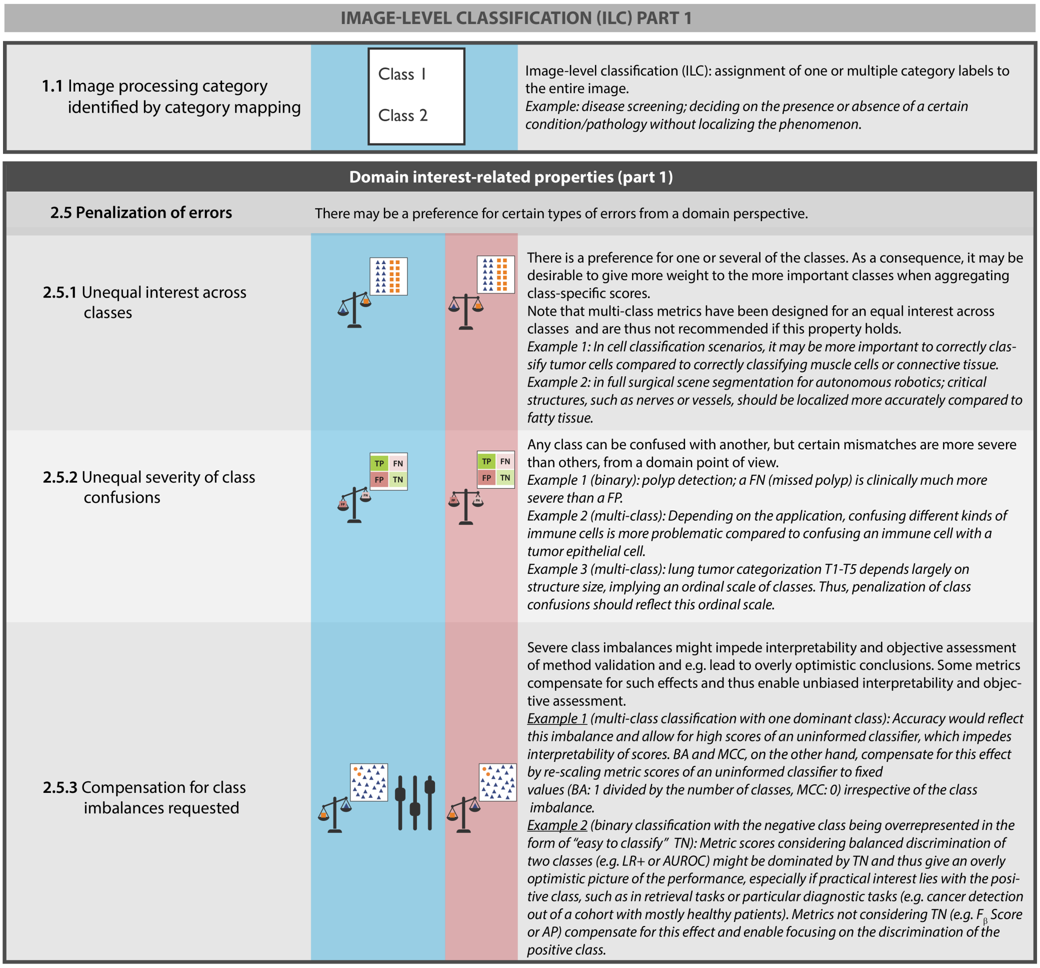


IMAGE-LEVEL CLASSIFICATION (ILC) PART 1

1.1 Image processing category identified by category mapping	Class I Class 2		Image-level classification (ILC): assignment of one or multiple category labels to the entire image. *Example: disease screening; deciding on the presence or absence of a certain condition/pathology without localizing the phenomenon.*

Domain interest-related properties (part 1)

2.5 Penalization of errors	There may be a preference for certain types of errors from a domain perspective.		
2.5.1 Unequal interest across classes			There is a preference for one or several of the classes. As a consequence, it may be desirable to give more weight to the more important classes when aggregating class-specific scores. Note that multi-class metrics have been designed for an equal interest across classes and are thus not recommended if this property holds. *Example 1: In cell classification scenarios, it may be more important to correctly classify tumor cells compared to correctly classifying muscle cells or connective tissue.* *Example 2: in full surgical scene segmentation for autonomous robotics; critical structures, such as nerves or vessels, should be localized more accurately compared to fatty tissue.*
2.5.2 Unequal severity of class confusions	TP FN FP TN	TP FN FP TN	Any class can be confused with another, but certain mismatches are more severe than others, from a domain point of view. *Example 1 (binary): polyp detection; a FN (missed polyp) is clinically much more severe than a FP.* *Example 2 (multi-class): Depending on the application, confusing different kinds of immune cells is more problematic compared to confusing an immune cell with a tumor epithelial cell.* *Example 3 (multi-class): lung tumor categorization T1-T5 depends largely on structure size, implying an ordinal scale of classes. Thus, penalization of class confusions should reflect this ordinal scale.*
2.5.3 Compensation for class imbalances requested			Severe class imbalances might impede interpretability and objective assessment of method validation and e.g. lead to overly optimistic conclusions. Some metrics compensate for such effects and thus enable unbiased interpretability and objective assessment. *Example 1 (multi-class classification with one dominant class): Accuracy would reflect this imbalance and allow for high scores of an uninformed classifier, which impedes interpretability of scores. BA and MCC, on the other hand, compensate for this effect by re-scaling metric scores of an uninformed classifier to fixed values (BA: 1 divided by the number of classes, MCC: 0) irrespective of the class imbalance.* *Example 2 (binary classification with the negative class being overrepresented in the form of "easy to classify" TN): Metric scores considering balanced discrimination of two classes (e.g. LR+ or AUROC) might be dominated by TN and thus give an overly optimistic picture of the performance, especially if practical interest lies with the positive class, such as in retrieval tasks or particular diagnostic tasks (e.g. cancer detection out of a cohort with mostly healthy patients). Metrics not considering TN (e.g. F_β Score or AP) compensate for this effect and enable focusing on the discrimination of the positive class.*

Fig. 10. **Fingerprint for image-level classification (Part 1).** In the case of binary fingerprints, the blue column shows examples for which the property is true while the red column shows counterexamples. Categorical fingerprints are only shown in blue.

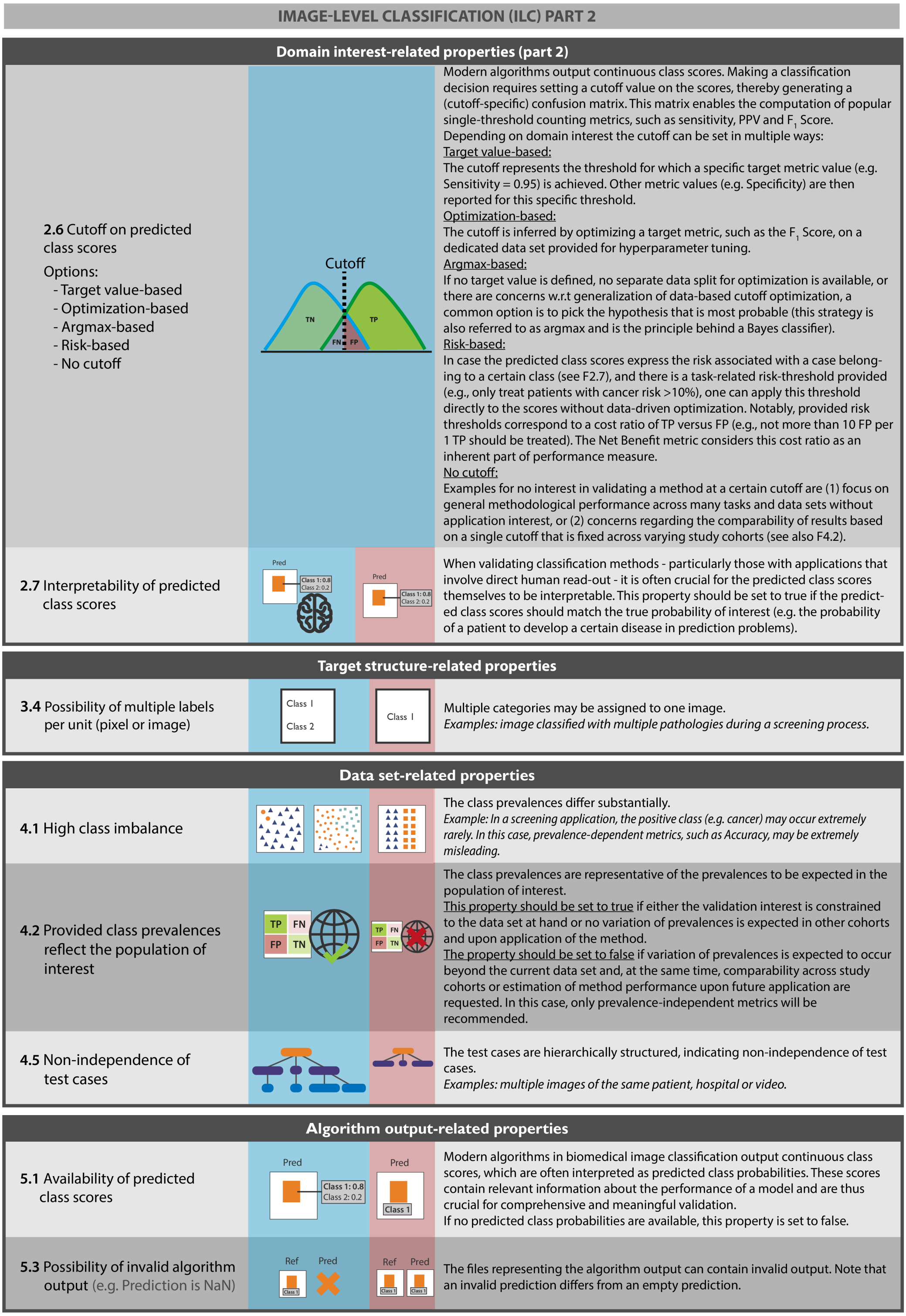


IMAGE-LEVEL CLASSIFICATION (ILC) PART 2

Domain interest-related properties (part 2)

2.6 Cutoff on predicted class scores

Options:
- Target value-based
- Optimization-based
- Argmax-based
- Risk-based
- No cutoff

Modern algorithms output continuous class scores. Making a classification decision requires setting a cutoff value on the scores, thereby generating a (cutoff-specific) confusion matrix. This matrix enables the computation of popular single-threshold counting metrics, such as sensitivity, PPV and F_1 Score. Depending on domain interest the cutoff can be set in multiple ways:

Target value-based:
The cutoff represents the threshold for which a specific target metric value (e.g. Sensitivity = 0.95) is achieved. Other metric values (e.g. Specificity) are then reported for this specific threshold.

Optimization-based:
The cutoff is inferred by optimizing a target metric, such as the F_1 Score, on a dedicated data set provided for hyperparameter tuning.

Argmax-based:
If no target value is defined, no separate data split for optimization is available, or there are concerns w.r.t generalization of data-based cutoff optimization, a common option is to pick the hypothesis that is most probable (this strategy is also referred to as argmax and is the principle behind a Bayes classifier).

Risk-based:
In case the predicted class scores express the risk associated with a case belonging to a certain class (see F2.7), and there is a task-related risk-threshold provided (e.g., only treat patients with cancer risk >10%), one can apply this threshold directly to the scores without data-driven optimization. Notably, provided risk thresholds correspond to a cost ratio of TP versus FP (e.g., not more than 10 FP per 1 TP should be treated). The Net Benefit metric considers this cost ratio as an inherent part of performance measure.

No cutoff:
Examples for no interest in validating a method at a certain cutoff are (1) focus on general methodological performance across many tasks and data sets without application interest, or (2) concerns regarding the comparability of results based on a single cutoff that is fixed across varying study cohorts (see also F4.2).

2.7 Interpretability of predicted class scores

When validating classification methods - particularly those with applications that involve direct human read-out - it is often crucial for the predicted class scores themselves to be interpretable. This property should be set to true if the predicted class scores should match the true probability of interest (e.g. the probability of a patient to develop a certain disease in prediction problems).

Target structure-related properties

3.4 Possibility of multiple labels per unit (pixel or image)

Multiple categories may be assigned to one image.
Examples: image classified with multiple pathologies during a screening process.

Data set-related properties

4.1 High class imbalance

The class prevalences differ substantially.
Example: In a screening application, the positive class (e.g. cancer) may occur extremely rarely. In this case, prevalence-dependent metrics, such as Accuracy, may be extremely misleading.

4.2 Provided class prevalences reflect the population of interest

The class prevalences are representative of the prevalences to be expected in the population of interest.
This property should be set to true if either the validation interest is constrained to the data set at hand or no variation of prevalences is expected in other cohorts and upon application of the method.
The property should be set to false if variation of prevalences is expected to occur beyond the current data set and, at the same time, comparability across study cohorts or estimation of method performance upon future application are requested. In this case, only prevalence-independent metrics will be recommended.

4.5 Non-independence of test cases

The test cases are hierarchically structured, indicating non-independence of test cases.
Examples: multiple images of the same patient, hospital or video.

Algorithm output-related properties

5.1 Availability of predicted class scores

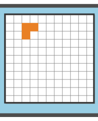
Modern algorithms in biomedical image classification output continuous class scores, which are often interpreted as predicted class probabilities. These scores contain relevant information about the performance of a model and are thus crucial for comprehensive and meaningful validation.
If no predicted class probabilities are available, this property is set to false.

5.3 Possibility of invalid algorithm output (e.g. Prediction is NaN)

The files representing the algorithm output can contain invalid output. Note that an invalid prediction differs from an empty prediction.

Fig. 11. **Fingerprint for image-level classification (Part 2).** In the case of binary fingerprints, the blue column shows examples for which the property is true while the red column shows counterexamples. Categorical fingerprints are only shown in blue.

SEMANTIC SEGMENTATION (SS) PART 1

1.1 Image processing category identified by category mapping			Semantic segmentation (SS): assignment of one or multiple category labels to each pixel. *Example: surgical scene segmentation for autonomous robotics; assigning each pixel the corresponding structure/organ/pathology label.*

Domain interest-related properties

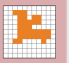
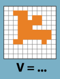
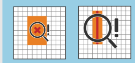
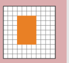
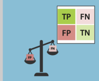
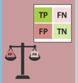
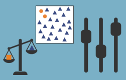
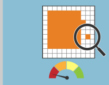
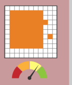
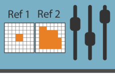
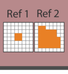
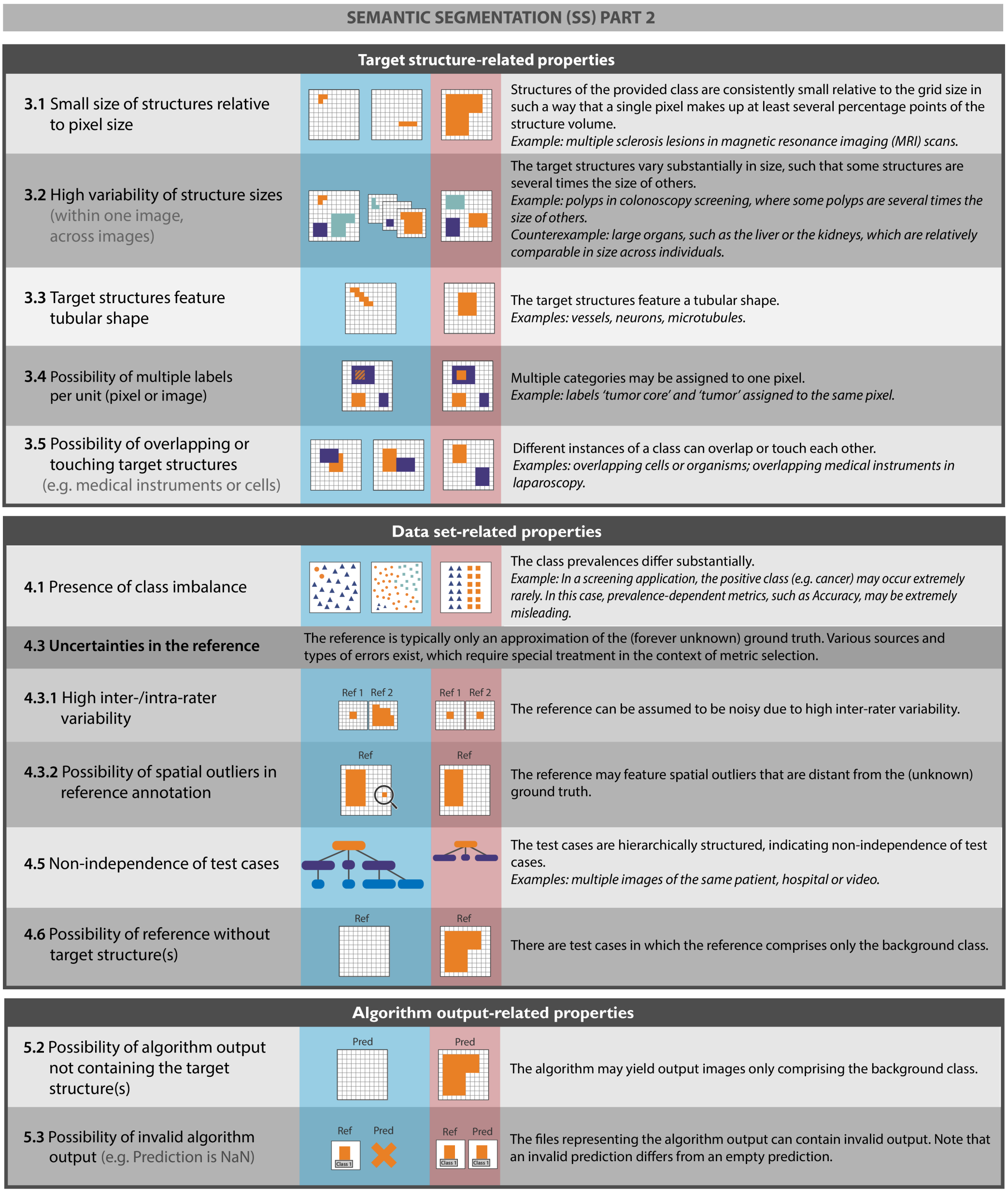
2.1 Particular importance of structure boundaries			The biomedical application requires exact structure boundaries. *Example: segmentation for radiotherapy planning; knowledge of exact structure boundaries is crucial to destroy the tumor while sparing healthy tissue.* Important: Overlap-based metrics do not measure shape agreement. In the case of complex shapes (high boundary-to-volume ratio) it is therefore typically advisable to set this property to true.
2.2 Particular importance of structure volume			The biomedical application requires accurate knowledge of structure volumes. *Example: liver segmentation as basis for remnant liver volume computation in surgical resection planning.*
2.3 Particular importance of structure center (e.g. in cells, vessels)			The biomedical application requires accurate knowledge of structure centers. *Example: cell centers are subsequently used for cell tracking and cell motion characterisation, so false center movement should be suppressed.*
2.5 Penalization of errors	There may be a preference for certain types of errors from a domain perspective.		
2.5.1 Unequal interest across classes			There is a preference for one or several of the classes. As a consequence, it may be desirable to give more weight to the more important classes when aggregating class-specific scores. *Example 1: In cell classification scenarios, it may be more important to correctly classify tumor cells compared to correctly classifying muscle cells or connective tissue.* *Example 2: in full surgical scene segmentation for autonomous robotics; critical structures, such as nerves or vessels, should be localized more accurately compared to fatty tissue.*
2.5.2 Unequal severity of class confusions			Any class can be confused with another, but certain mismatches are more severe than others, from a domain point of view. *Example 1 (binary): polyp detection; a FN (missed polyp) is clinically much more severe than a FP.* *Example 2 (multi-class): Depending on the application, confusing different kinds of immune cells is more problematic compared to confusing an immune cell with a tumor epithelial cell.* *Example 3 (multi-class): Lung tumor categorization T1-T5 depends largely on structure size, implying an ordinal scale of classes. Thus, penalization of class confusions should reflect this ordinal scale.*
2.5.3 Compensation for class imbalances requested			Severe class imbalances might impede interpretability and objective assessment of method validation and e.g. lead to overly optimistic conclusions. Some metrics compensate for such effects and thus enable unbiased interpretability and objective assessment. *Example (binary classification with the negative class being overrepresented in the form of "easy to classify" TN): Metric scores considering balanced discrimination of two classes (e.g. LR+ or AUROC) might be dominated by TN and thus give an overly optimistic picture of the performance, especially if practical interest lies with the positive class, such as in retrieval tasks or particular diagnostic tasks (e.g. cancer detection out of a cohort with mostly healthy patients). Metrics not considering TN (e.g. F_β Score or AP) compensate for this effect and enable focusing on the discrimination of the positive class.*
2.5.4 Handling of spatial outliers			Spatial outliers are FP predictions that feature a large distance to the reference. They can be handled in three different ways: Distance-based penalization with outlier focus: Outliers should be heavily penalized as a function of the distance to the reference. Distance-based penalization with contour focus: Outliers should be penalized as a function of the distance to the reference, but the assessment should focus on the general contour agreement rather than individual outliers. Existence-based penalization: The existence of spatial outliers should be penalized irrespective of their distance to the reference. Distance-based penalization is not possible when either the reference or the prediction is empty. In applications in which many such cases potentially occur, we therefore recommend an existence-based penalization.
2.5.5 Compensation for annotation imprecisions requested			The reference annotation is typically only an approximation of the (forever unknown) ground truth. It may be desirable to compensate for known uncertainties, such as intra-rater or inter-rater variability, by configuring the metric accordingly. This is only possible for some metrics.

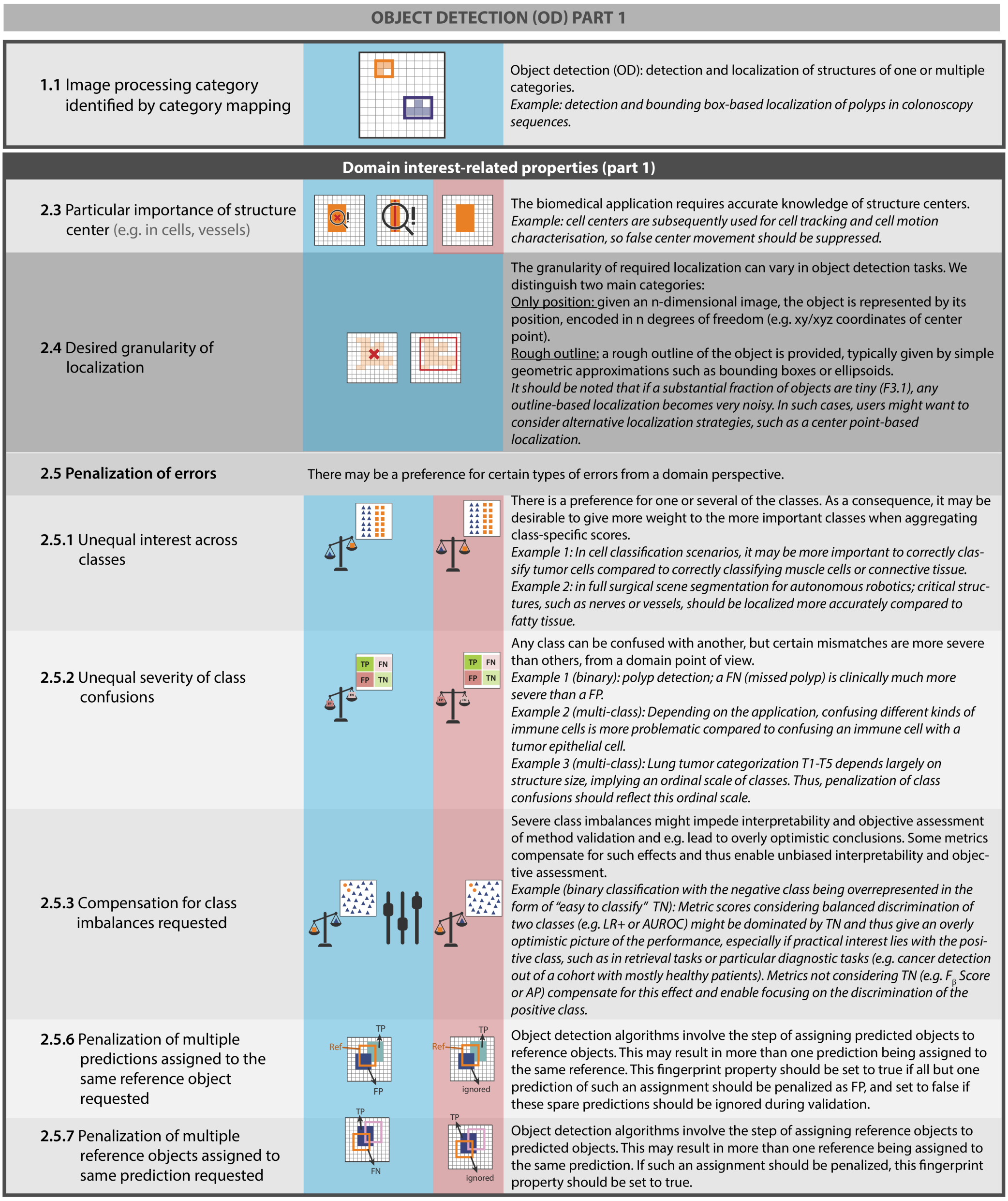
Fig. 12. **Fingerprint for semantic segmentation (Part 1).** In the case of binary fingerprints, the blue column shows examples for which the property is true while the red column shows counterexamples. Categorical fingerprints are only shown in blue.



SEMANTIC SEGMENTATION (SS) PART 2

Target structure-related properties			
3.1 Small size of structures relative to pixel size			Structures of the provided class are consistently small relative to the grid size in such a way that a single pixel makes up at least several percentage points of the structure volume. *Example: multiple sclerosis lesions in magnetic resonance imaging (MRI) scans.*
3.2 High variability of structure sizes (within one image, across images)			The target structures vary substantially in size, such that some structures are several times the size of others. *Example: polyps in colonoscopy screening, where some polyps are several times the size of others.* *Counterexample: large organs, such as the liver or the kidneys, which are relatively comparable in size across individuals.*
3.3 Target structures feature tubular shape			The target structures feature a tubular shape. *Examples: vessels, neurons, microtubules.*
3.4 Possibility of multiple labels per unit (pixel or image)			Multiple categories may be assigned to one pixel. *Example: labels 'tumor core' and 'tumor' assigned to the same pixel.*
3.5 Possibility of overlapping or touching target structures (e.g. medical instruments or cells)			Different instances of a class can overlap or touch each other. *Examples: overlapping cells or organisms; overlapping medical instruments in laparoscopy.*
Data set-related properties			
4.1 Presence of class imbalance			The class prevalences differ substantially. *Example: In a screening application, the positive class (e.g. cancer) may occur extremely rarely. In this case, prevalence-dependent metrics, such as Accuracy, may be extremely misleading.*
4.3 Uncertainties in the reference	The reference is typically only an approximation of the (forever unknown) ground truth. Various sources and types of errors exist, which require special treatment in the context of metric selection.		
4.3.1 High inter-/intra-rater variability	Ref 1 Ref 2	Ref 1 Ref 2	The reference can be assumed to be noisy due to high inter-rater variability.
4.3.2 Possibility of spatial outliers in reference annotation	Ref	Ref	The reference may feature spatial outliers that are distant from the (unknown) ground truth.
4.5 Non-independence of test cases			The test cases are hierarchically structured, indicating non-independence of test cases. *Examples: multiple images of the same patient, hospital or video.*
4.6 Possibility of reference without target structure(s)	Ref	Ref	There are test cases in which the reference comprises only the background class.
Algorithm output-related properties			
5.2 Possibility of algorithm output not containing the target structure(s)	Pred	Pred	The algorithm may yield output images only comprising the background class.
5.3 Possibility of invalid algorithm output (e.g. Prediction is NaN)	Ref Pred Class 1	Ref Pred Class 1 Class 1	The files representing the algorithm output can contain invalid output. Note that an invalid prediction differs from an empty prediction.

Fig. 13. **Fingerprint for semantic segmentation (Part 2).** In the case of binary fingerprints, the blue column shows examples for which the property is true while the red column shows counterexamples. Categorical fingerprints are only shown in blue.



TP
FN
FP
TN
Ref
TP
FP
ignored
TP
FN
ignored

OBJECT DETECTION (OD) PART 1

1.1 Image processing category identified by category mapping	Object detection (OD): detection and localization of structures of one or multiple categories. *Example: detection and bounding box-based localization of polyps in colonoscopy sequences.*

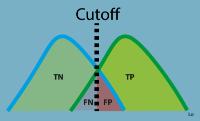
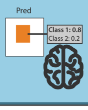
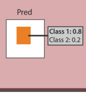
Domain interest-related properties (part 1)

2.3 Particular importance of structure center (e.g. in cells, vessels)	The biomedical application requires accurate knowledge of structure centers. *Example: cell centers are subsequently used for cell tracking and cell motion characterisation, so false center movement should be suppressed.*
2.4 Desired granularity of localization	The granularity of required localization can vary in object detection tasks. We distinguish two main categories: Only position: given an n-dimensional image, the object is represented by its position, encoded in n degrees of freedom (e.g. xy/xyz coordinates of center point). Rough outline: a rough outline of the object is provided, typically given by simple geometric approximations such as bounding boxes or ellipsoids. *It should be noted that if a substantial fraction of objects are tiny (F3.1), any outline-based localization becomes very noisy. In such cases, users might want to consider alternative localization strategies, such as a center point-based localization.*
2.5 Penalization of errors	There may be a preference for certain types of errors from a domain perspective.
2.5.1 Unequal interest across classes	There is a preference for one or several of the classes. As a consequence, it may be desirable to give more weight to the more important classes when aggregating class-specific scores. *Example 1: In cell classification scenarios, it may be more important to correctly classify tumor cells compared to correctly classifying muscle cells or connective tissue.* *Example 2: in full surgical scene segmentation for autonomous robotics; critical structures, such as nerves or vessels, should be localized more accurately compared to fatty tissue.*
2.5.2 Unequal severity of class confusions	Any class can be confused with another, but certain mismatches are more severe than others, from a domain point of view. *Example 1 (binary): polyp detection; a FN (missed polyp) is clinically much more severe than a FP.* *Example 2 (multi-class): Depending on the application, confusing different kinds of immune cells is more problematic compared to confusing an immune cell with a tumor epithelial cell.* *Example 3 (multi-class): Lung tumor categorization T1-T5 depends largely on structure size, implying an ordinal scale of classes. Thus, penalization of class confusions should reflect this ordinal scale.*
2.5.3 Compensation for class imbalances requested	Severe class imbalances might impede interpretability and objective assessment of method validation and e.g. lead to overly optimistic conclusions. Some metrics compensate for such effects and thus enable unbiased interpretability and objective assessment. *Example (binary classification with the negative class being overrepresented in the form of "easy to classify" TN): Metric scores considering balanced discrimination of two classes (e.g. LR+ or AUROC) might be dominated by TN and thus give an overly optimistic picture of the performance, especially if practical interest lies with the positive class, such as in retrieval tasks or particular diagnostic tasks (e.g. cancer detection out of a cohort with mostly healthy patients). Metrics not considering TN (e.g. F_β Score or AP) compensate for this effect and enable focusing on the discrimination of the positive class.*
2.5.6 Penalization of multiple predictions assigned to the same reference object requested	Object detection algorithms involve the step of assigning predicted objects to reference objects. This may result in more than one prediction being assigned to the same reference. This fingerprint property should be set to true if all but one prediction of such an assignment should be penalized as FP, and set to false if these spare predictions should be ignored during validation.
2.5.7 Penalization of multiple reference objects assigned to same prediction requested	Object detection algorithms involve the step of assigning reference objects to predicted objects. This may result in more than one reference being assigned to the same prediction. If such an assignment should be penalized, this fingerprint property should be set to true.

Fig. 14. **Fingerprint for object detection (Part 1).** In the case of binary fingerprints, the blue column shows examples for which the property is true while the red column shows counterexamples. Categorical fingerprints are only shown in blue.

OBJECT DETECTION (OD) PART 2

Domain interest-related properties (part 2)

2.6 Cutoff on predicted class scores Options: - Target value-based - Optimization-based - Argmax-based - No cutoff			Modern algorithms output continuous class scores. Making a classification decision requires setting a cutoff value on the scores, thereby generating a (cutoff-specific) confusion matrix. This matrix enables the computation of popular single-threshold counting metrics, such as sensitivity, PPV and F_1 Score. Depending on domain interest the cutoff can be set in multiple ways: Target value-based: The cutoff represents the threshold for which a specific target metric value (e.g. Sensitivity = 0.95) is achieved. Other metric values (e.g. Specificity) are then reported for this specific threshold. Optimization-based: The cutoff is inferred by optimizing a target metric, such as the F_1 Score, on a dedicated data set provided for hyperparameter tuning. Argmax-based: If no target value is defined, no separate data split for optimization is available, or there are concerns w.r.t generalization of data-based cutoff optimization, a common option is to pick the hypothesis that is most probable (this strategy is also referred to as argmax and is the principle behind a Bayes classifier). No cutoff: Examples for no interest in validating a method at a certain cutoff are (1) focus on general methodological performance across many tasks and data sets without application interest, or (2) concerns regarding the comparability of results based on a single cutoff that is fixed across varying study cohorts (see also F4.2).
2.7 Interpretability of predicted class scores			When validating classification methods - particularly those with applications that involve direct human read-out - it is often crucial for the predicted class scores themselves to be interpretable. This property should be set to true if the predicted class scores should match the true probability of interest (e.g. the probability of a patient to develop a certain disease in prediction problems).

Target structure-related properties

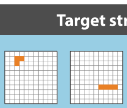
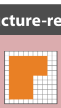
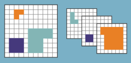
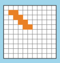
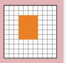
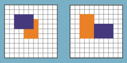
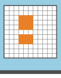
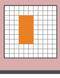
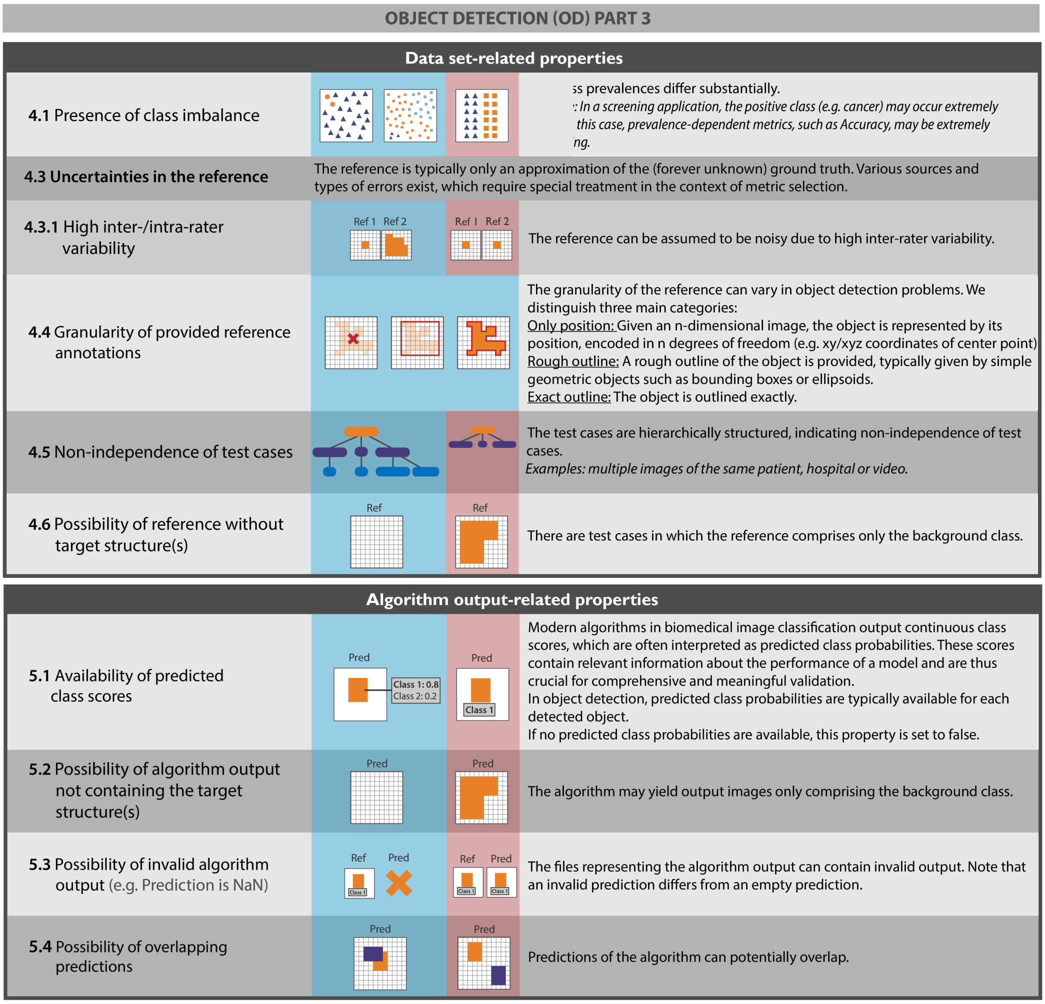
3.1 Small size of structures relative to pixel size			Structures of the provided class are consistently small relative to the grid size in such a way that a single pixel makes up at least several percentage points of the structure volume. *Example: multiple sclerosis lesions in magnetic resonance imaging (MRI) scans.*
3.2 High variability of structure sizes (within one image, across images)			The target structures vary substantially in size, such that some structures are several times the size of others. *Example: polyps in colonoscopy screening, where some polyps are several times the size of others.* *Counterexample: large organs, such as the liver or the kidneys, which are relatively comparable in size across individuals.*
3.3 Target structures feature tubular shape			The target structures feature a tubular shape. *Examples: vessels, neurons, microtubules.*
3.5 Possibility of overlapping or touching target structures (e.g. medical instruments or cells)			Different instances of a class can overlap or touch each other. *Examples: overlapping cells or organisms; overlapping medical instruments in laparoscopy.*
3.6 Possibility of disconnected target structure(s)			A given structure appears disconnected in the given image. *Examples: neurons in 2D microscopy of a slice of tissue; single tomographic image slice depicting complex vessels. vv*

Fig. 15. **Fingerprint for object detection (Part 2).** In the case of binary fingerprints, the blue column shows examples for which the property is true while the red column shows counterexamples. Categorical fingerprints are only shown in blue.

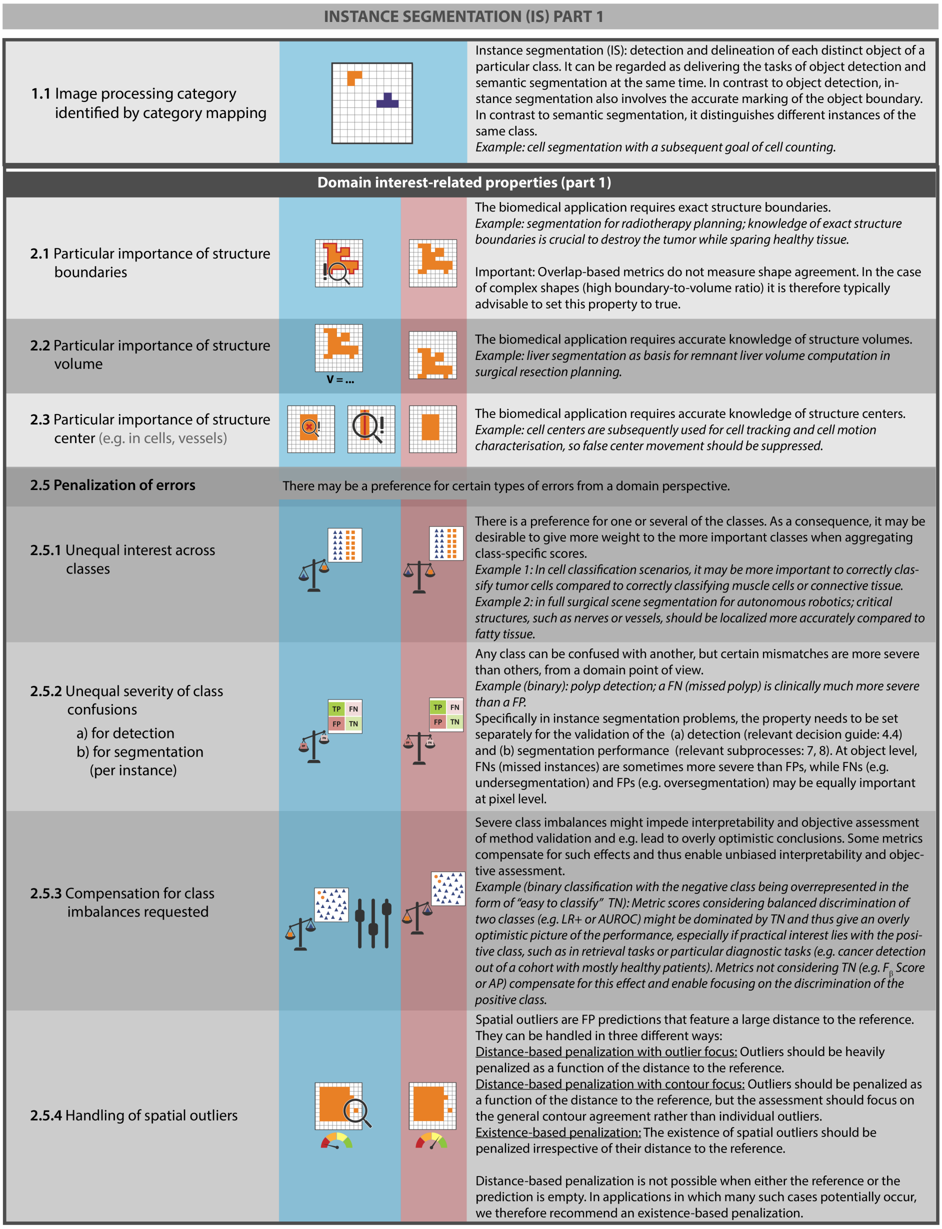


OBJECT DETECTION (OD) PART 3

Data set-related properties	
4.1 Presence of class imbalance	ss prevalences differ substantially. *: In a screening application, the positive class (e.g. cancer) may occur extremely this case, prevalence-dependent metrics, such as Accuracy, may be extremely ng.*
4.3 Uncertainties in the reference	The reference is typically only an approximation of the (forever unknown) ground truth. Various sources and types of errors exist, which require special treatment in the context of metric selection.
4.3.1 High inter-/intra-rater variability	The reference can be assumed to be noisy due to high inter-rater variability.
4.4 Granularity of provided reference annotations	The granularity of the reference can vary in object detection problems. We distinguish three main categories: Only position: Given an n-dimensional image, the object is represented by its position, encoded in n degrees of freedom (e.g. xy/xyz coordinates of center point) Rough outline: A rough outline of the object is provided, typically given by simple geometric objects such as bounding boxes or ellipsoids. Exact outline: The object is outlined exactly.
4.5 Non-independence of test cases	The test cases are hierarchically structured, indicating non-independence of test cases. *Examples: multiple images of the same patient, hospital or video.*
4.6 Possibility of reference without target structure(s)	There are test cases in which the reference comprises only the background class.

Algorithm output-related properties	
5.1 Availability of predicted class scores	Modern algorithms in biomedical image classification output continuous class scores, which are often interpreted as predicted class probabilities. These scores contain relevant information about the performance of a model and are thus crucial for comprehensive and meaningful validation. In object detection, predicted class probabilities are typically available for each detected object. If no predicted class probabilities are available, this property is set to false.
5.2 Possibility of algorithm output not containing the target structure(s)	The algorithm may yield output images only comprising the background class.
5.3 Possibility of invalid algorithm output (e.g. Prediction is NaN)	The files representing the algorithm output can contain invalid output. Note that an invalid prediction differs from an empty prediction.
5.4 Possibility of overlapping predictions	Predictions of the algorithm can potentially overlap.

Fig. 16. **Fingerprint for object detection (Part 3).** In the case of binary fingerprints, the blue column shows examples for which the property is true while the red column shows counterexamples. Categorical fingerprints are only shown in blue.



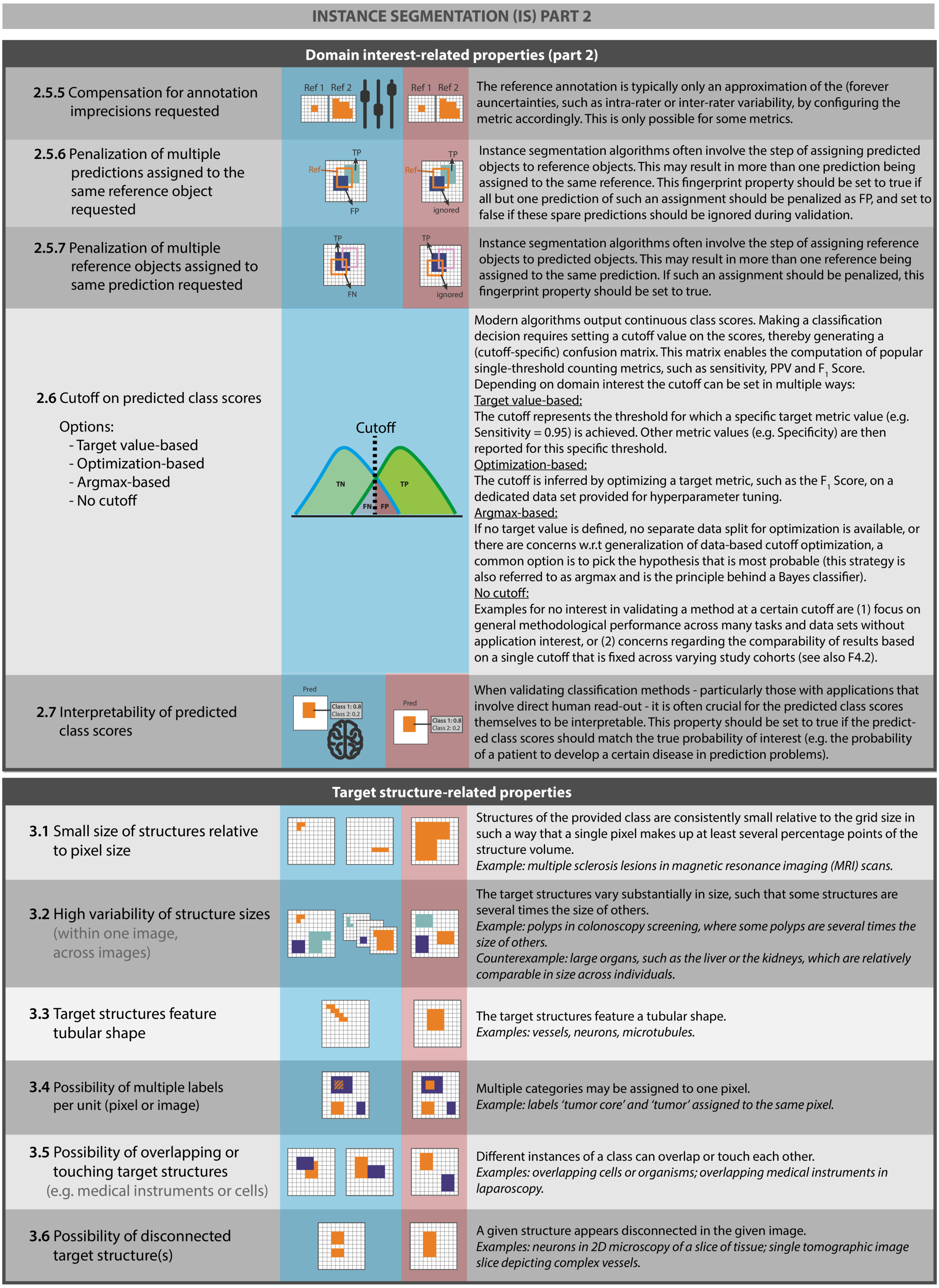
INSTANCE SEGMENTATION (IS) PART 1

1.1 Image processing category identified by category mapping	Instance segmentation (IS): detection and delineation of each distinct object of a particular class. It can be regarded as delivering the tasks of object detection and semantic segmentation at the same time. In contrast to object detection, instance segmentation also involves the accurate marking of the object boundary. In contrast to semantic segmentation, it distinguishes different instances of the same class. *Example: cell segmentation with a subsequent goal of cell counting.*

Domain interest-related properties (part 1)

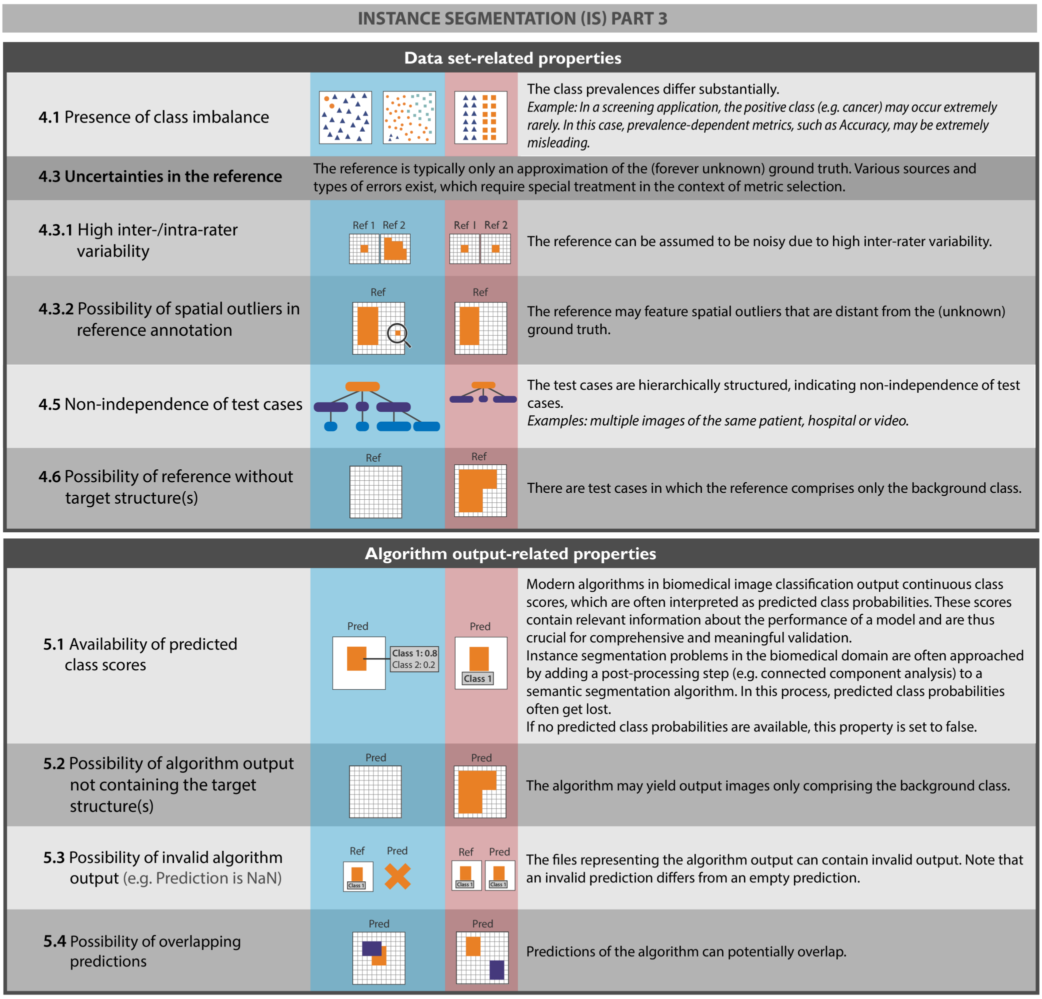
2.1 Particular importance of structure boundaries	The biomedical application requires exact structure boundaries. *Example: segmentation for radiotherapy planning; knowledge of exact structure boundaries is crucial to destroy the tumor while sparing healthy tissue.* Important: Overlap-based metrics do not measure shape agreement. In the case of complex shapes (high boundary-to-volume ratio) it is therefore typically advisable to set this property to true.
2.2 Particular importance of structure volume	The biomedical application requires accurate knowledge of structure volumes. *Example: liver segmentation as basis for remnant liver volume computation in surgical resection planning.*
2.3 Particular importance of structure center (e.g. in cells, vessels)	The biomedical application requires accurate knowledge of structure centers. *Example: cell centers are subsequently used for cell tracking and cell motion characterisation, so false center movement should be suppressed.*
2.5 Penalization of errors	There may be a preference for certain types of errors from a domain perspective.
2.5.1 Unequal interest across classes	There is a preference for one or several of the classes. As a consequence, it may be desirable to give more weight to the more important classes when aggregating class-specific scores. *Example 1: In cell classification scenarios, it may be more important to correctly classify tumor cells compared to correctly classifying muscle cells or connective tissue.* *Example 2: in full surgical scene segmentation for autonomous robotics; critical structures, such as nerves or vessels, should be localized more accurately compared to fatty tissue.*
2.5.2 Unequal severity of class confusions a) for detection b) for segmentation (per instance)	Any class can be confused with another, but certain mismatches are more severe than others, from a domain point of view. *Example (binary): polyp detection; a FN (missed polyp) is clinically much more severe than a FP.* Specifically in instance segmentation problems, the property needs to be set separately for the validation of the (a) detection (relevant decision guide: 4.4) and (b) segmentation performance (relevant subprocesses: 7, 8). At object level, FNs (missed instances) are sometimes more severe than FPs, while FNs (e.g. undersegmentation) and FPs (e.g. oversegmentation) may be equally important at pixel level.
2.5.3 Compensation for class imbalances requested	Severe class imbalances might impede interpretability and objective assessment of method validation and e.g. lead to overly optimistic conclusions. Some metrics compensate for such effects and thus enable unbiased interpretability and objective assessment. *Example (binary classification with the negative class being overrepresented in the form of "easy to classify" TN): Metric scores considering balanced discrimination of two classes (e.g. LR+ or AUROC) might be dominated by TN and thus give an overly optimistic picture of the performance, especially if practical interest lies with the positive class, such as in retrieval tasks or particular diagnostic tasks (e.g. cancer detection out of a cohort with mostly healthy patients). Metrics not considering TN (e.g. F_β Score or AP) compensate for this effect and enable focusing on the discrimination of the positive class.*
2.5.4 Handling of spatial outliers	Spatial outliers are FP predictions that feature a large distance to the reference. They can be handled in three different ways: Distance-based penalization with outlier focus: Outliers should be heavily penalized as a function of the distance to the reference. Distance-based penalization with contour focus: Outliers should be penalized as a function of the distance to the reference, but the assessment should focus on the general contour agreement rather than individual outliers. Existence-based penalization: The existence of spatial outliers should be penalized irrespective of their distance to the reference. Distance-based penalization is not possible when either the reference or the prediction is empty. In applications in which many such cases potentially occur, we therefore recommend an existence-based penalization.

Fig. 17. **Fingerprint for instance segmentation (Part 1).** In the case of binary fingerprints, the blue column shows examples for which the property is true while the red column shows counterexamples. Categorical fingerprints are only shown in blue.



INSTANCE SEGMENTATION (IS) PART 2	
Domain interest-related properties (part 2)	
2.5.5 Compensation for annotation imprecisions requested	The reference annotation is typically only an approximation of the (forever auncertainties, such as intra-rater or inter-rater variability, by configuring the metric accordingly. This is only possible for some metrics.
2.5.6 Penalization of multiple predictions assigned to the same reference object requested	Instance segmentation algorithms often involve the step of assigning predicted objects to reference objects. This may result in more than one prediction being assigned to the same reference. This fingerprint property should be set to true if all but one prediction of such an assignment should be penalized as FP, and set to false if these spare predictions should be ignored during validation.
2.5.7 Penalization of multiple reference objects assigned to same prediction requested	Instance segmentation algorithms often involve the step of assigning reference objects to predicted objects. This may result in more than one reference being assigned to the same prediction. If such an assignment should be penalized, this fingerprint property should be set to true.
2.6 Cutoff on predicted class scores Options: - Target value-based - Optimization-based - Argmax-based - No cutoff	Modern algorithms output continuous class scores. Making a classification decision requires setting a cutoff value on the scores, thereby generating a (cutoff-specific) confusion matrix. This matrix enables the computation of popular single-threshold counting metrics, such as sensitivity, PPV and F_1 Score. Depending on domain interest the cutoff can be set in multiple ways: Target value-based: The cutoff represents the threshold for which a specific target metric value (e.g. Sensitivity = 0.95) is achieved. Other metric values (e.g. Specificity) are then reported for this specific threshold. Optimization-based: The cutoff is inferred by optimizing a target metric, such as the F_1 Score, on a dedicated data set provided for hyperparameter tuning. Argmax-based: If no target value is defined, no separate data split for optimization is available, or there are concerns w.r.t generalization of data-based cutoff optimization, a common option is to pick the hypothesis that is most probable (this strategy is also referred to as argmax and is the principle behind a Bayes classifier). No cutoff: Examples for no interest in validating a method at a certain cutoff are (1) focus on general methodological performance across many tasks and data sets without application interest, or (2) concerns regarding the comparability of results based on a single cutoff that is fixed across varying study cohorts (see also F4.2).
2.7 Interpretability of predicted class scores	When validating classification methods - particularly those with applications that involve direct human read-out - it is often crucial for the predicted class scores themselves to be interpretable. This property should be set to true if the predicted class scores should match the true probability of interest (e.g. the probability of a patient to develop a certain disease in prediction problems).
Target structure-related properties	
3.1 Small size of structures relative to pixel size	Structures of the provided class are consistently small relative to the grid size in such a way that a single pixel makes up at least several percentage points of the structure volume. *Example: multiple sclerosis lesions in magnetic resonance imaging (MRI) scans.*
3.2 High variability of structure sizes (within one image, across images)	The target structures vary substantially in size, such that some structures are several times the size of others. *Example: polyps in colonoscopy screening, where some polyps are several times the size of others.* *Counterexample: large organs, such as the liver or the kidneys, which are relatively comparable in size across individuals.*
3.3 Target structures feature tubular shape	The target structures feature a tubular shape. *Examples: vessels, neurons, microtubules.*
3.4 Possibility of multiple labels per unit (pixel or image)	Multiple categories may be assigned to one pixel. *Example: labels 'tumor core' and 'tumor' assigned to the same pixel.*
3.5 Possibility of overlapping or touching target structures (e.g. medical instruments or cells)	Different instances of a class can overlap or touch each other. *Examples: overlapping cells or organisms; overlapping medical instruments in laparoscopy.*
3.6 Possibility of disconnected target structure(s)	A given structure appears disconnected in the given image. *Examples: neurons in 2D microscopy of a slice of tissue; single tomographic image slice depicting complex vessels.*

Fig. 18. **Fingerprint for instance segmentation (Part 2).** In the case of binary fingerprints, the blue column shows examples for which the property is true while the red column shows counterexamples. Categorical fingerprints are only shown in blue.



INSTANCE SEGMENTATION (IS) PART 3

Data set-related properties

Property	Description
4.1 Presence of class imbalance	The class prevalences differ substantially. *Example: In a screening application, the positive class (e.g. cancer) may occur extremely rarely. In this case, prevalence-dependent metrics, such as Accuracy, may be extremely misleading.*
4.3 Uncertainties in the reference	The reference is typically only an approximation of the (forever unknown) ground truth. Various sources and types of errors exist, which require special treatment in the context of metric selection.
4.3.1 High inter-/intra-rater variability	The reference can be assumed to be noisy due to high inter-rater variability.
4.3.2 Possibility of spatial outliers in reference annotation	The reference may feature spatial outliers that are distant from the (unknown) ground truth.
4.5 Non-independence of test cases	The test cases are hierarchically structured, indicating non-independence of test cases. *Examples: multiple images of the same patient, hospital or video.*
4.6 Possibility of reference without target structure(s)	There are test cases in which the reference comprises only the background class.

Algorithm output-related properties

Property	Description
5.1 Availability of predicted class scores	Modern algorithms in biomedical image classification output continuous class scores, which are often interpreted as predicted class probabilities. These scores contain relevant information about the performance of a model and are thus crucial for comprehensive and meaningful validation. Instance segmentation problems in the biomedical domain are often approached by adding a post-processing step (e.g. connected component analysis) to a semantic segmentation algorithm. In this process, predicted class probabilities often get lost. If no predicted class probabilities are available, this property is set to false.
5.2 Possibility of algorithm output not containing the target structure(s)	The algorithm may yield output images only comprising the background class.
5.3 Possibility of invalid algorithm output (e.g. Prediction is NaN)	The files representing the algorithm output can contain invalid output. Note that an invalid prediction differs from an empty prediction.
5.4 Possibility of overlapping predictions	Predictions of the algorithm can potentially overlap.

Fig. 19. **Fingerprint for instance segmentation (Part 3).** In the case of binary fingerprints, the blue column shows examples for which the property is true while the red column shows counterexamples. Categorical fingerprints are only shown in blue.

E Metric Selection

Based on the problem fingerprinting (App. D), the user is guided through the process of selecting an appropriate set validation metrics from different families while being made aware of potential pitfalls related to individual choices. As a foundation for this process, the *Metrics Reloaded* consortium compiled a set of common reference-based validation metrics that are recommended (Tab. 2). A comprehensive introduction and discussion of the individual metrics can be found in the sister publication of this work [72]. Here, we focus on which metric(s) to recommend under which circumstances. The Results section summarized the core of our recommendations. The following Figures detail the subprocesses that are referenced in Fig. 2 of the main paper.

Table 2. **Overview of recommended reference-based metrics.** For each metric, a name, acronym, synonyms, reference to the definition and illustration, range and corresponding problem categories are provided. The direction of each arrow in the column *range* indicates whether higher (up) or lower scores (down) are better. A detailed introduction and discussion of the metrics can be found in the sister publication of this work [72]. ILS: image-level classification; SS: semantic segmentation; OD: object detection; IS: instance segmentation.

Metric	Acronym	Synonyms	Definition	Range	Recommended for			
					ILC	SS	OD	IS
Counting Metrics								
Accuracy			[37, 85]	[0, 1] ↑	x			
Balanced Accuracy	BA		[37, 85]	[0, 1] ↑	x			
Weighted Cohen's Kappa	WCK	Cohen's Kappa Coefficient, Kappa Statistic, Kappa Score	[28]	[-1, 1] ↑	x			
Center line Dice Similarity Coefficient	clDice		[79]	[0, 1] ↑		x		x
Dice Similarity Coefficient	DSC	Sørensen–Dice Coefficient, F_1 Score, Balanced F Score	[32]	[0, 1] ↑		x		x
Expected Cost	EC		[14]	(-∞, ∞) ↓	x			
F_β Score			[24]	[0, 1] ↑	x	x	x	x
False Positives per Image*	FPPI		[8, 89]	[0, ∞) ↓			x	x
Intersection over Union	IoU	Jaccard Index, Tanimoto Coefficient	[44]	[0, 1] ↑		x		x
Matthews Correlation Coefficient	MCC	Phi Coefficient	[63]	[-1, 1] ↑	x			
Panoptic Quality	PQ		[51]	[0, 1] ↑				x
Net Benefit	NB		[90]	(-∞,∞) ↑	x			
Negative Predictive Value*			[13, 85]	[0, 1] ↑	x			
Positive Likelihood Ratio	LR+	Likelihood Ratio Positive, Likelihood Ratio for Positive Results	[6]	[0, ∞) ↑	x			
Positive Predictive Value*	PPV	Precision	[13, 37, 85]	[0, 1] ↑	x		x	x
Sensitivity*		Recall, Hit Rate, True Positive Rate (TPR)	[13, 37, 85]	[0, 1] ↑	x		x	x
Specificity*		Selectivity, True Negative Rate (TNR)	[13, 37, 85]	[0, 1] ↑	x			
Multi-threshold Metrics								
Area under the Receiver Operating Characteristic Curve	AUROC	Area under the curve (AUC), AUC Receiver Operating Characteristic (ROC), C-Index, C-Statistics	[39]	[0, 1] ↑	x			
Average Precision	AP		[57]	[0, 1] ↑	x		x	x
Free-Response Receiver Operating Characteristic Score	FROC Score		[8, 89]	[0, 1] ↑			x	x
Distance-based Metrics								
Average Symmetric Surface Distance	ASSD		[94]	[0, ∞) ↓		x		x
Boundary Intersection over Union	Boundary IoU		[21]	[0, 1] ↑		x		x
Hausdorff Distance	HD	Hausdorff Metric, Pompeiu–Hausdorff Distance, Maximum Symmetric Surface Distance	[42]	[0, ∞) ↓		x		x
Mean Average Surface Distance	MASD		[11]	[0, ∞) ↓		x		x
Normalized Surface Distance	NSD	Normalized Surface Dice, Surface Distance, Surface Dice	[70]	[0, 1] ↑		x		x
X^{th} Percentile Hausdorff Distance	X^{th} Percentile HD		[42]	[0, ∞) ↓		x		x
Calibration Metrics								
Proper Scoring Rules	PSR		[35]	varying**	x		x	x
Expected Calibration Error	ECE		[38]	[0, 1] ↑	x		x	x

*: This metric is best used in combination with another metric using a predefined target value (see "Target value-based cutoff" in the definition of *F2.6: Cutoff on predicted class scores* (App. D).

**: The range depends on the chosen Proper Scoring Rule.

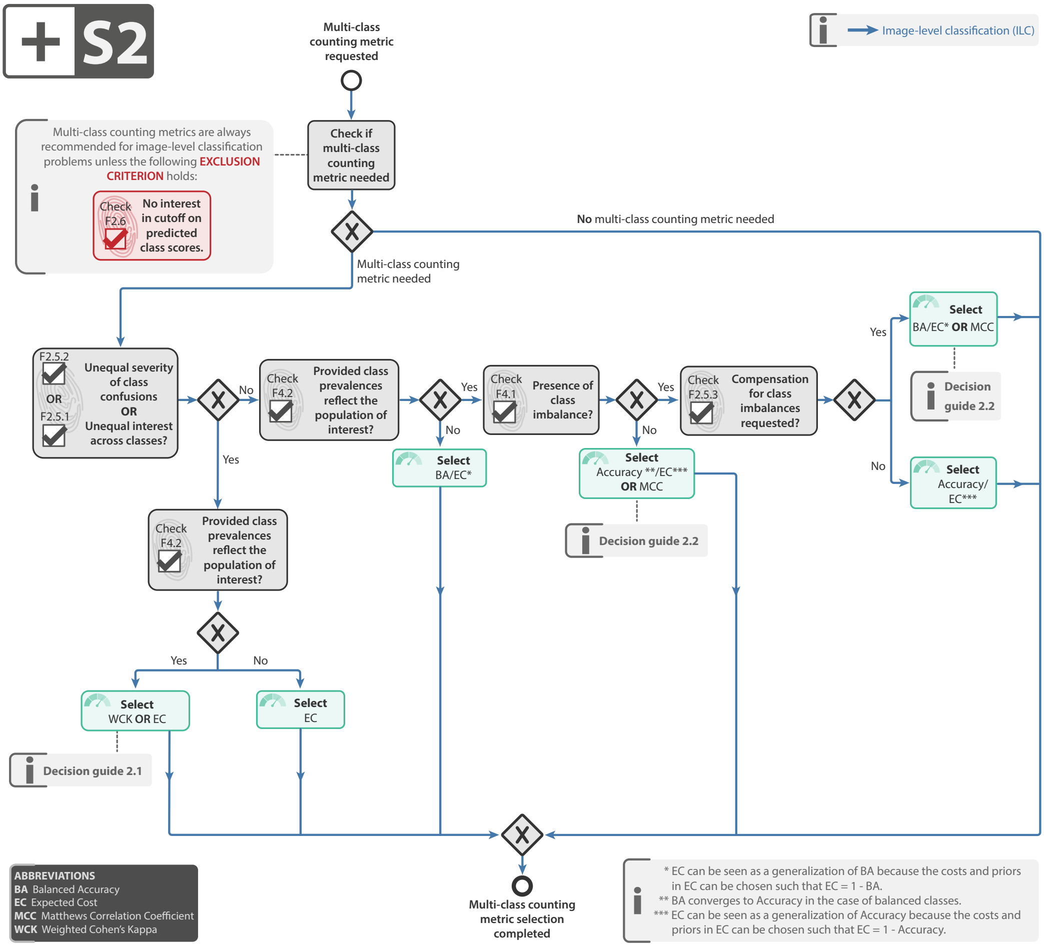

S2
Image-level classification (ILC)
Multi-class counting metric requested
Check if multi-class counting metric needed
Multi-class counting metrics are always recommended for image-level classification problems unless the following EXCLUSION CRITERION holds:
Check F2.6 No interest in cutoff on predicted class scores.
No multi-class counting metric needed
Multi-class counting metric needed
F2.5.2 Unequal severity of class confusions
OR
F2.5.1 Unequal interest across classes?
No
Yes
Check F4.2 Provided class prevalences reflect the population of interest?
Yes
No
Check F4.1 Presence of class imbalance?
Yes
No
Check F2.5.3 Compensation for class imbalances requested?
Yes
No
Select BA/EC*
Select Accuracy **/EC*** OR MCC
Decision guide 2.2
Select BA/EC* OR MCC
Decision guide 2.2
Select Accuracy/ EC***
Check F4.2 Provided class prevalences reflect the population of interest?
Yes
No
Select WCK OR EC
Decision guide 2.1
Select EC
Multi-class counting metric selection completed
ABBREVIATIONS
BA Balanced Accuracy
EC Expected Cost
MCC Matthews Correlation Coefficient
WCK Weighted Cohen's Kappa
* EC can be seen as a generalization of BA because the costs and priors in EC can be chosen such that EC = 1 - BA.
** BA converges to Accuracy in the case of balanced classes.
*** EC can be seen as a generalization of Accuracy because the costs and priors in EC can be chosen such that EC = 1 - Accuracy.

Fig. 20. **Subprocess S2 for selecting multi-class metrics (if any)**. Applies to: image-level classification (ILC). In the case of presence of class imbalance and no compensation of class imbalance being requested, one should follow the 'no' branch. Decision guides are provided in App. F.2.

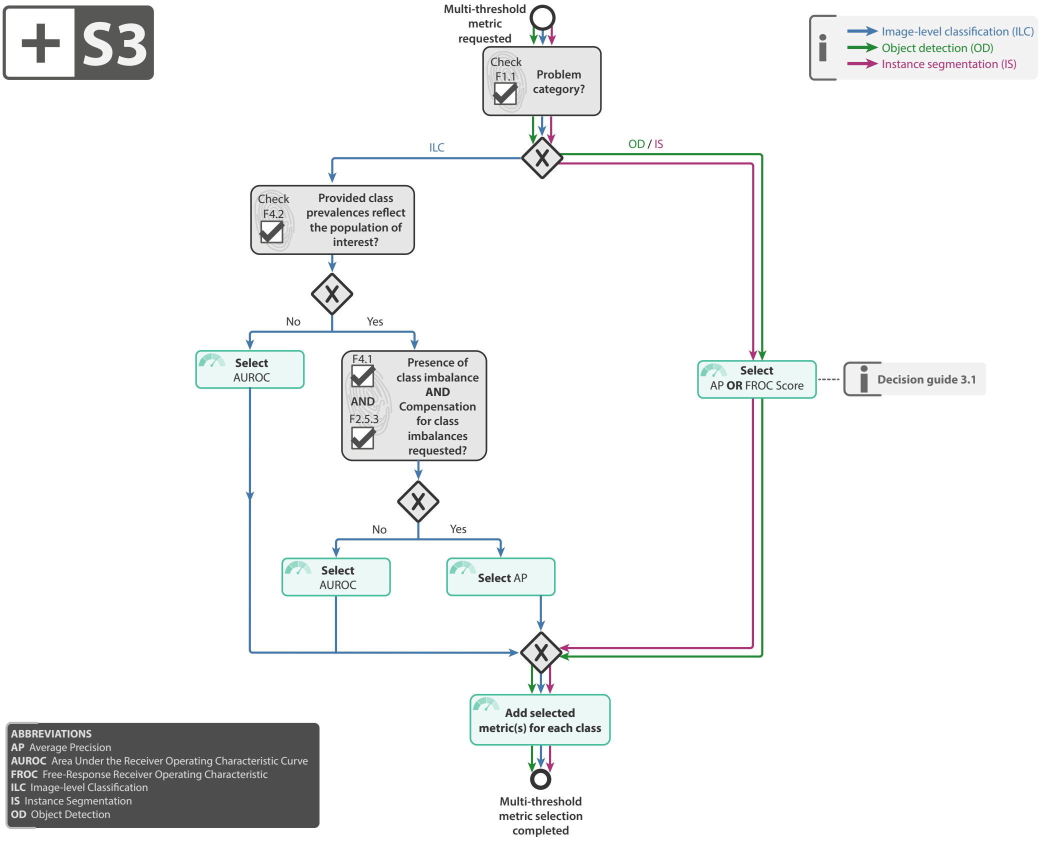

+ S3
Multi-threshold metric requested
Check F1.1
Problem category?
Image-level classification (ILC)
Object detection (OD)
Instance segmentation (IS)
ILC
OD / IS
Check F4.2
Provided class prevalences reflect the population of interest?
No
Yes
Select AUROC
F4.1
AND
F2.5.3
Presence of class imbalance AND Compensation for class imbalances requested?
Select AP OR FROC Score
Decision guide 3.1
No
Yes
Select AUROC
Select AP
Add selected metric(s) for each class
Multi-threshold metric selection completed
ABBREVIATIONS
AP Average Precision
AUROC Area Under the Receiver Operating Characteristic Curve
FROC Free-Response Receiver Operating Characteristic
ILC Image-level Classification
IS Instance Segmentation
OD Object Detection

Fig. 21. **Subprocess S3 for selecting a multi-threshold metric (if any)**. Applies to: image-level classification (ILC), object detection (OD) and instance segmentation (IS). Decision guides are provided in App. F.3.

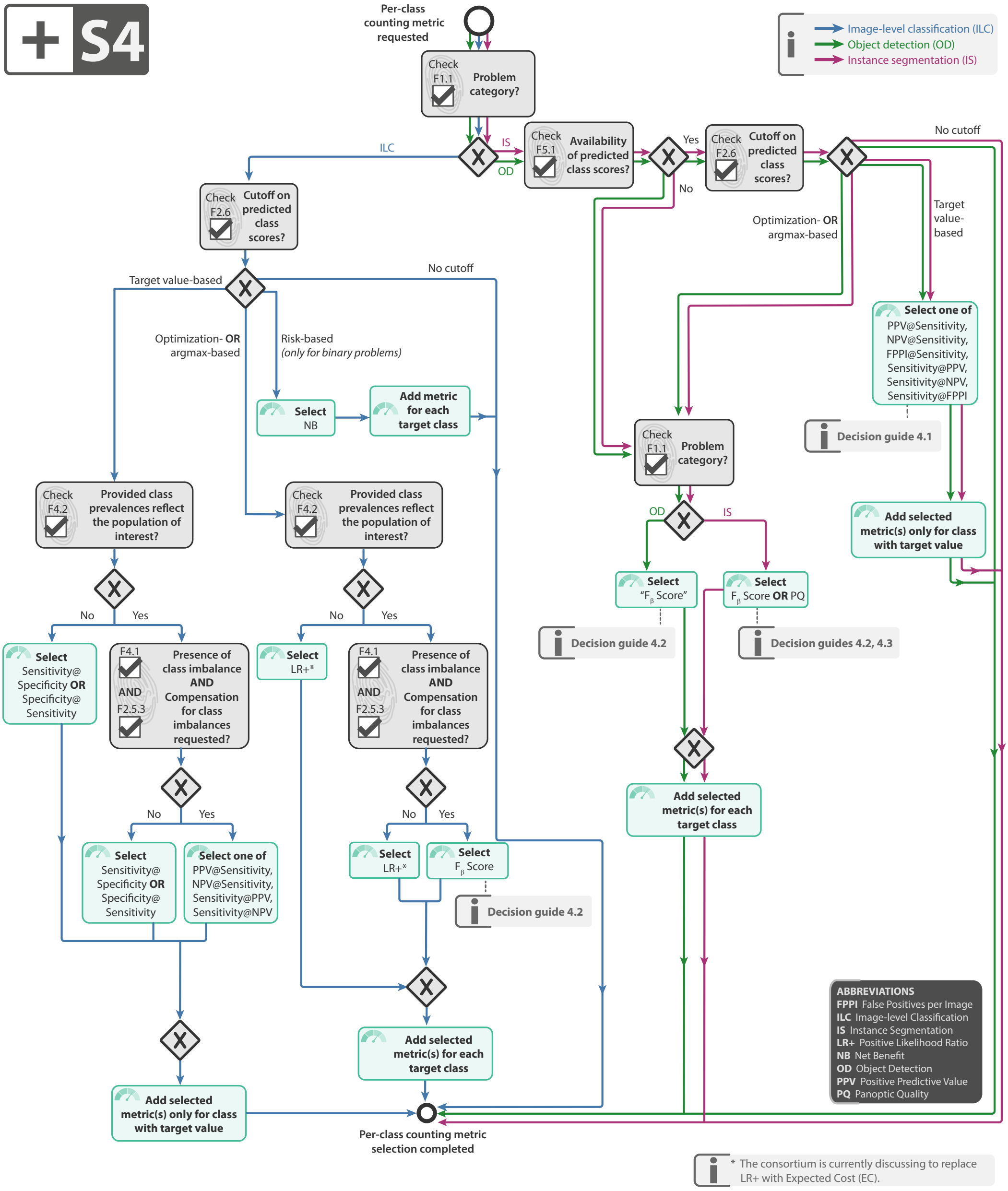

+ S4
Per-class counting metric requested
Image-level classification (ILC)
Object detection (OD)
Instance segmentation (IS)
Check F1.1 Problem category?
ILC
IS
OD
Check F5.1 Availability of predicted class scores?
Yes
No
Check F2.6 Cutoff on predicted class scores?
No cutoff
Target value-based
Optimization- OR argmax-based
Select one of PPV@Sensitivity, NPV@Sensitivity, FPPI@Sensitivity, Sensitivity@PPV, Sensitivity@NPV, Sensitivity@FPPI
Decision guide 4.1
Add selected metric(s) only for class with target value
Check F1.1 Problem category?
OD
IS
Select "F_β Score"
Select F_β Score OR PQ
Decision guide 4.2
Decision guides 4.2, 4.3
Add selected metric(s) for each target class
Check F2.6 Cutoff on predicted class scores?
Target value-based
No cutoff
Optimization- OR argmax-based
Risk-based (only for binary problems)
Select NB
Add metric for each target class
Check F4.2 Provided class prevalences reflect the population of interest?
No
Yes
Select Sensitivity@ Specificity OR Specificity@ Sensitivity
F4.1 AND F2.5.3 Presence of class imbalance AND Compensation for class imbalances requested?
No
Yes
Select Sensitivity@ Specificity OR Specificity@ Sensitivity
Select one of PPV@Sensitivity, NPV@Sensitivity, Sensitivity@PPV, Sensitivity@NPV
Add selected metric(s) only for class with target value
Check F4.2 Provided class prevalences reflect the population of interest?
No
Yes
Select LR+*
F4.1 AND F2.5.3 Presence of class imbalance AND Compensation for class imbalances requested?
No
Yes
Select LR+*
Select F_β Score
Decision guide 4.2
Add selected metric(s) for each target class
Per-class counting metric selection completed
ABBREVIATIONS
FPPI False Positives per Image
ILC Image-level Classification
IS Instance Segmentation
LR+ Positive Likelihood Ratio
NB Net Benefit
OD Object Detection
PPV Positive Predictive Value
PQ Panoptic Quality
* The consortium is currently discussing to replace LR+ with Expected Cost (EC).

Fig. 22. **Subprocess S4 for selecting a per-class counting metric (if any)**. Applies to: image-level classification (ILC), object detection (OD) and instance segmentation (IS). Decision guides are provided in App. F.4.

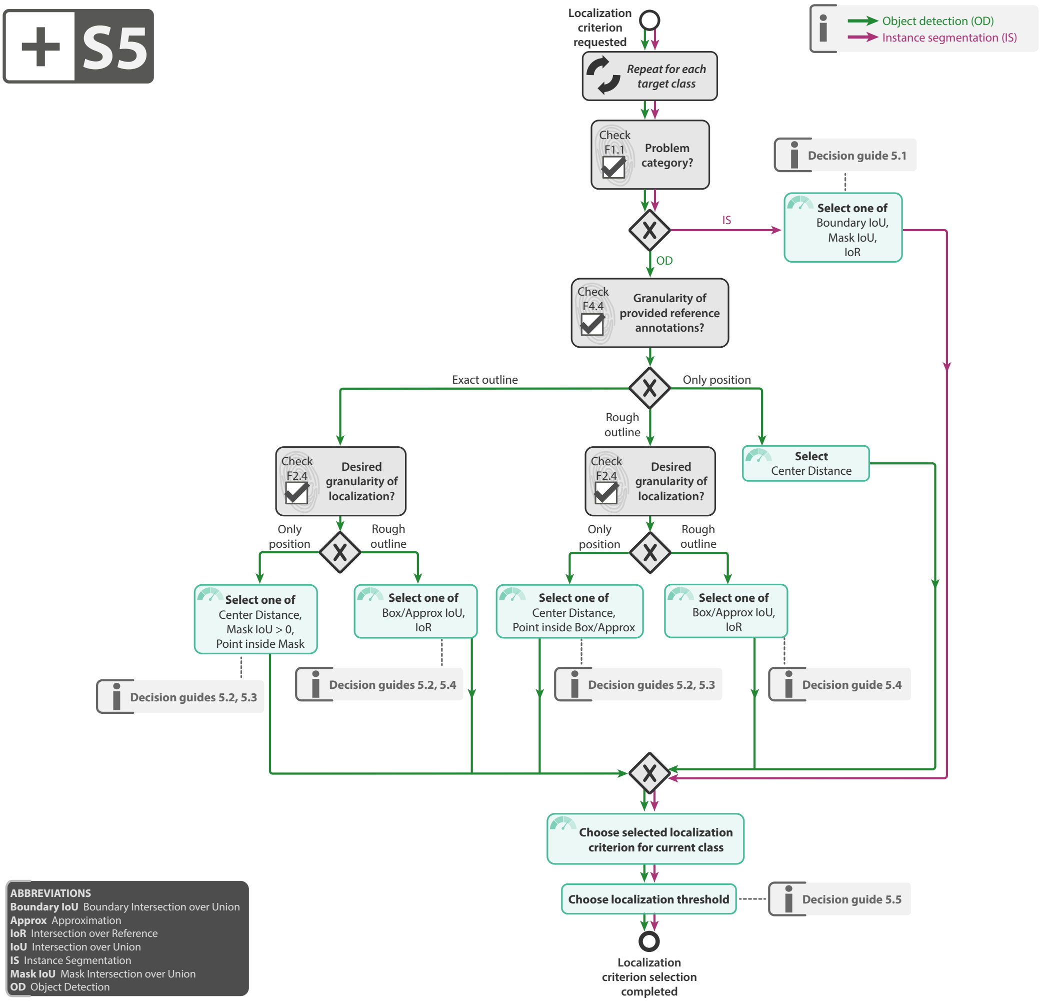

+ S5
Localization criterion requested
Object detection (OD)
Instance segmentation (IS)
Repeat for each target class
Check F1.1
Problem category?
Decision guide 5.1
IS
Select one of Boundary IoU, Mask IoU, IoR
OD
Check F4.4
Granularity of provided reference annotations?
Exact outline
Only position
Rough outline
Select Center Distance
Check F2.4
Desired granularity of localization?
Only position
Rough outline
Check F2.4
Desired granularity of localization?
Only position
Rough outline
Select one of Center Distance, Mask IoU > 0, Point inside Mask
Select one of Box/Approx IoU, IoR
Select one of Center Distance, Point inside Box/Approx
Select one of Box/Approx IoU, IoR
Decision guides 5.2, 5.3
Decision guides 5.2, 5.4
Decision guides 5.2, 5.3
Decision guide 5.4
Choose selected localization criterion for current class
Choose localization threshold
Decision guide 5.5
Localization criterion selection completed
ABBREVIATIONS
Boundary IoU Boundary Intersection over Union
Approx Approximation
IoR Intersection over Reference
IoU Intersection over Union
IS Instance Segmentation
Mask IoU Mask Intersection over Union
OD Object Detection

Fig. 23. **Subprocess S5 for selecting the localization criterion**. Applies to: object detection (OD) and instance segmentation (IS). Definitions of the localization criteria can be found in [72]. Decision guides are provided in App. F.5.

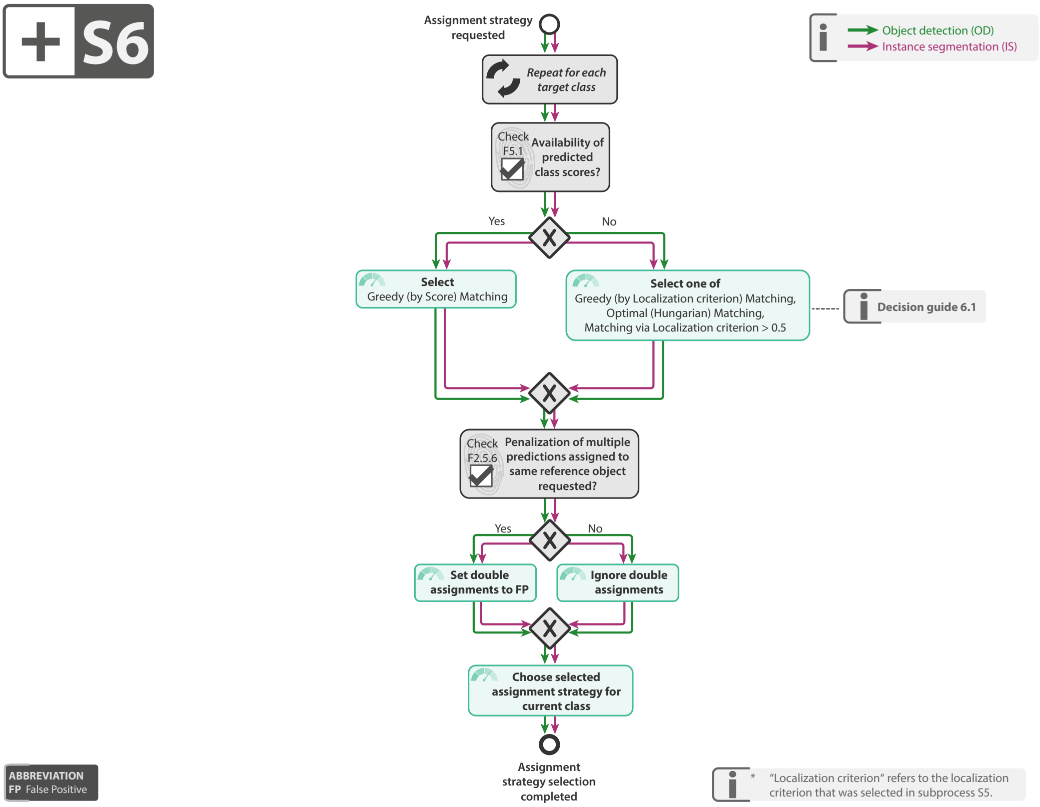
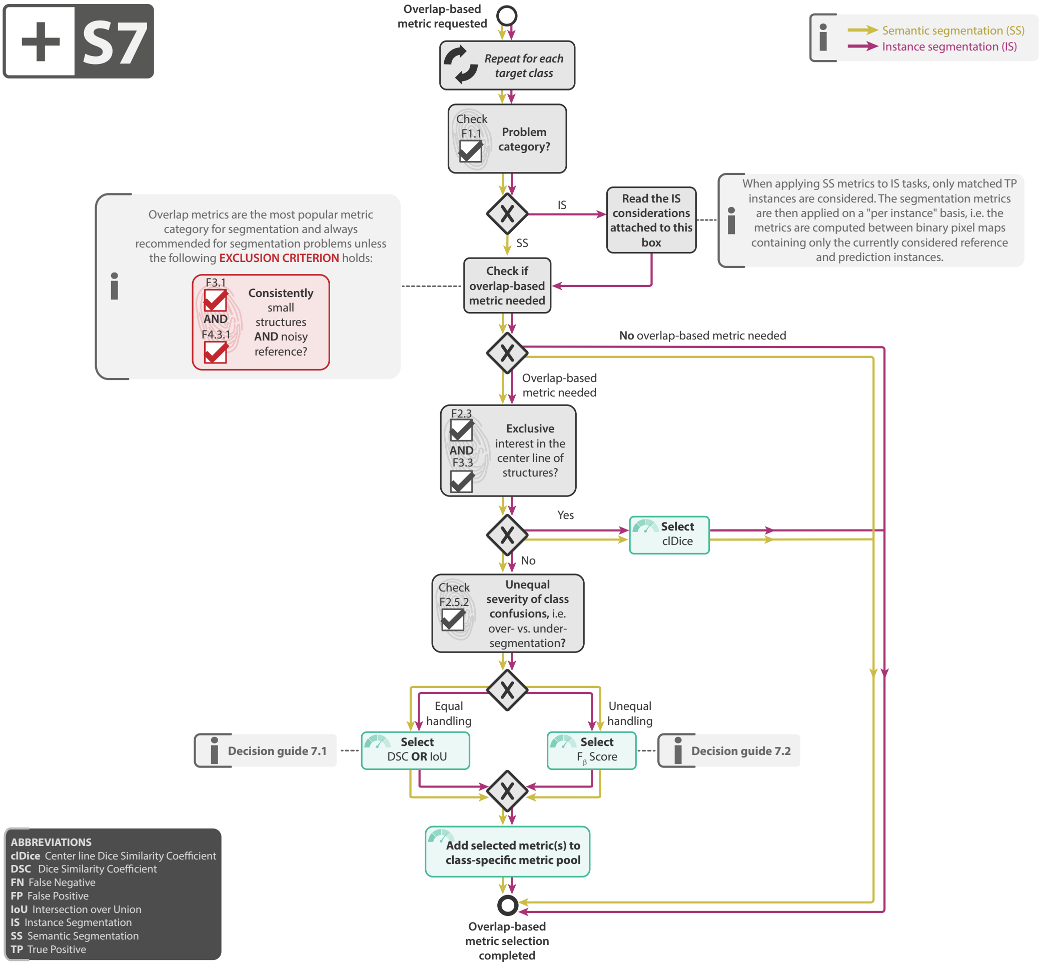

S6
Assignment strategy requested
Object detection (OD)
Instance segmentation (IS)
Repeat for each target class
Check F5.1
Availability of predicted class scores?
Yes
No
Select
Greedy (by Score) Matching
Select one of
Greedy (by Localization criterion) Matching,
Optimal (Hungarian) Matching,
Matching via Localization criterion > 0.5
Decision guide 6.1
Check F2.5.6
Penalization of multiple predictions assigned to same reference object requested?
Yes
No
Set double assignments to FP
Ignore double assignments
Choose selected assignment strategy for current class
Assignment strategy selection completed
ABBREVIATION
FP False Positive
* "Localization criterion" refers to the localization criterion that was selected in subprocess S5.

Fig. 24. **Subprocess S6 for selecting the assignment strategy.** Applies to: object detection (OD) and instance segmentation (IS). Assignment strategies are defined in [72]. Decision guides are provided in App. F.6.

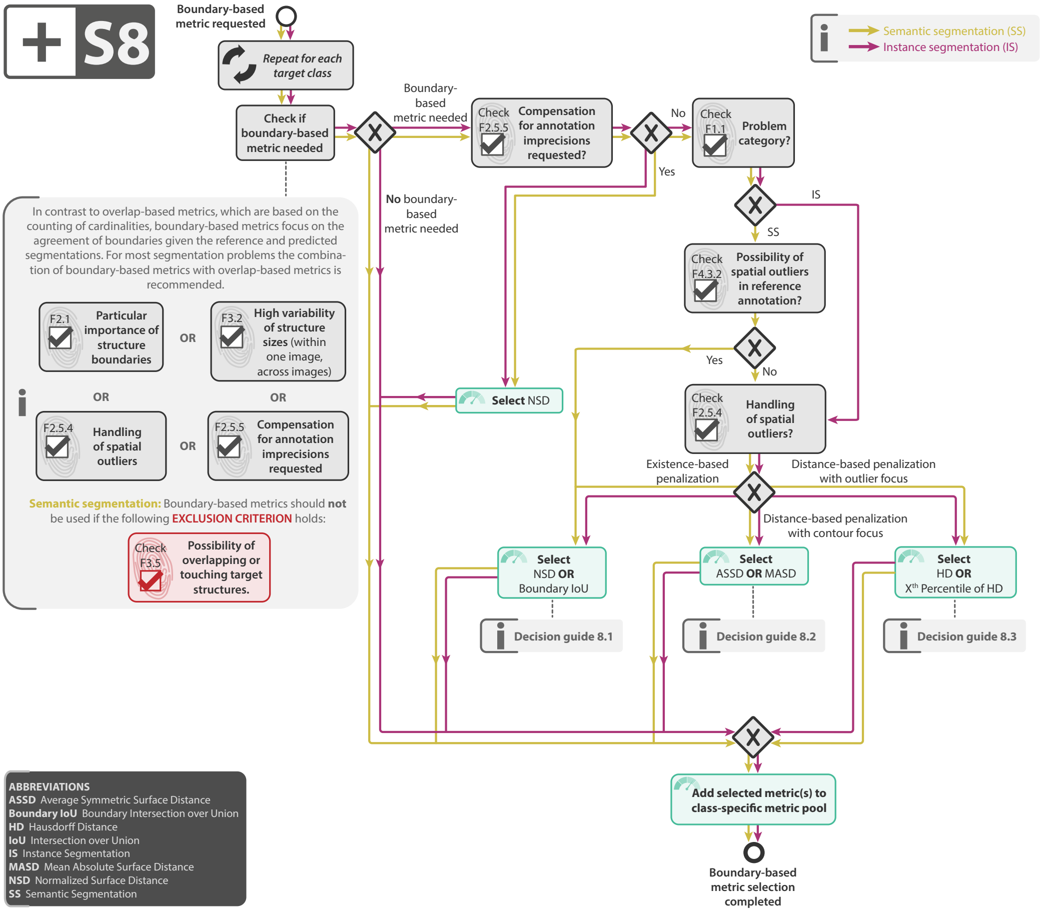

S7
Overlap-based metric requested
Repeat for each target class
Check F1.1 Problem category?
IS
SS
Read the IS considerations attached to this box
Check if overlap-based metric needed
F3.1
AND
F4.3.1
Consistently small structures AND noisy reference?
No overlap-based metric needed
Overlap-based metric needed
F2.3
AND
F3.3
Exclusive interest in the center line of structures?
Yes
Select clDice
No
Check F2.5.2 Unequal severity of class confusions, i.e. over- vs. under-segmentation?
Equal handling
Unequal handling
Decision guide 7.1
Select DSC OR IoU
Select F_β Score
Decision guide 7.2
Add selected metric(s) to class-specific metric pool
Overlap-based metric selection completed
Semantic segmentation (SS)
Instance segmentation (IS)

When applying SS metrics to IS tasks, only matched TP instances are considered. The segmentation metrics are then applied on a "per instance" basis, i.e. the metrics are computed between binary pixel maps containing only the currently considered reference and prediction instances.

Overlap metrics are the most popular metric category for segmentation and always recommended for segmentation problems unless the following **EXCLUSION CRITERION** holds:

ABBREVIATIONS
clDice Center line Dice Similarity Coefficient
DSC Dice Similarity Coefficient
FN False Negative
FP False Positive
IoU Intersection over Union
IS Instance Segmentation
SS Semantic Segmentation
TP True Positive

Fig. 25. **Subprocess S7 for selecting overlap-based segmentation metrics (if any)**. Applies to: semantic segmentation (SS) and instance segmentation (IS). Decision guides are provided in App. F.7.


S8
Boundary-based metric requested
Repeat for each target class
Check if boundary-based metric needed
Boundary-based metric needed
No boundary-based metric needed
Check F2.5.5 Compensation for annotation imprecisions requested?
No
Yes
Check F1.1 Problem category?
IS
SS
Check F4.3.2 Possibility of spatial outliers in reference annotation?
Yes
No
Check F2.5.4 Handling of spatial outliers?
Existence-based penalization
Distance-based penalization with outlier focus
Distance-based penalization with contour focus
Select NSD
Select NSD OR Boundary IoU
Select ASSD OR MASD
Select HD OR X^th Percentile of HD
Decision guide 8.1
Decision guide 8.2
Decision guide 8.3
Add selected metric(s) to class-specific metric pool
Boundary-based metric selection completed
Semantic segmentation (SS)
Instance segmentation (IS)
F2.1 Particular importance of structure boundaries
OR
F3.2 High variability of structure sizes (within one image, across images)
OR
OR
F2.5.4 Handling of spatial outliers
OR
F2.5.5 Compensation for annotation imprecisions requested
Check F3.5 Possibility of overlapping or touching target structures.

In contrast to overlap-based metrics, which are based on the counting of cardinalities, boundary-based metrics focus on the agreement of boundaries given the reference and predicted segmentations. For most segmentation problems the combination of boundary-based metrics with overlap-based metrics is recommended.

Semantic segmentation: Boundary-based metrics should **not** be used if the following **EXCLUSION CRITERION** holds:

ABBREVIATIONS
ASSD Average Symmetric Surface Distance
Boundary IoU Boundary Intersection over Union
HD Hausdorff Distance
IoU Intersection over Union
IS Instance Segmentation
MASD Mean Absolute Surface Distance
NSD Normalized Surface Distance
SS Semantic Segmentation

Fig. 26. **Subprocess S8 for selecting a boundary-based segmentation metric (if any)**. Applies to: semantic segmentation (SS) and instance segmentation (IS). Decision guides are provided in App. F.8.

F Decision Guides

F.1 Decision guide Overview Mapping.

DM.1: Select calibration metric

There are two main concepts for measuring whether predicted class scores of a classifier correctly express the probability of class membership.

- **Expected Calibration Error**: A *calibrated model* outputs predicted class scores that match the empirical success rate (e.g. outputs with score 0.8 for a specific class empirically belong to this class in 80% of the cases). Measuring whether a model is calibrated is commonly done by dividing the scale of class scores into bins and subsequently (1) plotting the empirical success rate over bins for visual assessment (see [38]) and (2) computing the expected calibration error (ECE), i.e. the calibration error per bin averaged over bins. Crucially, calibration should generally not be assessed without validating a model's discrimination ability in parallel. This is because a model might be perfectly calibrated but provide no discrimination (e.g. always output score 0.5 on a dataset with two balanced classes), which reduces the value of the model to stating the class prevalence of the data set (i.e. the expressed class membership probability per case is correct but not useful).
- **Proper Scoring Rules**: PSR allow to validate discrimination and calibration in a single score. Specifically, these metrics require predicted class scores to match the true posterior probability in each individual case. Example metrics are the Brier score (BS) and the Logarithmic Score (LS)).

Calibration and class prevalence: Crucially, calibration errors as well as PSR are generally prevalence-dependent metrics. Thus, if the prevalence of the data set does not represent the population of interest (see F4.2), the quality of the risk expressed by predicted class scores is not comparable across data sets and needs to be re-validated on each new study cohort (see Fig. 27).

Pitfall: Prevalence-dependent metrics cannot be compared across data sets with different prevalences

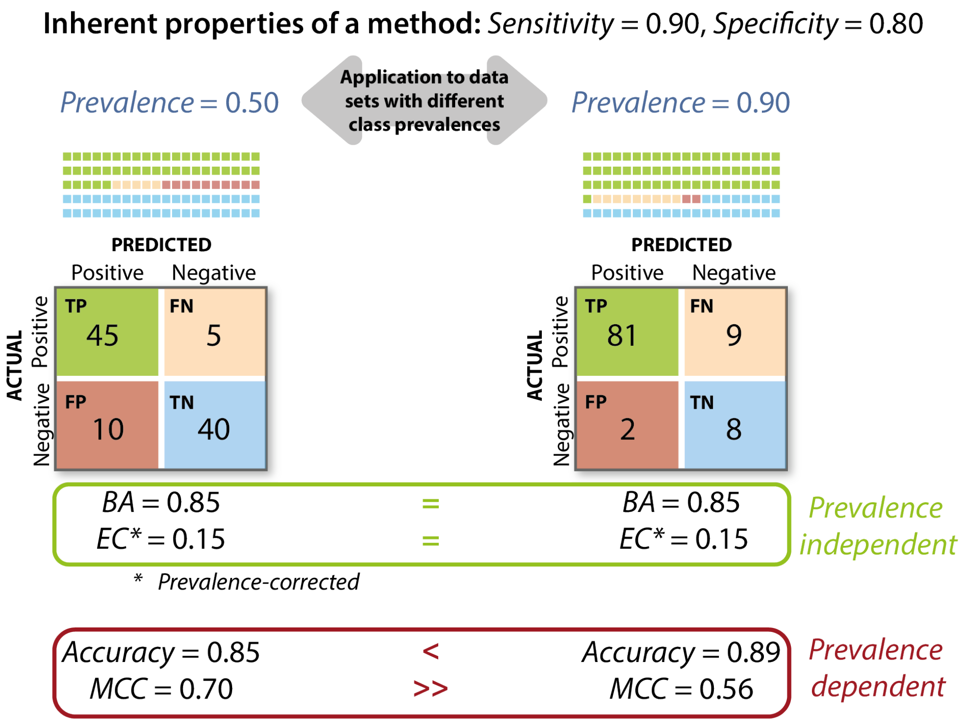


Fig. 27. Effect of prevalence dependency. An algorithm with specific inherent properties (here: Sensitivity of 0.9 and Specificity of 0.8), may perform completely differently on different data sets if the prevalences differ (here: 50% (left) and 90% (right)) and prevalence-dependent metrics are used for validation (here: Accuracy and Matthews Correlation Coefficient (MCC)). In contrast, prevalence-independent metrics (here: Balanced Accuracy (BA) and the prevalence-corrected Expected Cost (EC)) can be used to compare validation results across different data sets. Used abbreviations: True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN).

F.2 Decision guide S2.

D2.1: Weighted Cohen's Kappa (WCK) versus EC

The following aspects should be taken into account when deciding between WCK and EC: Both metrics allow for incorporating task-specific penalties for confusions between individual pairs of classes. Common use cases for this property are tasks with ordinal classes or diagnostic decisions with errors of varying clinical severity. However, there are substantial differences between the two metrics:

- **Symmetry**: Kappa statistics in general and WCK in particular were originally proposed to compare annotations/guesses of two raters on an ordinal scale, which is a symmetric problem by nature. Application of such measures to benchmarking studies is hence problematic because these involve the comparison of a prediction to a reference, which is considered to be/approximate the truth (asymmetric setting). Therefore, WCK does conceptually not match the comparison we are attempting to make and could thus lead to misleading results. EC does not suffer from this problem.
- **Interpretability**: WCK is arguably hard to interpret. The fact that the "agreement by chance" is subtracted to account for class imbalances seems intuitive, but the behaviour of resulting scores is not independent of class imbalances (leading to generally higher scores for more balanced data). EC, on the other hand, constitutes an intuitive extension of Accuracy where customized penalties for individual class confusions can be integrated. Further, class priors (i.e. expected class prevalences in the target population) in EC can be adapted if known [56].
- **Undesired behaviour in practice**: Using WCK with quadratic weights, as often done for ordinal tasks, has been found to lead to "paradoxical results" [92].
- **Popularity**: WCK is often used in the biomedical domain, whenever customized penalties for class confusions are required. EC, on the other hand, is currently mostly found either in statistical text books or in non-related domains such as speech recognition.
- **Theoretical foundation**: EC comes with a comprehensive theoretical foundation based on Bayesian decision theory. As a consequence, it is possible to analytically derive the optimal cutoff on the predicted class scores. As this requires calibrated scores (i.e. the scores are expected to represent posterior probabilities of the class given the data), any performance mismatch between the theoretical threshold and an empirical threshold on the final test data can be interpreted as a calibration error.

D2.2: Accuracy and BA versus MCC

Accuracy and BA can both be seen as an instantiation of EC. Accuracy, in turn, is identical to BA in the case of balanced classes (F4.1 = FALSE). In the provided decision tree (Subprocess S2, Fig. 20), the decisions to be made thus primarily boil down to deciding between BA and MCC. In this context, the following aspects should be taken into account:

- **Commonalities:** It is well-known that some metrics, such as Accuracy, can yield an unreasonably high score for an uninformed classifier when class imbalance is high [72]. BA and MCC both compensate for this effect by re-scaling metric scores of an uninformed classifier to fixed values (BA: 1 divided by the number of classes, MCC: 0) irrespective of the class imbalance, thereby enhancing interpretability. Accuracy, in turn, is identical to BA in the case of balanced classes. Note that we recommend Accuracy only if there is no class balance (F4.1 = FALSE) or the class imbalance should not be compensated for (F2.5.3 = FALSE).
- **Prevalence dependency:** A main difference between BA and MCC is the integration of the predictive values from the confusion matrix: While BA only averages the recalls for each class (e.g. TPR and TNR in the binary case), MCC takes into account the predictive values

(e.g. PPV and NPV) in the binary case) as well. This makes MCC a prevalence-dependent metric and thus unsuited for comparison between datasets with different prevalences (Fig. 27). We therefore recommend MCC only if the provided class prevalences reflect the population of interest (F4.2 = TRUE). On the other hand, the integration of the predictive values can also be seen as an advantage. In fact,MCC is a comprehensive measure that only yields high scores if all basic rates (TPR, TNR, PPV, NPV in the binary case), (and thus also F_1 Score, Accuracy and BA) yield high scores [23]. In other words, low MCC scores can be interpreted as a warning that at least one of the related metrics must be low as well.

- **Interpretability:** BA is the macro-averaged TPR and can thus be interpreted as the chance for correct prediction if the classes have a uniform prior. If this uniform prior is also representative for the underlying medical application (F4.1 = FALSE), BA and Accuracy are identical and and represent the fraction of correctly classified samples, which is straightforward to interpret. The discrepancy of the actual BA and the theoretical value for an uninformed classifier (1 divided by the number of classes) constitutes an objective estimation of the "randomness" of a classifier irrespective of class prevalences [23]. MCC always maps such an uninformed "random" classifier to 0 and a perfect classifier to 1. Values below 0 indicate a performance worse than the simple uninformed classifier. However, the interpretation of intermediate MCC values is more difficult. It can not be used to measure the "closeness" to random guessing[23, 97]. Finally it should be noted that the behavior of MCC has been mostly studied for the binary case and the literature is lacking broad comparison for the multi-class context.
- **Bias**: It has been shown that MCC can be expressed in terms of BA with an additional factor depending on class prevalence and model bias [23]. One interpretation of this formula for the binary case is, that MCC favors models whose predictions are biased towards one class, where BA only cares about the mean of TPR and TNR. At the same time, this bias enforces a higher average of PPV and NPV, thus increasing the predictive performance of the model.
- **Example**: Imagine a breast cancer screening program with a dominance of negative cases, where FN indicate a missed cancer case and FP lead to unnecessary biopsy and stress for the patients. Both need to be avoided, butFP (even considerable amounts) might go unnoticed in BA because the large amount of TN dominates the TNR. Thus, in order to detect these FP it is crucial to check for a high PPV. MCC addresses this problem by considering all basic rates including PPV and NPV.
- **Popularity:** BA and Accuracy are far more popular than MCC, probably because of the easier interpretability and the possibility to compare across datasets.

F.3 Decision guide S3.

D3.1: Average Precision (AP) versus Free-Response Receiver Operating Characteristic (FROC) Score

The following aspects should be taken into account when deciding between AP and FROC Score:

- **Community preferences:** While AP constitutes the undisputed standard metric for object detection and instance segmentation in the computer vision community, the FROC Score is often favoured in the clinical context due to its easier interpretability despite its lack of standardization (employed False Positives per Image (FPPI) Scores vary across studies [10, 47, 78]). Thus, the decision between the two metrics often boils down to a decision between a standardized and technical validation versus an interpretable and application-focused validation.
- **Data set size awareness:** In contrast to AP, the FROC Score takes into account the total number of images in the data set (see also Fig. 28). This property does not affect relative method comparison and can be related to the underlying question "at which scale are matched objects (cardinalities) aggregated/counted?". We discuss this question and implication for associated metrics in [72]. While AP originally (i.e. in the computer vision community) counts matched objects over the entire data set (as opposed to per image), [72] demonstrates how to apply AP and other object detection metrics to (e.g. clinical) scenarios requiring per image aggregation. FROC score is a hybrid metric in this context, where *Sensitivity* is computed per data set while FP are averaged over single images (FPPI).
- **Dealing with low-confidence predictions:** It is often desired to filter low confidence predictions (e.g. objects with high confidence of being background) prior to metric computation. For AP computation, this requires a cutoff on the confidence score or upper limits of considered predictions per image or per data set. For FROC, however, with typical values of FPPI, such low-confidence predictions naturally go unconsidered, thus allowing to avoid additional filtering measures.
- **FPPI:** Different FPPI values are used in the field for computing the FROC Score, yielding non-standardized results (see Fig. 29). A potential default are the values 1/8, 1/4, 1/2, 1, 2, 4, 8, as used for multiple popular benchmarks [78, 89]. Here, lower FPPI values (smaller than one) are weighted equally to higher FPPI values (greater than 1; four values each). Deviation from this weighting might be appropriate depending on the application, but should be explained. In the biostatistics community, areas under the curve are sometimes computed constraining the FPPI range to [0,1] [74].

Pitfall: Average Precision (AP) disregards total number of images

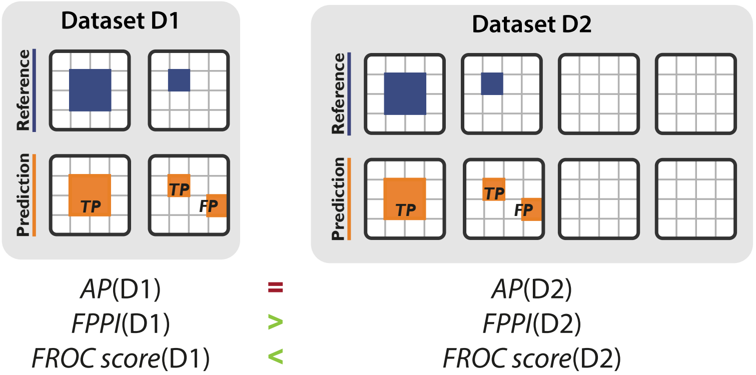


Fig. 28. Effect of the number of images per data set on the metric scores. The Average Precision (AP) metric does not take into account the total number of images, yielding the same score for data sets D1 and D2. The Free-Response Receiver Operating Characteristic (FROC) curve plots the average number of False Positives per Image (FPPI) against the Sensitivity, therefore accounting for the number of images. The FPPI is lower for D2, yielding a higher FROC score.

Pitfall: *Free-Response Receiver Operating Characteristic (FROC)* not standardized

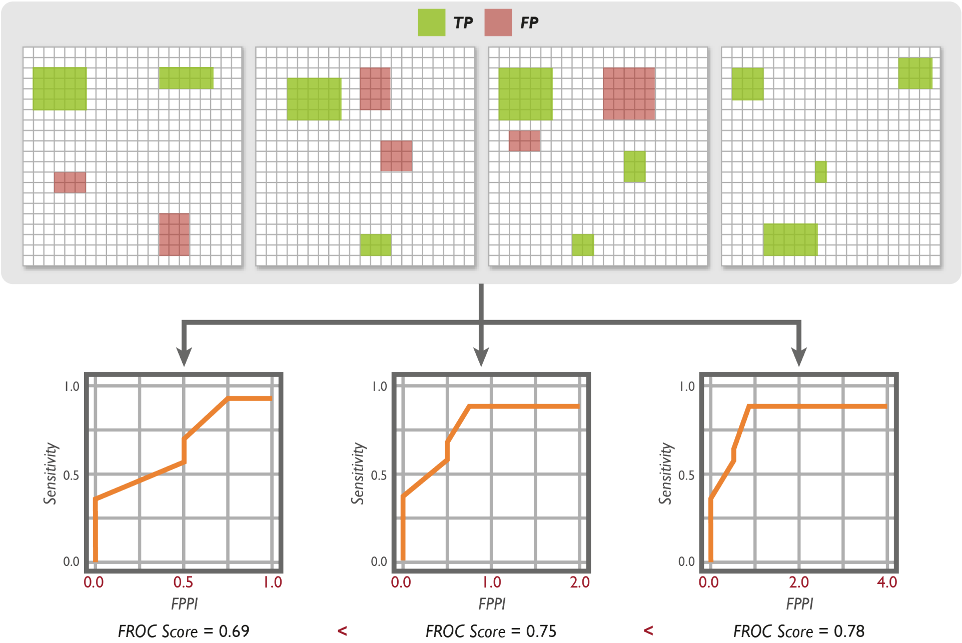


Fig. 29. Effect of defining different ranges for the False Positives per Image (FPPI) used to draw the Free-Response Receiver Operating Characteristic (FROC) curve for the same prediction (top). The resulting FROC Scores will change for different boundaries of the x-axis.

F.4 Decision guide S4.

D4.1: PPV@Sensitivity versus FPPI@Sensitivity

The decision between PPV@Sensitivity and FPPI@Sensitivity relates to most aspects described in D3.2 (AP versus FROC Score): While PPV is a generic metric applied in various communities, FPPI originated in the medical community and features improved interpretability. One important difference is that PPV is typically computed once over the entire data set, while FPPI counts FP per image and subsequently averages over the data set. In [72] we discuss the implications of these two aggregation strategies and further present an approach on how to apply various object detection metrics (including PPV) in the context of per-image aggregation.

D4.2: How to determine β in F$_\beta$ Score

The F$_\beta$ Score is defined as:

$$F_\beta = (1+\beta^2) \cdot \frac{\text{PPV} \cdot \text{Sensitivity}}{(\beta^2 \cdot \text{PPV}) + \text{Sensitivity}} = \frac{(1+\beta^2) \cdot \text{TP}}{(1+\beta^2) \cdot \text{TP} + \beta^2 \cdot \text{FN} + \text{FP}} \tag{1}$$

The most common choice is to set β to 1 resulting in equal weighting of FP and FN penalties. If unequal penalization of class confusions is desired (see F2.5.2), higher values of β result in higher weights on FN penalties compared to FP penalties and thus imply a focus on Sensitivity compared to PPV.

D4.3: F$_\beta$ Score versus Panoptic Quality (PQ)

The F$_\beta$ Score is a pure detection metric counting TP, FP, and FN detections on instance level (specifically, it represents the harmonic mean of PPV and Sensitivity, see also [72]. The "segmentation aspect" of IS is here only incorporated via a prior cutoff on the localization criterion operating on pixel level (e.g. "IoU > 0.5"). In case shifting the focus of validation more towards the segmentation quality of successfully matched (TP) instances is desired, there are two options:

(1) **Complementary segmentation metric:** One option is to select separate segmentation metrics (in addition to object detection metrics such as F$_\beta$ Score) on a per-instance basis (e.g. "DSC per TP-instance"). This selection is enabled by following the IS path in Fig. 2 traversing the SS subroutines S7 (Fig. 25) and S8 (Fig. 26).

(2) **Hybrid metric:** An alternative is to select PQ instead of F$_\beta$ Score, which allows expressing both interests (detection performance and segmentation quality) in a single score. Essentially, PQ is a modified F$_1$ Score, where TP instances do not count as "1" in the calculation, but the "1" is replaced with the associated DSC score (range [0,1]) of the instance. While combining the two aspects in a single score might be desirable, e.g. for method benchmarking or ranking, on the downside, such combined metrics make it harder to trace back performance to individual aspects (in this case: object detection versus segmentation; see Fig. 30).

Pitfall: Assessing segmentation and detection quality simultaneously

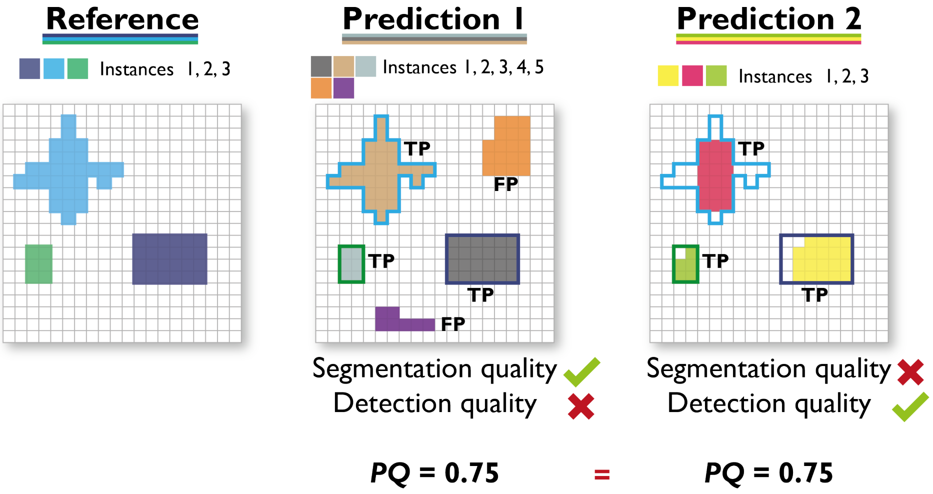


Fig. 30. Effect of assessing segmentation and detection quality in a single score. *Prediction 1* achieves a high segmentation but low detection quality (with several False Positive (FP) predictions); vice versa for *Prediction 2* (only predicting True Positive (TP) instances; no FP but low segmentation quality). However, both yield the same Panoptic Quality (PQ) score.

F.5 Decision guide S5.

D5.1: Mask IoU versus Boundary IoU versus Intersection over Reference (IoR)
The following aspects should be taken into account when deciding between Mask IoU, Boundary IoU, and IoR in instance segmentation problems. We will first focus on the more subtle distinction between Mask IoU and Boundary IoU, and finally discuss scenarios for potential usage of IoR:

Boundary versus Mask IoU

- Boundary focus: While Mask IoU measures the overlap of structures in general, Boundary IoU allows to focus on the correctness of boundaries (F2.1, see Fig. 31). Note that the focus on boundaries also comes with pitfalls. Boundary IoU can even be fooled to result in a perfect value of 1.0 despite an imperfect prediction (see Fig. 32).
- Small structures: Mask IoU over-penalizes small structures in tasks with high variability of structure sizes (F3.2) because boundary pixels increase linearly (or quadratically) with size, while total pixels increase quadratically (or cubically) with size. Boundary IoU [21] addresses this issue by selecting only pixels with a maximum distance of “d” with regard to the boundary for validation (see Fig. 31).
- Hyperparameters: For the computation of Boundary IoU, the distance “d” constitutes an additional and sensitive hyperparameter to be determined. It can be determined based on inter-rater variability, for example.
- Popularity: While Mask IoU represents an established concept that is well-known to the community, Boundary IoU is a recently proposed modification [21] that might thus require specific introduction when used in validation.

IoR In the case of a high ratio of touching reference objects, “non-split errors” (one prediction overlaps with multiple reference objects) might occur frequently. While the IoU criterion can potentially heavily penalize this scenario resulting in FN and multiple FP, a less severe penalization might be desired (check F2.5.7), e.g. in the form of the Intersection over Reference (IoR) [64]. IoR essentially considers the ratio of the area of a reference object that is covered by a prediction (see Fig. 33), allowing for multiple TP matches of the same prediction. Appropriate penalization in these cases is then ensured either by separating such errors as "merge errors" [18], or by means of additional segmentation metrics. IoR shares the behaviour of Mask IoU regarding the above discussions on boundaries and small structures. Wide-spread usage ofIoR is currently limited to the field of cell segmentation, where images with high density of structures are present [64].

Pitfall: Effect of small structures on localization criterion

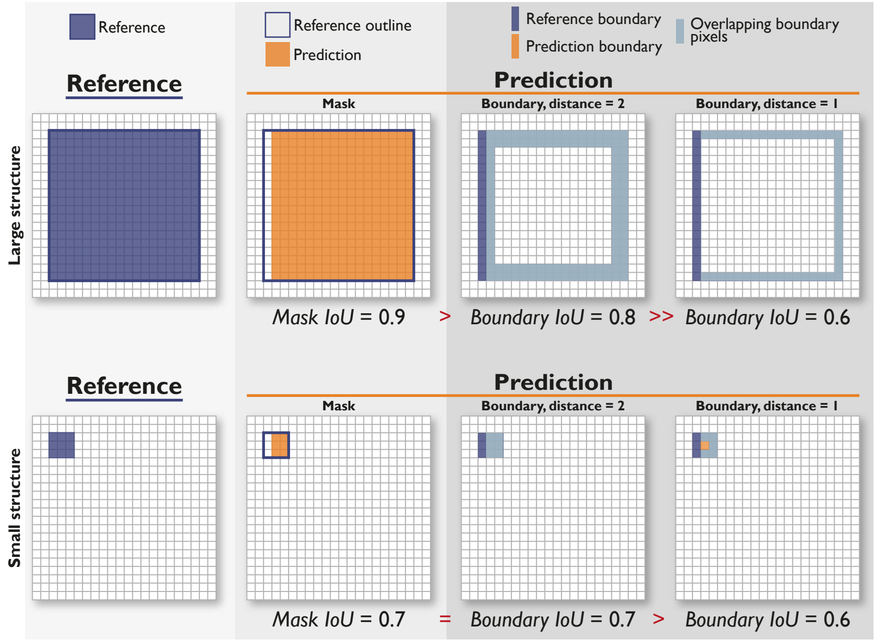


Fig. 31. Compared to the Mask Intersection over Union (IoU), the Boundary IoU (third and fourth column, representing two different thresholds) (1) specifically penalizes errors in the boundaries and (2) is more invariant to structure sizes (top: large; bottom: small).

Pitfall: Perfect Boundary IoU for imperfect prediction

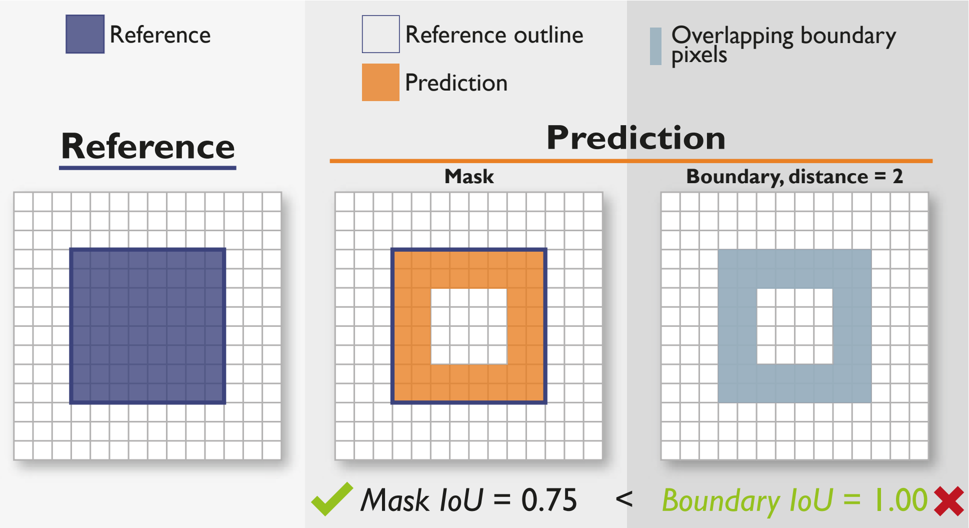


Fig. 32. Example of a perfect Boundary Intersection over Union (IoU) score for an imperfect prediction. Overlapping pixels from the reference and prediction are shown in light blue. For a prediction with a hole in the middle, the Boundary IoU may result in a score of 1.00 if the distance to border contains all mask pixels (here: distance = 2). However, the Mask IoU spots the problem and yields a lower score.

Pitfall: One prediction assigned to multiple reference objects

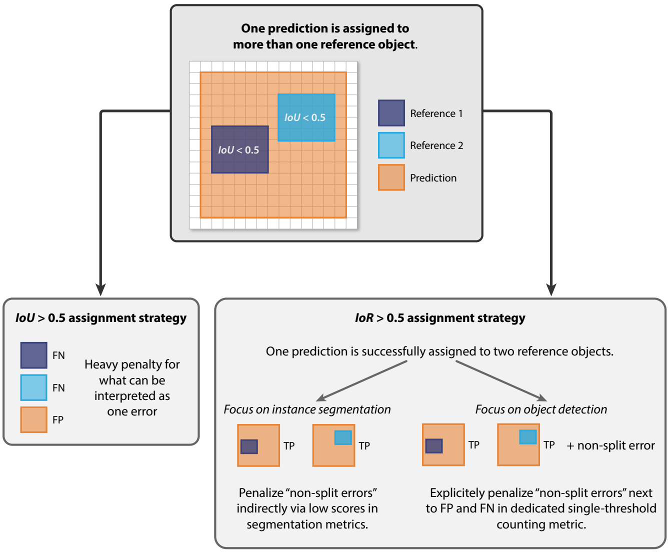

One prediction is assigned to more than one reference object.
IoU < 0.5
IoU < 0.5
Reference 1
Reference 2
Prediction
IoU > 0.5 assignment strategy
FN
FN
FP
Heavy penalty for what can be interpreted as one error
IoR > 0.5 assignment strategy
One prediction is successfully assigned to two reference objects.
Focus on instance segmentation
Focus on object detection
TP
TP
TP
TP
+ non-split error
Penalize "non-split errors" indirectly via low scores in segmentation metrics.
Explicitely penalize "non-split errors" next to FP and FN in dedicated single-threshold counting metric.

Fig. 33. In case of one prediction assigned to multiple reference objects, an assignment strategy needs to be chosen. This may be based, for example, on the *Intersection over Union (IoU)*>0.5 strategy, which may result in a heavy penalty (two False Negatives (FN) and one False Positive (FP)). Another option is to use the *Intersection over Union (IoU)*>0.5 strategy, which examines whether the prediction was successfully assigned to the reference objects. In an additional step, the "non-split errors" will be penalized. Used abbreviations: False Negative (FN), False Positive (FP) and True Positive (TP).

D5.2: Discarding information of provided annotations

Selecting a localization criterion operating on a lower/coarser resolution with regard to provided reference annotations effectively discards spatial information and should be well motivated by the given task (see Fig. 34). For instance, Box IoU is sometimes employed despite access to pixel-mask annotations (F4.4) because associated models (object detectors) are considered simpler approaches (compared to instance segmentation models). Such simplification may cause problems if structures are not well-approximated by a box shape (especially for 3D shapes, boxes usually constitute poor approximations), or if structures can overlap (F3.5), causing multi-component masks (see Fig. 35).

Discarding information of provided annotations

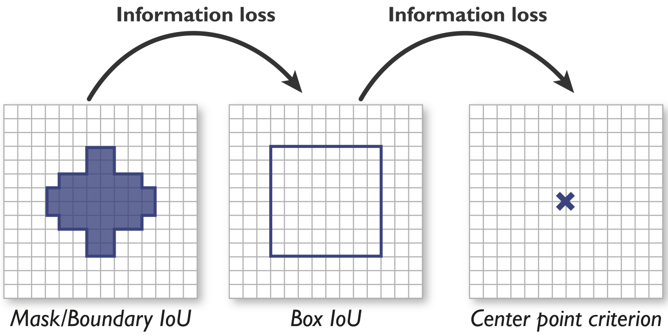


Fig. 34. Selection of a localization criterion that discards spatial information should be well motivated by the given task.

Pitfall: Effect of annotation type

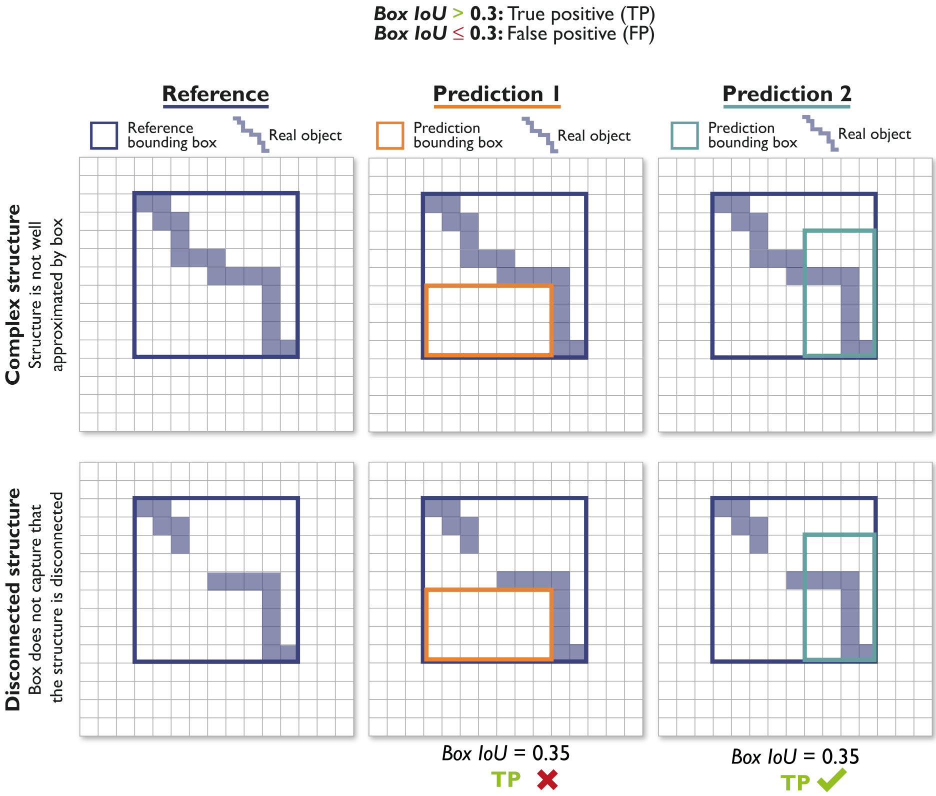


Fig. 35. Bounding boxes are not well-suited for representing complex (top) and disconnected (bottom) shapes. Specifically, they are not well-suited for capturing multi-component structures. *Predictions 1* and *2* would both end up in a True Positive (TP) detection, as the Box Intersection over Union (IoU) is larger than the threshold 0.3. However, *Prediction 1* is not hitting the real objects at all, as the given annotation does not represent them well.

D5.3: Localization without outlining of structures

When choosing a localization criterion for tasks where the mere existence of objects is of interest (as opposed to the outlining of objects), the following aspects should be considered:

- **Loose criterion:** The intuitive choice of a very loose IoU criterion (e.g. "IoU > 0" or "at least one pixel overlap") comes with the pitfall that the size of the predicted structure is in theory unbounded, i.e. the predicted location can be ambiguous (see Fig. 36).
- **Point-based criteria:** A preferable alternative for the case of pure localization (without interest in outlines) is to constrain the prediction to a single coordinate. A common criterion for this scenario is the distance to the center point of the structure (which can also be of explicit interest, see F2.3, Fig. 37). The center point[4], however, might not be a good reference for tubular structures (check F3.3) or disconnected structures (check F3.6). In such cases (and if annotations are provided in the form of masks), a binary "Point inside Mask" criterion might be the better choice. On the other hand, the "Point inside Mask" criterion does not allow for a variation of the criterion's strictness (i.e. threshold). Application despite this shortcoming should be well-justified.

Pitfall: Loose Intersection over Union (IoU) criterion

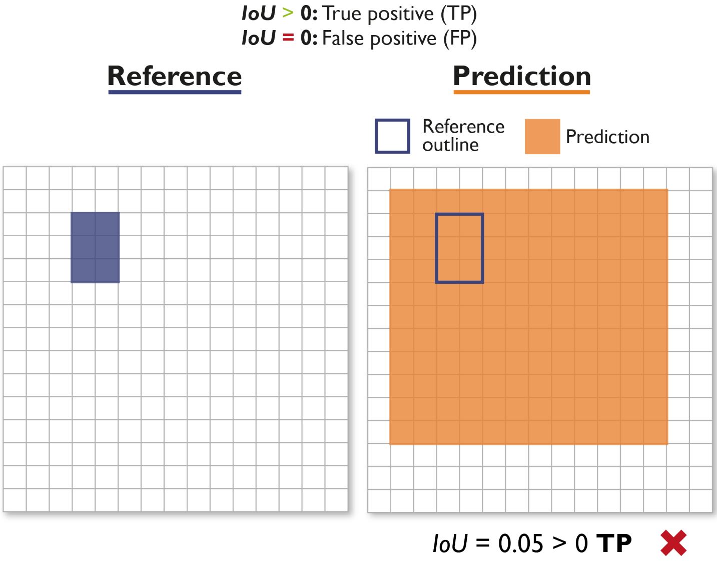


Fig. 36. Effect of a loose Intersection over Union (IoU) criterion. When defining a True Positive (TP) by an IoU > 0, the resulting localizations may be fooled by very large predictions.

[4] Depending on what kind of information the center point is derived from, different definitions are possible, for instance: (1) geometric center of the box/approximation shape, (2) geometric center of a binary mask (i.e. average of positions of all pixels), (3) center of mass of a binary mask overlaid with the original image, i.e. weighted average of positions of all pixels with weight equal to (or derived from) the intensity of a particular pixel.

Pitfall: Problems of center distance-based criteria

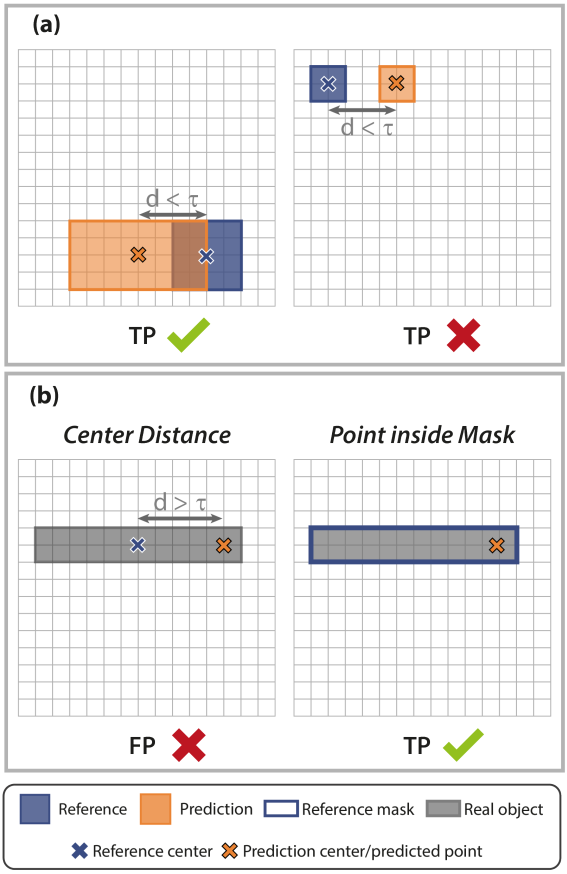


Fig. 37. Pitfalls of the Center Distance. **(a) Ignoring overlap between objects.** Both predictions have the same distance to their corresponding reference center. The Center Distance, which requires a threshold distance τ between center points not be exceeded, does not take into account the overlap between objects. However, the right prediction does not overlap with the reference and should, thus, not be considered a True Positive (TP). **(b) Tubular structures.** The Center Distance is not an ideal criterion because it implies that the prediction at the top would result in a False Positive (FP), although it is hitting the elongated structure. This could be overcome by a Point inside Mask criterion.

D5.4: Box IoU versus IoR

The following aspects should be taken into account when deciding between Box IoU and IoR:

- **Default choice:** The default choice in this setting is Box IoU.
- **Touching structures:** In the case of a high ratio of touching reference objects, "non-split errors" (one prediction overlaps with multiple reference objects) might occur frequently. In this case, the Intersection over Reference (IoR) [64], where only the ratio of a reference object's area covered by a prediction is measured, might be considered as an alternative to IoU (see Fig. 33).

D5.5: Choose localization threshold

Note that most localization criteria require a threshold to be set (e.g. IoU > 0.5 counts as detected). However, such cutoff renders the validity of results limited to the specific threshold. To increase robustness of reported metrics, it is common practice in the computer vision community to average metrics over multiple cutoff values (default for IoU criteria: from 0.5 until 0.9 in steps of 0.05). On the other hand, certain properties of the problem at hand may limit the relevance of cutoff values to lower or higher values.

The following properties might warrant validation with lower thresholds: interest in the existence of objects rather than their precise localization, small size of structures (F3.1), high variability of structure sizes (F3.2), 3D input images (as volume increases cubically with size, the desired overlap ratio might require adaptation), uncertainties in the reference (F4.3.1). Conversely, these properties typically warrant validation at higher thresholds: interest in precise localization, dense distribution of structures in images (F3.5).

It should be noted that no threshold is needed for the Point inside Mask/Box/Approx and Mask IoU > 0 criteria.

F.6 Decision guide S6.

D6.1: Assignment without predicted class probabilities on instance level

The following aspects should be considered when selecting the assignment strategy:

- **"Localization Criterion" > 0.5:** If overlapping predictions are not possible (F5.4 = FALSE), sophisticated matching strategies are often avoided in the biomedical domain by setting the the threshold for the localization criterion (Mask IoU, Boundary IoU, or IoR) to > 0.5. With this strategy, assignment ambiguities are inherently avoided. However, if either overlapping predictions are possible, a non-overlap based criterion is employed, or a criterion with a threshold above 0.5 is not appropriate, one of the following strategies should be chosen.
- **Greedy Matching:** A greedy approach can be taken, in which each reference is assigned to the best matching prediction. If predicted class scores are available (F5.1 = TRUE) this is typically achieved based on the class score ("Greedy by Score Matching"). In the given scenario with F5.1 = FALSE, an intuitive alternative is to rank predictions by the localization criterion score ("Greedy by localization criterion Matching").Assignment is then achieved by stepping through the ranked list, matching the current prediction with the most overlapping reference object, and removing the reference object from the assignment process.
- **Optimal (Hungarian) Matching:** The Hungarian algorithm optimizes the matching between predictions and reference objects while minimizing a given cost function, such as the average overlap for all matched pairs. Notably, this optimization generally leads to optimistic interpretation/validation of ambiguous model outputs, but might not represent the most realistic approximation of model performance upon applications.

F.7 Decision guide S7.

D7.1: Dice versus IoU

The DSC is identical to the F_1 Score on pixel level and closely related to the IoU, which, in turn, is identical to the Jaccard Index (see equations 2 and 3). The two metrics will yield the same ranking (of aggregated metric values) in most applications (theoretically, deviations are possible), such that there is no value in combining them. Commonly, the computer vision community prefers the IoU, while the medical image community favors the DSC.

$$IoU = \frac{DSC}{2 - DSC} \quad (2) \qquad\qquad DSC = \frac{2IoU}{1 + IoU} \quad (3)$$

D7.2: How to determine β in $\mathbf{F}_\beta$ Score

The F_β Score is defined as:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{PPV} \cdot \text{Sensitivity}}{(\beta^2 \cdot \text{PPV}) + \text{Sensitivity}} = \frac{(1 + \beta^2) \cdot \text{TP}}{(1 + \beta^2) \cdot \text{TP} + \beta^2 \cdot \text{FN} + \text{FP}} \quad (4)$$

The most common choice is to set β to 1 resulting in equal weighting of FP and FN penalties. Higher values of β result in higher weights on FN penalties (undersegmentation) compared to FP penalties (oversegmentation) and thus imply a focus on Sensitivity compared to PPV.

F.8 Decision guide S8.

D8.1: NSD and Boundary IoU

The following aspects should be considered when deciding between NSD and Boundary IoU:

- **Different research questions:** Both metrics set the focus on the boundary/contour of structures, but fundamentally differ in what they measure: NSD measures the DSC score on the surface voxels (often interpreted as the ratio of correctly predicted contour), where the strictness for what constitutes a correct boundary is controlled by a tolerance parameter. This way, noise in the image, limited resolution (propagating to noise in the reference annotations) or imprecise reference annotations can be accounted for. Boundary IoU directly measures the overlap between predicted and reference contours (without tolerance) up to a certain width (which is controlled by a width parameter). Thus, NSD is preferable if a tolerance accounting for imprecise annotations is requested. Boundary IoU, on the other hand, is preferable if errors at the contour are thought of as crucial inconsistencies that should be assessed, or if a wider area around the contour line is of interest (dynamic transition to the classical IoU).
- **Setting the hyperparameter:** The NSD and Boundary IoU both require users to manually set a hyperparameter. **NSD:** Boundary distances below the tolerance threshold will be considered TP (deviations do not count as errors). This parameter can be set according to the inter-rater variability or, if not available, heuristics. **Boundary IoU:** The distance parameter determines the thickness of the considered boundary and thus also influences the sensitivity to contour errors (the smaller the distance, the higher the sensitivity). This parameter can also be set according to the inter-rater variability (here in order to capture potential inconsistencies as opposed to disregarding noise like in NSD) or, if not available, heuristics.

D8.2: Mean Average Surface Distance (MASD) versus Average Symmetric Surface Distance (ASSD)

The ASSD puts all boundary distances (all distances from boundary A to boundary B and all distances from boundary B to boundary A) in a list, then takes the mean. Thus, if one boundary is much larger than the other, this boundary will impact the mean much more. The MASD computes the sum of the mean distances from boundary A to boundary B and the mean distances from boundary B to boundary A. Therefore, the reference and prediction boundaries contribute equally (see Fig. 38). While there are corner cases in which MASD features disadvantages compared to ASSD as well, (see Fig. 39) we generally recommend MASD because of the aforementioned advantage.

Average Symmetric Surface Distance (ASSD) versus the Mean Average Surface Distance (MASD)

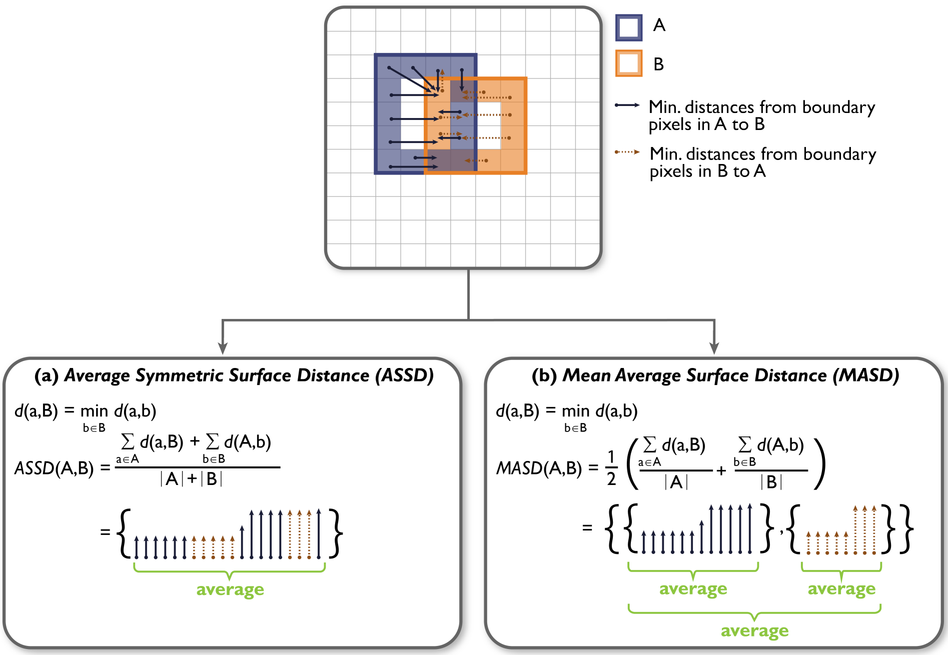

A
B
Min. distances from boundary pixels in A to B
Min. distances from boundary pixels in B to A
(a) Average Symmetric Surface Distance (ASSD)
d(a,B) = min_{b∈B} d(a,b)
ASSD(A,B) = (Σ_{a∈A} d(a,B) + Σ_{b∈B} d(A,b)) / (|A|+|B|)
average
(b) Mean Average Surface Distance (MASD)
d(a,B) = min_{b∈B} d(a,b)
MASD(A,B) = 1/2 (Σ_{a∈A} d(a,B) / |A| + Σ_{b∈B} d(A,b) / |B|)
average
average
average

Fig. 38. Most commonly used distance-based segmentation metrics: **(a)** the Average Symmetric Surface Distance (ASSD) and **(b)** the Mean Average Surface Distance (MASD). The Euclidean distance between boundary pixels a and b is defined as $d(a, b)$. Only the True Positives (TP) are considered.

Pitfall: High variability of sizes

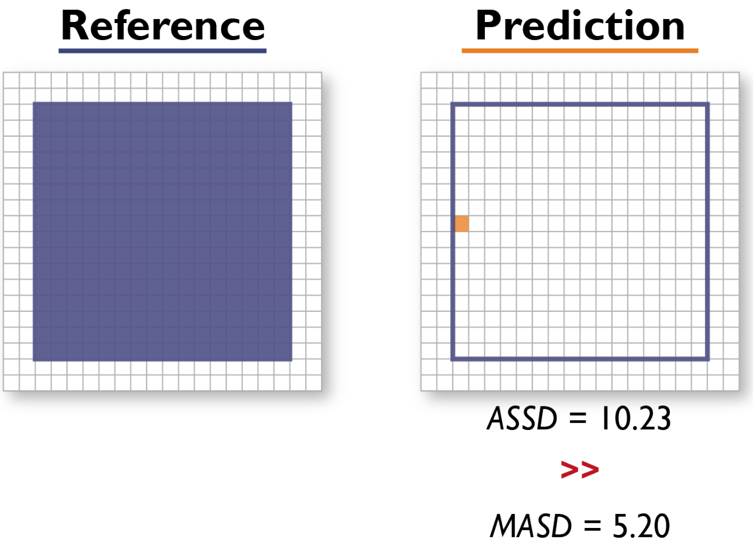


Fig. 39. Effect of one structure being much smaller than the other. If the *Prediction* is very small (here: one pixel) and located close to the reference boundary, the Mean Average Surface Distance (MASD) will be much lower compared to the ASSD.

D8.3: Hausdorff Distance (HD) versus Xth Percentile Hausdorff Distance (Xth Percentile HD) The HD calculates the maximum of all shortest distances for all points from one object boundary to the other, which is why it is also known as the Maximum Symmetric Surface Distance [95]. The Xth Percentile HD calculates the Xth percentile (e.g. 95% percentile, the Hausdorff Distance 95% Percentile (HD95)) instead of the maximum, and should therefore be used instead if spatial outliers should be disregarded (F2.5.4, see Fig. 40).

Pitfall: Noise and artifacts

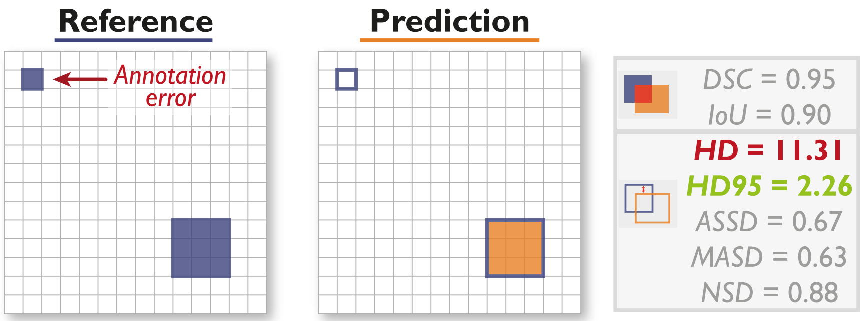


Fig. 40. Effect of annotation errors/noise. A single erroneously annotated pixel may lead to a large decrease in performance, especially in the case of the Hausdorff Distance (HD) when applied to small structures. The Hausdorff Distance 95% Percentile (HD95), on the other hand, was designed to deal with spatial outliers. Further abbreviations: Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Average Symmetric Surface Distance (ASSD), Normalized Surface Distance (NSD).

G Expected formats of reference and algorithm output

- **Image-level Classification:** The metric mapping expects the following format for image-level classification with C classes: For each image I there is a reference annotation y_I that is either indicating the class for the image ($y_I \in \{1, ..., C\}$), or, in the case of multi-label classification, indicating presence for each class ($y_I \in \{0, 1\}^C$). If the algorithm does not provide predicted class scores (F5.1 = FALSE) the algorithm output should be provided in identical format. Otherwise, for each image I, the continuous class scores for each of the classes ($\hat{y}_I \in [0, 1]^C$), indicating the predicted class probability, should be provided.
- **Semantic Segmentation** We assume the reference annotation and the algorithm output to be in the same coordinate system with identical spacing. The metric mapping expects the following format for semantic segmentation with C classes: For each pixel P there is a reference annotation y_P that either assigns a single class to P ($y_P \in \{1, ..., C\}$) or, in case of possible multiple labels per pixel, indicates assignment for each class ($y_P \in \{0, 1\}^C$). As for the algorithm output, for each pixel P there is expected to be either a single prediction ($\hat{y}_P \in \{1, ..., C\}$) or, in case of multiple possible labels per pixel, a prediction for each class ($\hat{y}_P \in \{0, 1\}^C$). Some segmentation metrics require object boundaries. For each class, boundaries are expected to be provided as a list of boundary pixels for both the reference and the prediction.
- **Object detection:** The metric mapping expects the following format for object detection with C classes: For each object O the reference consists of a tuple (y_O, l_O), where $y_O \in \{1, .., C\}$ indicates the class of the object and l_O is some location information (box, center point, radius, etc.). The algorithm output for an object prediction O is expected to comprise a tuple $(\hat{y}_O, \hat{l}_O)$ as well, where $\hat{y}_O$ indicates a single predicted class ($\hat{y}_O \in \{1, .., C\}$) optionally accompanied by an associated predicted class score ($\hat{c}_O \in [0, 1]$). See F5.1 in case no predicted class score is provided. $\hat{l}_O$ is expected to provide location information about the prediction in a similar format as the reference (box, center point, radius, etc.). In case reference objects are represented by rough outlines (F4.3) we assume that the chosen shapes (e.g. bounding box or ellipsoid) represent the underlying object adequately. Particular attention needs to be given to this aspect if objects feature a tubular shape (F3.3) or can potentially appear disconnected (F3.6).
- **Instance Segmentation** The metric mapping expects the following format for instance segmentation with C classes: For each object O the reference consists of a tuple (y_O, m_O), where $y_O \in \{1, .., C\}$ indicates the class of the object and $m_O \in \{0, 1\}^{H \times W}$ is a binary pixel map per instance matching the size of the image (height H and width W) and indicating pixel-wise location. The algorithm output for an object prediction O is expected to comprise a tuple $(\hat{y}_O, \hat{m}_O)$, where, similarly to object detection, $\hat{y}_O$ indicates a single predicted class ($\hat{y}_O \in \{1, .., C\}$) optionally accompanied by an associated predicted class score ($\hat{y}_O \in [0, 1]$). $\hat{m}_O$ denotes a binary pixel map per instance analogously to m_O. For both the reference and the predictions, object boundaries should be provided as list of boundary pixels separately for each instance. Note that annotations from semantic segmentation (not distinguishing instances of the same class) can be transformed to the instance segmentation format via connected component analysis (in case of purely non-touching instances).

In case the provided reference annotations deviate from the expected format, matching can be achieved via various measures (e.g. aggregation of pixel-level reference to required image-level reference).

H Recommendations on Metric Aggregation

In the final version of this paper, the recommendations on metric aggregation will be detailed here.

I Instantiation of the Framework for Concrete Biomedical Problems

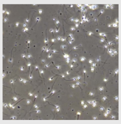
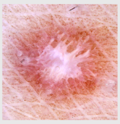
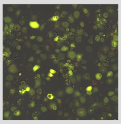
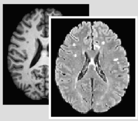
We instantiated the framework for several biological and medical image analysis use cases. The list of use cases with a link to the figures representing the recommendations is provided below:

Image-level classification

The following use cases have been instantiated for image-level classification problems. The metric recommendations can be found in Fig. 41.

- Frame-based sperm motility classification based on microscopy time-lapse video containing human spermatozoa [40] (the corresponding traversals through the decision trees are provided in Fig. 42)
- Disease classification in dermoscopic images [27] (the corresponding traversals through the decision trees are provided in Fig. 42)
- Classification of the overall autophagy stage for a collection of cells [68, 96]
- Diagnostic standard plane classification in ultrasound images [9]
- Identification of new lesions in brain multi-modal MRI images of patients with Multiple Sclerosis (MS) [30, 53]

IMAGE-LEVEL CLASSIFICATION

SCENARIO	SAMPLE INPUT IMAGE	POTENTIAL OUPUT	RECOMMENDATION
Frame-based sperm motility classification based on microscopy time-lapse video containing human spermatozoa		**Progressive motility: 0.5** Non-progressive motility: 0.4 Immotile: 0.1	**Multi-class counting metric (S2):** BA **Multi-threshold metric (S3):** AUROC **Output calibration:** ECE **Per-class counting metric (S4):** LR+
Disease classification in dermoscopic images		**Dermatofibroma: 0.6** Melanocytic nevus: 0.2 Melanoma: 0.1 Basal cell carcinoma: 0.0 Actinic keratosis: 0.0 Benign keratosis: 0.0 Vascular lesion: 0.1	
Classification of overall autophagy stage for a collection of cells		**Sequestration: 0.7** Transport to lysosomes: 0.2 Degradation: 0.1 Utilization of degradation products: 0.0	**Multi-class counting metric (S2):** MCC **Multi-threshold metric (S3):** AUROC **No output calibration** **Per-class counting metric (S4):** LR+
Diagnostic standard plane classification in ultrasound images		**Spine (sag.): 0.65** Background: 0.165, Femur 0.01, 3VV 0.01, Spine (cor.) 0.05, RVOT 0.05, LVOT 0.05	**S2:** EC **Multi-threshold metric (S3):** AUROC **Output calibration** *needed if used in interactive imaging guidance mode:* ECE **Per-class counting metric (S4):** Sensitivity@Specificity
Identification of new lesions in brain multi-modal MRI images of patients with MS		**Lesion: 0.9** No lesion: 0.1	**Multi-class counting metric (S2):** EC **Multi-threshold metric (S3):** AP **Output calibration:** PSR **Per-class counting metric (S4):** F_β Score

ABBREVIATIONS

AP Average Precision	**ECE** Expected Calibration Error
AUROC Area Under the Receiver Operating Characteristic Curve	**LR+** Positive Likelihood Ratio
BA Balanced Accuracy	**MCC** Matthews Correlation Coefficient
EC Expected Cost	**PSR** Proper Scoring Rules

Fig. 41. **Instantiation of the framework with recommendations for concrete biomedical image-level classification problems.** From top to bottom: **(1)** Frame-based sperm motility classification based on microscopy time-lapse video containing human spermatozoa [40]. **(2)** Disease classification in dermoscopic images [27]. **(3)** Classification of the overall autophagy stage for a collection of cells [68, 96]. **(4)** Diagnostic standard plane classification in ultrasound images [9]. **(5)** Identification of new lesions in brain multi-modal MRI images of patients with MS [30, 53]

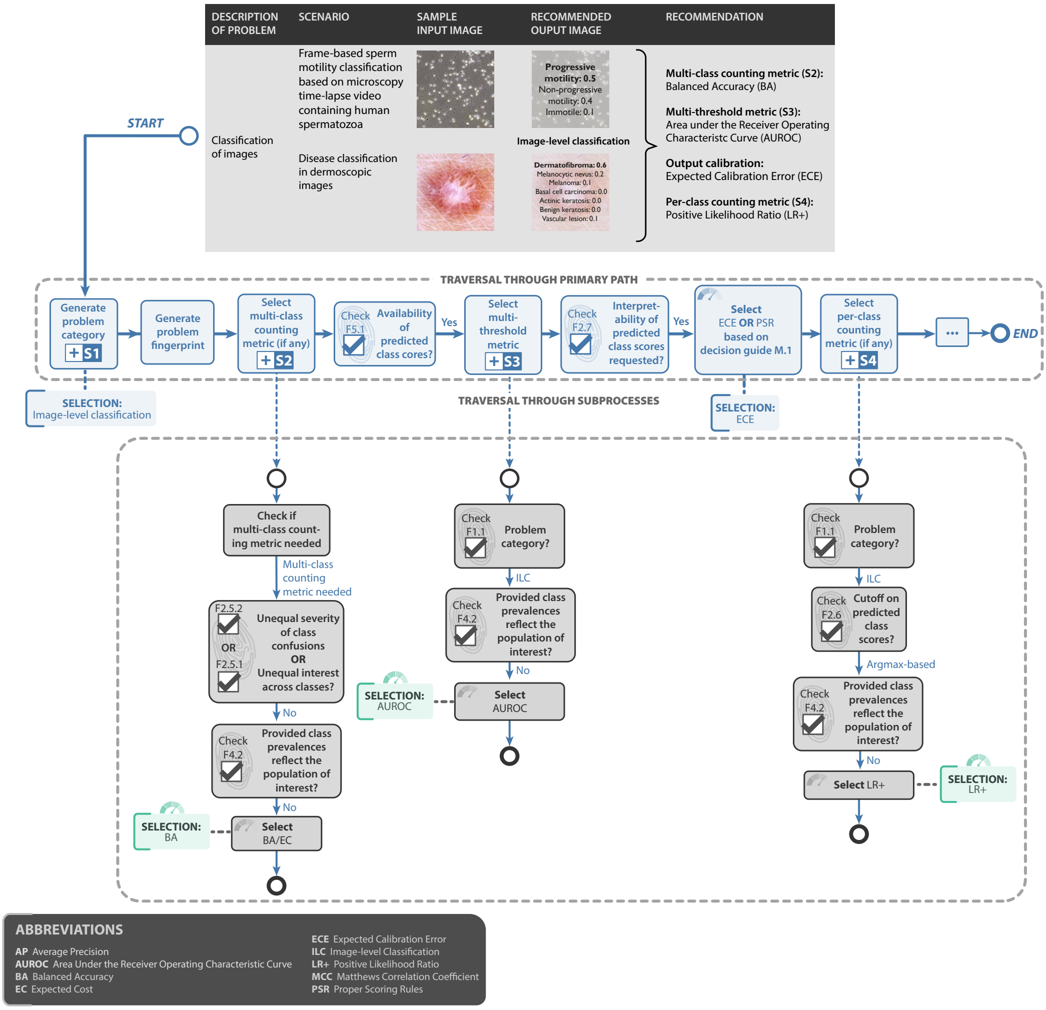
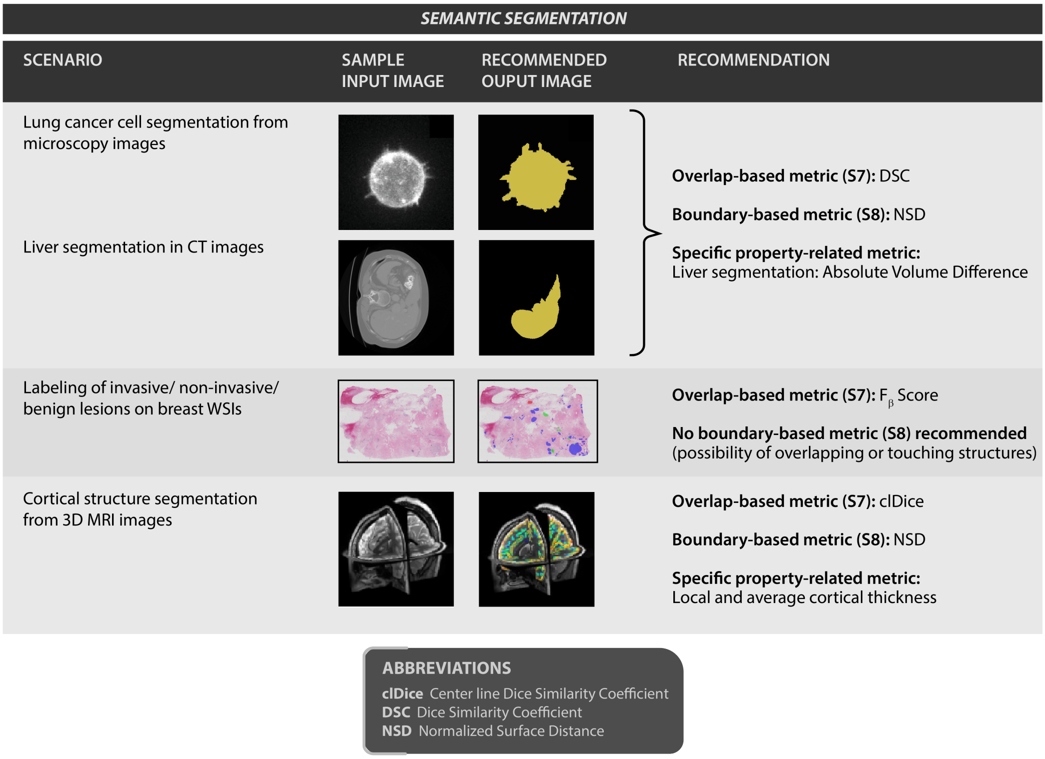

DESCRIPTION OF PROBLEM
SCENARIO
SAMPLE INPUT IMAGE
RECOMMENDED OUPUT IMAGE
RECOMMENDATION
Classification of images
Frame-based sperm motility classification based on microscopy time-lapse video containing human spermatozoa
Disease classification in dermoscopic images
Progressive motility: 0.5
Non-progressive motility: 0.4
Immotile: 0.1
Image-level classification
Dermatofibroma: 0.6
Melanocytic nevus: 0.2
Melanoma: 0.1
Basal cell carcinoma: 0.0
Actinic keratosis: 0.0
Benign keratosis: 0.0
Vascular lesion: 0.1
Multi-class counting metric (S2): Balanced Accuracy (BA)
Multi-threshold metric (S3): Area under the Receiver Operating Characteristc Curve (AUROC)
Output calibration: Expected Calibration Error (ECE)
Per-class counting metric (S4): Positive Likelihood Ratio (LR+)
START
TRAVERSAL THROUGH PRIMARY PATH
Generate problem category + S1
Generate problem fingerprint
Select multi-class counting metric (if any) + S2
Check F5.1 Availability of predicted class cores?
Yes
Select multi-threshold metric + S3
Check F2.7 Interpret-ability of predicted class scores requested?
Yes
Select ECE OR PSR based on decision guide M.1
Select per-class counting metric (if any) + S4
...
END
SELECTION: Image-level classification
SELECTION: ECE
TRAVERSAL THROUGH SUBPROCESSES
Check if multi-class count-ing metric needed
Multi-class counting metric needed
F2.5.2 OR F2.5.1 Unequal severity of class confusions OR Unequal interest across classes?
No
Check F4.2 Provided class prevalences reflect the population of interest?
No
SELECTION: BA
Select BA/EC
Check F1.1 Problem category?
ILC
Check F4.2 Provided class prevalences reflect the population of interest?
No
SELECTION: AUROC
Select AUROC
Check F1.1 Problem category?
ILC
Check F2.6 Cutoff on predicted class scores?
Argmax-based
Check F4.2 Provided class prevalences reflect the population of interest?
No
Select LR+
SELECTION: LR+
ABBREVIATIONS
AP Average Precision
AUROC Area Under the Receiver Operating Characteristic Curve
BA Balanced Accuracy
EC Expected Cost
ECE Expected Calibration Error
ILC Image-level Classification
LR+ Positive Likelihood Ratio
MCC Matthews Correlation Coefficient
PSR Proper Scoring Rules

Fig. 42. **Instantiation of the framework for the problems of frame-based sperm motility classification based on microscopy time-lapse video containing human spermatozoa [40] and disease classification in dermoscopic images [27].** The upper part shows the recommendations. The middle part shows the traversal through the main path of Fig. 2 based on the generated fingerprint for the specific problem. The bottom part shows the traversal through the subprocesses S2 (Fig. 20), S3 (Fig. 21) and S4 (Fig. 22) and the resulting recommended metrics for both use cases.

Semantic segmentation

The following use cases have been instantiated for semantic segmentation problems. The metric recommendations can be found in Fig. 43.

- Lung cancer cell segmentation from microscopy images [20] (the corresponding traversals through the decision trees are provided in Fig. 44)
- Liver segmentation in CT images [2, 80] (the corresponding traversals through the decision trees are provided in Fig. 44)
- Labeling of invasive/ non-invasive/ benign lesions on breast Whole Slide Imaging (WSI) [3]
- Cortical structure segmentation from 3D MRI images[16]



SEMANTIC SEGMENTATION

SCENARIO	SAMPLE INPUT IMAGE	RECOMMENDED OUPUT IMAGE	RECOMMENDATION
Lung cancer cell segmentation from microscopy images Liver segmentation in CT images			**Overlap-based metric (S7):** DSC **Boundary-based metric (S8):** NSD **Specific property-related metric:** Liver segmentation: Absolute Volume Difference
Labeling of invasive/ non-invasive/ benign lesions on breast WSIs			**Overlap-based metric (S7):** F_β Score **No boundary-based metric (S8) recommended** (possibility of overlapping or touching structures)
Cortical structure segmentation from 3D MRI images			**Overlap-based metric (S7):** clDice **Boundary-based metric (S8):** NSD **Specific property-related metric:** Local and average cortical thickness

ABBREVIATIONS
clDice Center line Dice Similarity Coefficient
DSC Dice Similarity Coefficient
NSD Normalized Surface Distance

Fig. 43. **Instantiation of the framework with recommendations for concrete biomedical semantic segmentation problems.** From top to bottom: **(1)** Lung cancer cell segmentation from microscopy images [20]. **(2)** Liver segmentation in Computed Tomography (CT) images [2, 80]. **(3)** Labeling of invasive/ non-invasive/ benign lesions on breast Whole Slide Imaging (WSI) [3]. **(4)** Cortical structure segmentation from 3D Magnetic Resonance Imaging (MRI) images[16].

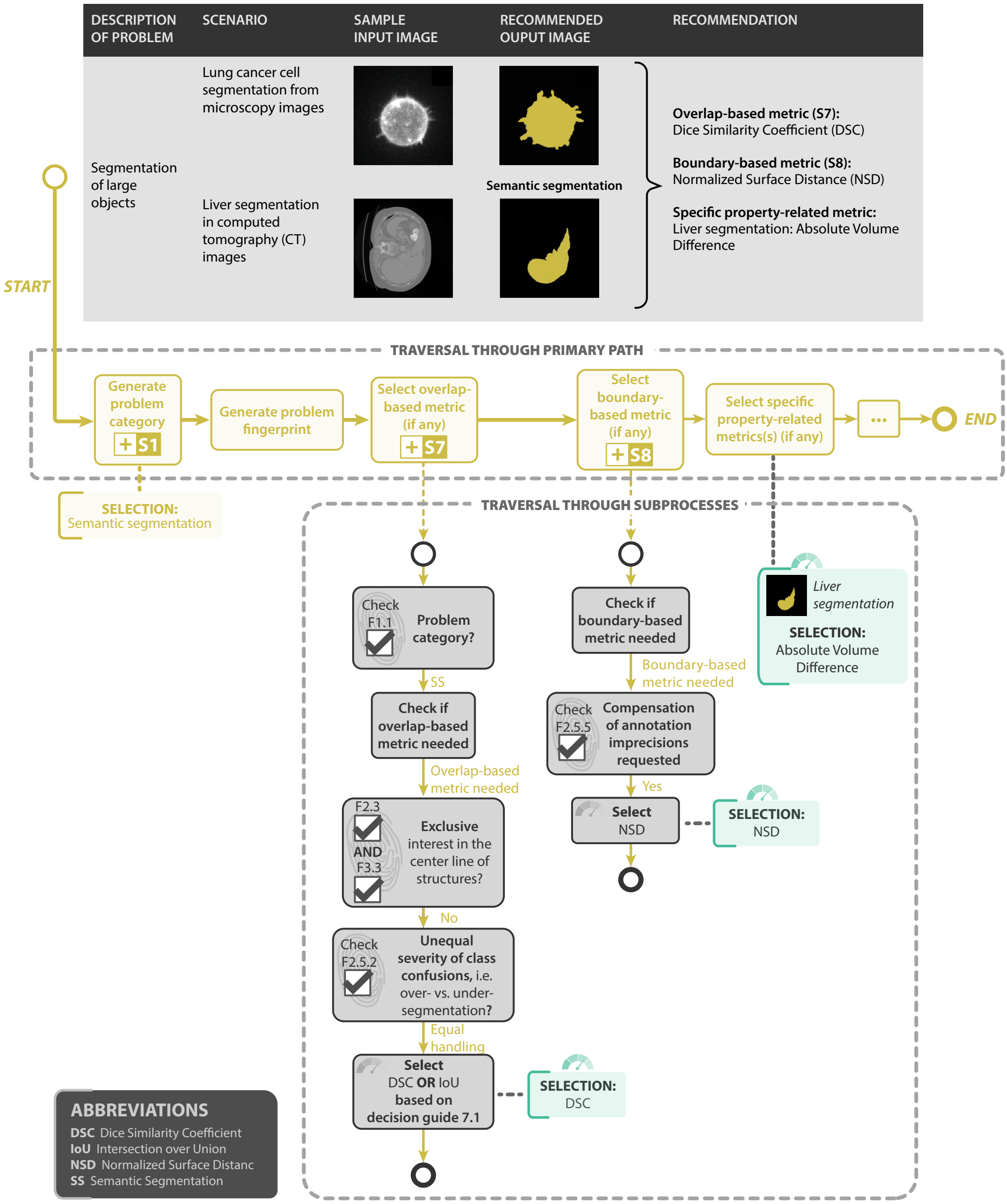
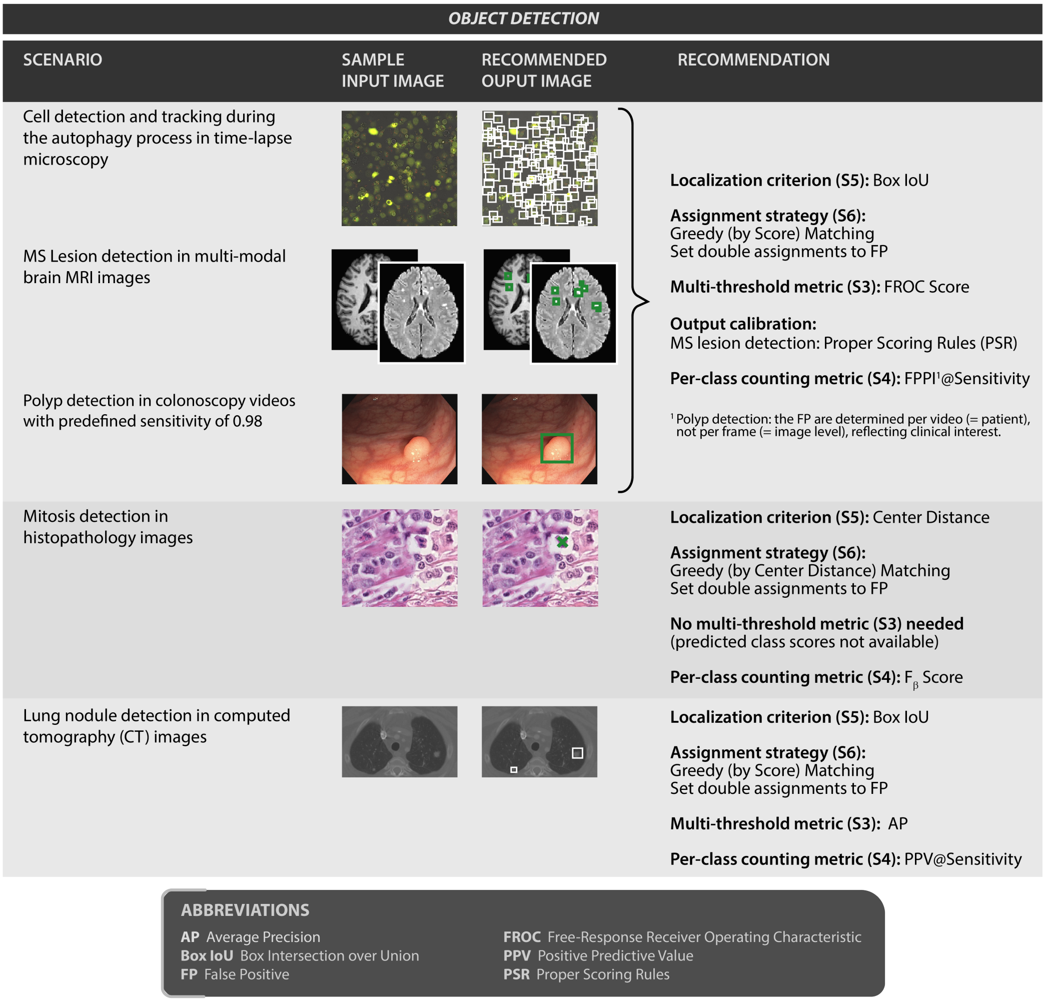

DESCRIPTION OF PROBLEM
SCENARIO
SAMPLE INPUT IMAGE
RECOMMENDED OUPUT IMAGE
RECOMMENDATION
Segmentation of large objects
Lung cancer cell segmentation from microscopy images
Liver segmentation in computed tomography (CT) images
Semantic segmentation
Overlap-based metric (S7): Dice Similarity Coefficient (DSC)
Boundary-based metric (S8): Normalized Surface Distance (NSD)
Specific property-related metric: Liver segmentation: Absolute Volume Difference
START
TRAVERSAL THROUGH PRIMARY PATH
Generate problem category S1
Generate problem fingerprint
Select overlap-based metric (if any) S7
Select boundary-based metric (if any) S8
Select specific property-related metrics(s) (if any)
...
END
SELECTION: Semantic segmentation
TRAVERSAL THROUGH SUBPROCESSES
Check F1.1 Problem category?
SS
Check if overlap-based metric needed
Overlap-based metric needed
F2.3 AND F3.3 Exclusive interest in the center line of structures?
No
Check F2.5.2 Unequal severity of class confusions, i.e. over- vs. under-segmentation?
Equal handling
Select DSC OR IoU based on decision guide 7.1
SELECTION: DSC
Check if boundary-based metric needed
Boundary-based metric needed
Check F2.5.5 Compensation of annotation imprecisions requested
Yes
Select NSD
SELECTION: NSD
Liver segmentation
SELECTION: Absolute Volume Difference
ABBREVIATIONS
DSC Dice Similarity Coefficient
IoU Intersection over Union
NSD Normalized Surface Distanc
SS Semantic Segmentation

Fig. 44. **Instantiation of the framework for the problems of lung cancer cell segmentation from microscopy images [20] and liver segmentation in computed tomography images [2, 80]**. The upper part shows the recommendations. The middle part shows the traversal through the main path of Fig. 2 based on the generated fingerprint for the specific problem. The bottom part shows the traversal through the subprocesses S7 (Fig. 25) and S8 (Fig. 26) and the resulting recommended metrics for both use cases.

Object detection

The following use cases have been instantiated for object detection problems. The metric recommendations can be found in Fig. 45.

- Cell detection and tracking during the autophagy process in time-lapse microscopy [68, 96] (the corresponding traversals through the decision trees are provided in Fig. 46)
- MS lesion detection in multi-modal brain MRI images [30, 53] (the corresponding traversals through the decision trees are provided in Fig. 46)
- Polyp detection in colonoscopy videos with predefined sensitivity of 0.9 [12, 75]
- Mitosis detection in histopathology images [7]
- Lung nodule detection in CT images [4, 5, 26]



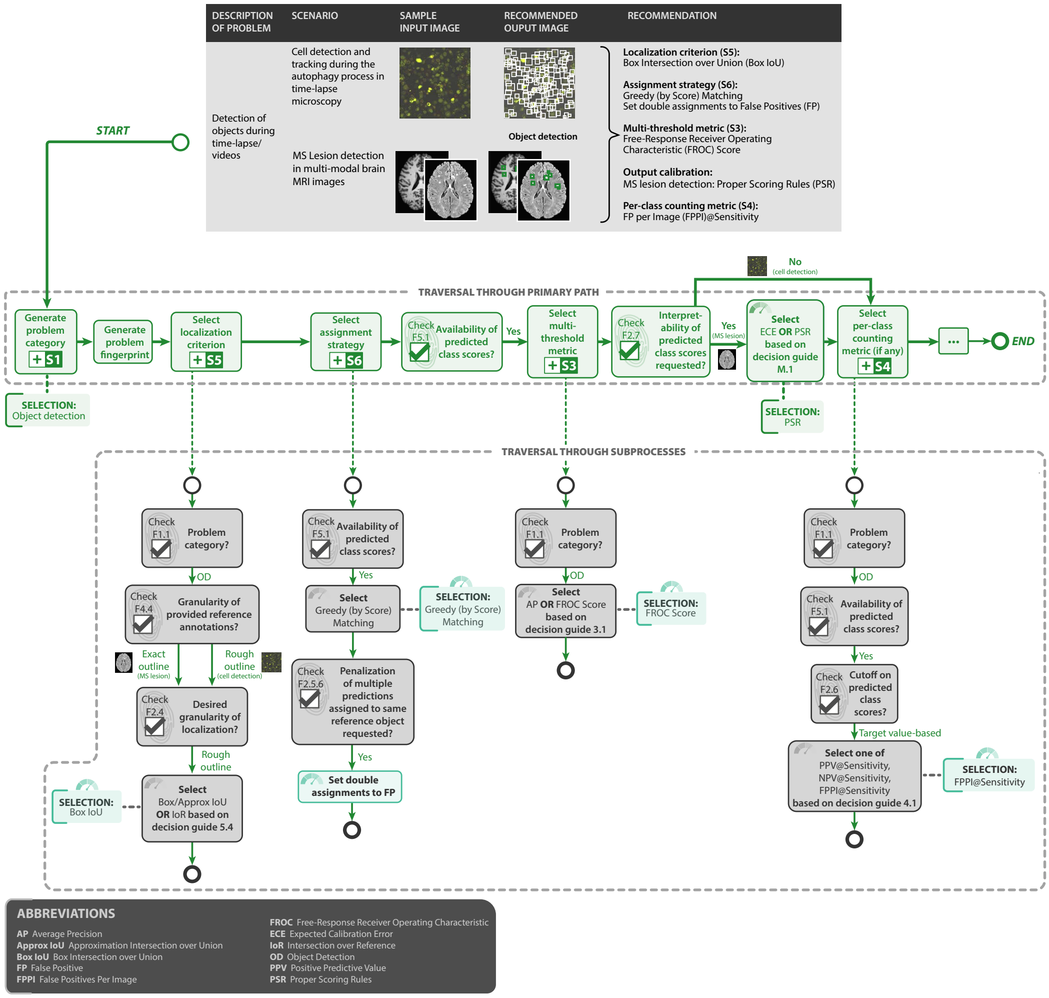
OBJECT DETECTION

SCENARIO	SAMPLE INPUT IMAGE	RECOMMENDED OUPUT IMAGE	RECOMMENDATION
Cell detection and tracking during the autophagy process in time-lapse microscopy			**Localization criterion (S5):** Box IoU **Assignment strategy (S6):** Greedy (by Score) Matching Set double assignments to FP **Multi-threshold metric (S3):** FROC Score **Output calibration:** MS lesion detection: Proper Scoring Rules (PSR) **Per-class counting metric (S4):** FPPI[1]@Sensitivity [1] Polyp detection: the FP are determined per video (= patient), not per frame (= image level), reflecting clinical interest.
MS Lesion detection in multi-modal brain MRI images			
Polyp detection in colonoscopy videos with predefined sensitivity of 0.98			
Mitosis detection in histopathology images			**Localization criterion (S5):** Center Distance **Assignment strategy (S6):** Greedy (by Center Distance) Matching Set double assignments to FP **No multi-threshold metric (S3) needed** (predicted class scores not available) **Per-class counting metric (S4):** F_β Score
Lung nodule detection in computed tomography (CT) images			**Localization criterion (S5):** Box IoU **Assignment strategy (S6):** Greedy (by Score) Matching Set double assignments to FP **Multi-threshold metric (S3):** AP **Per-class counting metric (S4):** PPV@Sensitivity

ABBREVIATIONS

AP Average Precision	**FROC** Free-Response Receiver Operating Characteristic
Box IoU Box Intersection over Union	**PPV** Positive Predictive Value
FP False Positive	**PSR** Proper Scoring Rules

Fig. 45. **Instantiation of the framework with recommendations for concrete biomedical object detection problems.** From top to bottom: **(1)** Cell detection and tracking during the autophagy process in time-lapse microscopy [68, 96]. **(2)** Multiple Sclerosis (MS) Lesion detection in multi-modal brain Magnetic Resonance Imaging (MRI) images [30, 53]. **(3)** Polyp detection in colonoscopy videos with predefined sensitivity of 0.9 [12, 75]. **(4)** Mitosis detection in histopathology images [7]. **(5)** Lung nodule detection in Computed Tomography (CT) images [4, 5, 26].



DESCRIPTION OF PROBLEM	SCENARIO	SAMPLE INPUT IMAGE	RECOMMENDED OUPUT IMAGE	RECOMMENDATION
Detection of objects during time-lapse/ videos	Cell detection and tracking during the autophagy process in time-lapse microscopy		Object detection	**Localization criterion (S5):** Box Intersection over Union (Box IoU) **Assignment strategy (S6):** Greedy (by Score) Matching Set double assignments to False Positives (FP) **Multi-threshold metric (S3):** Free-Response Receiver Operating Characteristic (FROC) Score **Output calibration:** MS lesion detection: Proper Scoring Rules (PSR) **Per-class counting metric (S4):** FP per Image (FPPI)@Sensitivity
	MS Lesion detection in multi-modal brain MRI images			

START
TRAVERSAL THROUGH PRIMARY PATH
Generate problem category + S1
Generate problem fingerprint
Select localization criterion + S5
Select assignment strategy + S6
Check F5.1 Availability of predicted class scores?
Yes
Select multi-threshold metric + S3
Check F2.7 Interpret-ability of predicted class scores requested?
No (cell detection)
Yes (MS lesion)
Select ECE OR PSR based on decision guide M.1
Select per-class counting metric (if any) + S4
...
END
SELECTION: Object detection
SELECTION: PSR
TRAVERSAL THROUGH SUBPROCESSES
Check F1.1 Problem category?
OD
Check F4.4 Granularity of provided reference annotations?
Exact outline (MS lesion)
Rough outline (cell detection)
Check F2.4 Desired granularity of localization?
Rough outline
Select Box/Approx IoU OR IoR based on decision guide 5.4
SELECTION: Box IoU
Check F5.1 Availability of predicted class scores?
Yes
Select Greedy (by Score) Matching
SELECTION: Greedy (by Score) Matching
Check F2.5.6 Penalization of multiple predictions assigned to same reference object requested?
Yes
Set double assignments to FP
Check F1.1 Problem category?
OD
Select AP OR FROC Score based on decision guide 3.1
SELECTION: FROC Score
Check F1.1 Problem category?
OD
Check F5.1 Availability of predicted class scores?
Yes
Check F2.6 Cutoff on predicted class scores?
Target value-based
Select one of PPV@Sensitivity, NPV@Sensitivity, FPPI@Sensitivity based on decision guide 4.1
SELECTION: FPPI@Sensitivity

ABBREVIATIONS

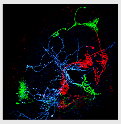
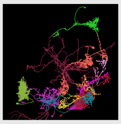
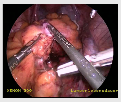
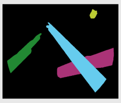
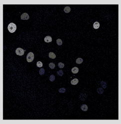
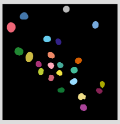
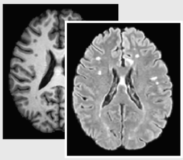
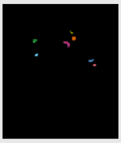
AP Average Precision
Approx IoU Approximation Intersection over Union
Box IoU Box Intersection over Union
FP False Positive
FPPI False Positives Per Image
FROC Free-Response Receiver Operating Characteristic
ECE Expected Calibration Error
IoR Intersection over Reference
OD Object Detection
PPV Positive Predictive Value
PSR Proper Scoring Rules

Fig. 46. **Instantiation of the framework for the problems of cell detection and tracking during the autophagy process in time-lapse microscopy [68, 96] and MS lesion detection in multi-modal brain MRI images [30, 53]**. The upper part shows the recommendations. The middle part shows the traversal through the main path of Fig. 2 based on the generated fingerprint for the specific problem. The bottom part shows the traversal through the subprocesses S5 (Fig. 23), S6 (Fig. 24), S3 (Fig. 21) and S4 (Fig. 22) as well as the resulting recommended localization/assignment methods and metrics for both use cases.

Instance segmentation

The following use cases have been instantiated for instance segmentation problems. The metric recommendations can be found in Fig. 47.

- Instance segmentation of neurons from the fruit fly in 3D multi-color light microscopy images [61, 65, 86] (the corresponding traversals through the decision trees are provided in Fig. 48)
- Surgical instrument instance segmentation in colonoscopy videos [60] (the corresponding traversals through the decision trees are provided in Fig. 48)
- Cell nuclei instance segmentation in time-lapse light microscopy with a subsequent goal of cell tracking [87]
- MS lesion segmentation in multi-modal brain MRI images [30, 53]

INSTANCE SEGMENTATION			
SCENARIO	SAMPLE INPUT IMAGE	RECOMMENDED OUPUT IMAGE	RECOMMENDATION
Instance segmentation of neurons from the fruit fly in 3D multi-color light microscopy images			**Localization criterion (S5):** Neuron segmentation: Mask IoU Instrument segmentation: Boundary IoU **Assignment strategy (S6):** Greedy (by Score) Matching, set double assignments to FP **Multi-threshold metric (S3):** AP **Per-class counting metric (S4):** F_β Score **Overlap-based metric (S7):** clDice **Boundary-based metric (S8):** NSD
Surgical instrument instance segmentation in colonoscopy videos			
Cell nuclei instance segmentation in time-lapse light microscopy with a subsequent goal of cell tracking			**Localization criterion (S5):** IoR **Assignment strategy (S6):** Matching via IoR > 0.5 Set double assignments to FP **No multi-threshold metric (S3) needed** (predicted class score not available) **Per-class counting metric (S4):** F_β Score **Overlap-based metric (S7):** IoU **No boundary-based metric (S8) needed** (no interest in structure boundaries)
MS Lesion segmentation in multi-modal brain MRI images			**Localization criterion (S5):** Boundary IoU **Assignment strategy (S6):** Greedy (by Score) Matching, set double assignments to FP **Multi-threshold metric (S3):** FROC Score **Output calibration:** PSR **Per-class counting metric (S4):** FPPI@Sensitivity **Overlap-based metric (S7):** DSC **Boundary-based metric (S8):** NSD

ABBREVIATIONS

AP Average Precision	**FPPI** False Positives Per Image
Boundary IoU Boundary Intersection over Union	**IoR** Intersection over Reference
clDice Center line Dice Similarity Coefficient	**IoU** Intersection over Union
DSC Dice Similarity Coefficient	**Mask IoU** Mask Intersection over Union
FROC Free-Response Receiver Operating Characteristic	**NSD** Normalized Surface Distance
FP False Positive	**PSR** Proper Scoring Rules

Fig. 47. **Instantiation of the framework with recommendations for concrete biomedical instance segmentation problems.** From top to bottom: **(1)** Instance segmentation of neurons from the fruit fly in 3D multi-color light microscopy images [61, 65, 86]. **(2)** Surgical instrument instance segmentation in colonoscopy videos [60]. **(3)** Cell nuclei instance segmentation in time-lapse light microscopy with a subsequent goal of cell tracking [87]. **(4)** Multiple Sclerosis (MS) Lesion segmentation in multi-modal brain Magnetic Resonance Imaging (MRI) images [30, 53].

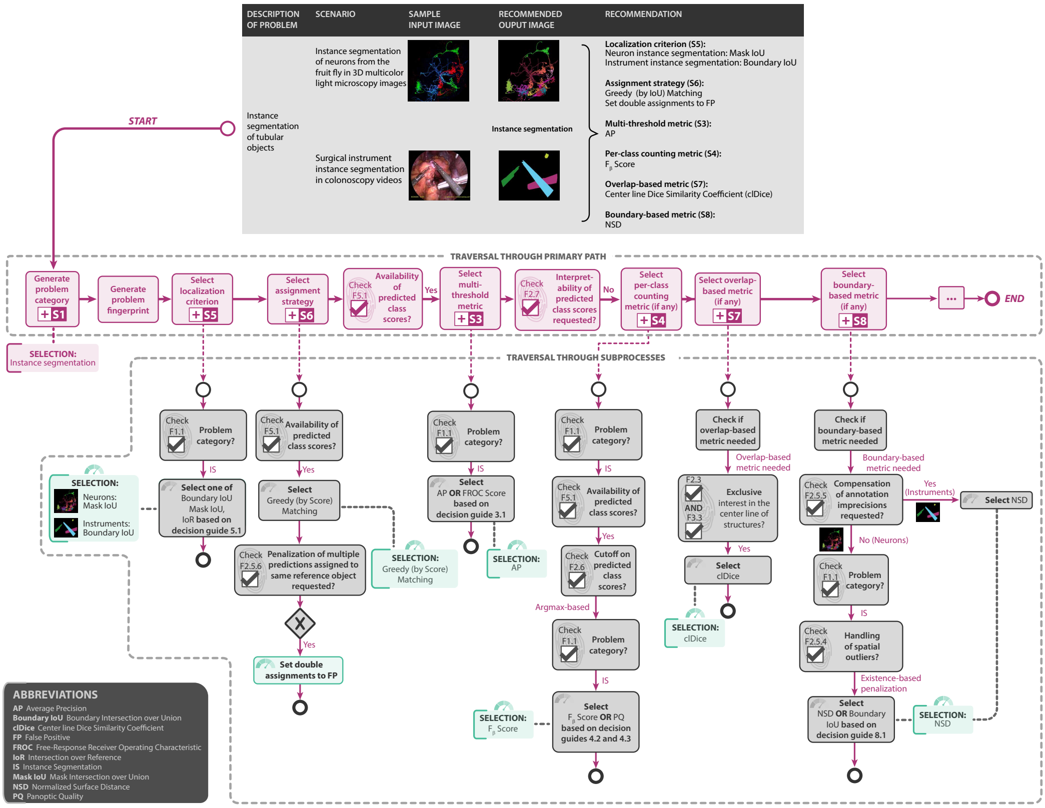

DESCRIPTION OF PROBLEM
SCENARIO
SAMPLE INPUT IMAGE
RECOMMENDED OUPUT IMAGE
RECOMMENDATION
Instance segmentation of tubular objects
Instance segmentation of neurons from the fruit fly in 3D multicolor light microscopy images
Surgical instrument instance segmentation in colonoscopy videos
Instance segmentation
Localization criterion (S5):
Neuron instance segmentation: Mask IoU
Instrument instance segmentation: Boundary IoU
Assignment strategy (S6):
Greedy (by IoU) Matching
Set double assignments to FP
Multi-threshold metric (S3):
AP
Per-class counting metric (S4):
F$_\beta$ Score
Overlap-based metric (S7):
Center line Dice Similarity Coefficient (clDice)
Boundary-based metric (S8):
NSD
START
TRAVERSAL THROUGH PRIMARY PATH
Generate problem category + S1
Generate problem fingerprint
Select localization criterion + S5
Select assignment strategy + S6
Check F5.1 Availability of predicted class scores?
Yes
Select multi-threshold metric + S3
Check F2.7 Interpret-ability of predicted class scores requested?
No
Select per-class counting metric (if any) + S4
Select overlap-based metric (if any) + S7
Select boundary-based metric (if any) + S8
...
END
SELECTION: Instance segmentation
TRAVERSAL THROUGH SUBPROCESSES
Check F1.1 Problem category?
IS
Select one of Boundary IoU Mask IoU, IoR based on decision guide 5.1
SELECTION: Neurons: Mask IoU Instruments: Boundary IoU
Check F5.1 Availability of predicted class scores?
Yes
Select Greedy (by Score) Matching
SELECTION: Greedy (by Score) Matching
Check F2.5.6 Penalization of multiple predictions assigned to same reference object requested?
Yes
Set double assignments to FP
Check F1.1 Problem category?
IS
Select AP OR FROC Score based on decision guide 3.1
SELECTION: AP
Check F1.1 Problem category?
IS
Check F5.1 Availability of predicted class scores?
Yes
Check F2.6 Cutoff on predicted class scores?
Argmax-based
Check F1.1 Problem category?
IS
Select F$_\beta$ Score OR PQ based on decision guides 4.2 and 4.3
SELECTION: F$_\beta$ Score
Check if overlap-based metric needed
Overlap-based metric needed
F2.3 AND F3.3 Exclusive interest in the center line of structures?
Yes
Select clDice
SELECTION: clDice
Check if boundary-based metric needed
Boundary-based metric needed
Check F2.5.5 Compensation of annotation imprecisions requested?
Yes (Instruments)
Select NSD
No (Neurons)
Check F1.1 Problem category?
IS
Check F2.5.4 Handling of spatial outliers?
Existence-based penalization
Select NSD OR Boundary IoU based on decision guide 8.1
SELECTION: NSD
ABBREVIATIONS
AP Average Precision
Boundary IoU Boundary Intersection over Union
clDice Center line Dice Similarity Coefficient
FP False Positive
FROC Free-Response Receiver Operating Characteristic
IoR Intersection over Reference
IS Instance Segmentation
Mask IoU Mask Intersection over Union
NSD Normalized Surface Distance
PQ Panoptic Quality

Fig. 48. **Instantiation of the framework for the problems of instance segmentation of neurons from the fruit fly in 3D multi-color light microscopy images [61, 65, 86] and surgical instrument instance segmentation in colonoscopy videos [60].** The upper part shows the recommendations. The middle part shows the traversal through the main path of Fig. 2 based on the generated fingerprint for the specific problem. The bottom part shows the traversal through the subprocesses S5 (Fig. 23), S6 (Fig. 24), S3 (Fig. 21), S4 (Fig. 22), S7 (Fig. 25) and S8 (Fig. 26) as well as the resulting recommended localization/assignment methods and metrics for both use cases.

J Glossary

- **Bounding box:** A bounding box is a (typically the smallest) rectangle drawn around and completely surrounding an object to be detected.
- **Challenge:** A challenge is an international competition, commonly hosted by individual researchers, an institute, or a professional society, that aims to comparatively assess the performance of competing algorithms on an identical data set, and thus serves to validate them. This validation is a crucial step towards the translation of an algorithm into practice.
- **Classification task:** A classification task is the task of giving categorical labels to an image or parts thereof. We distinguish classification at different scales, e.g. at image level, pixel level or object level.
- **Confidence:** See Predicted class scores.
- **Continuous class scores:** See Predicted class scores.
- **Evaluation:** See Validation.
- **FPPI@Sensitivity** FPPI at the predefined Sensitivity level, see Metric1@Metric2.
- **Hierarchical structure of classes/data:** A hierarchical structure of classes/data is present when classes or data are dependent on each other or paired, e.g. when data have been derived from the same patient, or from the same center. It requires interpretation and statistical efforts different from those suitable for independent data.
- **Hyperparameter:** A hyperparameter is a parameter whose value is optimized to control the training of an algorithm. In contrast to other parameters, it is not derived through the training process itself, but rather set before the training procedure.
- **Inference:** In the context of ML, inference denotes the processing of data by an algorithm to produce the desired output.
- **Instance:** An instance refers to a dedicated object, structure or entity in an image, such as an individual cell, tumor or medical instrument.
- **Image-level classification:** Image-level classification is the assignment of one or multiple category labels to an entire image, as detailed in App. C.
- **Instance segmentation:** Instance segmentation is the detection and delineation of each distinct object of a particular class in an image, as detailed in App. C.
- **Instantiation:** Instantiation here refers to the act of creating a specific application case of a general principle/framework.
- **Macro/micro averaging:** Macro averaging is the process of computing a metric (e.g. Sensitivity) for each class and subsequently averaging the metric scores. Micro averaging is the process of aggregating an average metric score over all classes.
- **Meta-information:** Meta-information refers to data about an image that is not explicitly contained within the image, e.g. Protected Health Information (PHI) data about the patient in radiology images.
- **Metric:** Metrics are the measures according to which performance of algorithms is quantified and validated. Depending on the domain-specific validation goal and property of interest, we distinguish between different types of metrics, e.g. reference-based (assessment of the algorithm output in comparison to the image reference) vs. non-reference based (assessment of complementary properties such as runtime or carbon footprint). Metrics can further be subdivided into different families based on their mathematical properties.
- **Metric1@Metric2** (e.g. FPPI@Sensitivity): Once a cutoff value for the predicted class probabilities has been set in such a way that the target metric value is achieved (here Metric 2: Sensitivity), other metric values (here Metric 1: FPPI) are obtained from the corresponding

fixed confusion matrix. In the example, this yields the FPPI at the predefined Sensitivity level, denoted as FPPI@Sensitivity.

- **NPV@Sensitivity** NPV at a predefined Sensitivity level, see Metric1@Metric2.
- **Object detection:** Object detection is the detection and localization of structures of one or multiple categories in an image, as detailed in App. C.
- **(Output) Calibration:** In application scenarios that involve interpreting the raw algorithm output (specifically the predicted class scores), output calibration can be used to obtain a reliable measure of confidence associated with the decision (see description of F2.7 in App. D.
- **PPV@Sensitivity:** PPV at a predefined Sensitivity level, see Metric1@Metric2.
- **Precision:** Precision is a term used differently in different scientific communities. In the medical community, for example, it commonly refers to the confidence of an output. Here, we use the term to denote the PPV.
- **Predicted class scores:** Modern neural network-based approaches usually output predicted class scores (also referred to as continuous class scores, confidence scores or pseudo-probabilities) between 0 and 1 for every image/object/pixel and class, indicating the probability of the image/object/pixel belonging to a specific class.
- **Prediction:** Prediction refers to the output of an algorithm. It is not used in the temporal sense in this paper.
- **Problem category:** Biomedical image analysis problems can be subdivided into problem categories according to the procedures performed. The category a problem falls into informs the appropriate choice of metrics. In this paper, we focus on four problem categories: Image-level classification, Semantic Segmentation, Object Detection, and Instance Segmentation.
- **Pseudo-probabilities:** See Predicted class scores.
- **Reference / Reference-based metrics:** We assume that the validation process is based on the comparison of the algorithm output and a **reference** (sometimes called **gold standard**), which is assumed to be close or equal to the correct result; the (often forever unknown) **ground truth**. In terms of metrics, we distinguish between *reference-based metrics* [46], which use the image-based reference, and *non-reference-based metrics* that assess complementary properties, such as runtime, memory consumption, or carbon footprint. The reference, sometimes also referred to as the gold standard, is a value assumed to be close or equal to the correct result, the (often forever unknown) ground truth.
- **Semantic Segmentation:** Semantic segmentation is the assignment of one or multiple category labels to each pixel in an image, as detailed in App. C.
- **Sensitivity@FPPI:** Sensitivity at the predefined FPPI level, see Metric1@Metric2.
- **Sensitivity@Specificity:** Sensitivity at the predefined Specificity level, see Metric1@Metric2.
- **Sensitivity@PPV** Sensitivity at the predefined FPPI level, see Metric1@Metric2.
- **Specificity@Sensitivity:** Specificity at the predefined Sensitivity level, see Metric1@Metric2.
- **Structure instance:** See Instance.
- **Training/Test case:** The data sets used in the process of algorithm development and validation comprise training/test cases. A case refers to the data (typically an n-dimensional image, possibly enhanced with clinical context information) that is required for an algorithm to produce one result (e.g. a segmentation or classification). A training case refers to a data set that includes reference annotations and is thus used for training an algorithm. A test case refers to a data set that is used for performance assessment
- **Type 1 and Type 2 error:** A type 1 error is a FP result, e.g. a false detection of something that is not present. A type 2 error is a FN result, e.g. a non-detection of something that is present.

- **Validation:** Validation is the process of assessing that the validated algorithm is effectively doing what it is expected to do and what it was developed for, for example that a segmentation method is actually segmenting. Evaluation is the process of assessing that the algorithm is valuable, i.e. that it brings quantifiable added value for the clinical user in a dedicated clinical context [43].

K Full Author Affiliations

Lena Maier-Hein, l.maier-hein@dkfz.de, German Cancer Research Center (DKFZ), Div. Intelligent Medical Systems and HI Helmholtz Imaging, Heidelberg, Germany and Heidelberg University, Faculty of Mathematics and Computer Science and Medical Faculty, Heidelberg, Germany and National Center for Tumor Diseases (NCT), Heidelberg, Germany
Annika Reinke*, German Cancer Research Center (DKFZ), Div. Intelligent Medical Systems and HI Helmholtz Imaging, Heidelberg, Germany and Heidelberg University, Faculty of Mathematics and Computer Science, Heidelberg, Germany
Patrick Godau, German Cancer Research Center (DKFZ), Div. Intelligent Medical Systems, Heidelberg, Germany and Heidelberg University, Faculty of Mathematics and Computer Science, Heidelberg, Germany
Minu D. Tizabi, German Cancer Research Center (DKFZ), Div. Intelligent Medical Systems and HI Helmholtz Imaging, Heidelberg, Germany
Evangelia Christodoulou, German Cancer Research Center (DKFZ), Div. Intelligent Medical Systems, Heidelberg, Germany
Ben Glocker, Imperial College London, Biomedical Image Analysis Group, Department of Computing, London, UK
Fabian Isensee, German Cancer Research Center (DKFZ), HI Applied Computer Vision Lab, Div. of Medical Image Computing, Heidelberg, Germany
Jens Kleesiek, University Medicine Essen, Translational Image-guided Oncology (TIO), Institute for AI in Medicine (IKIM), Essen, Germany
Michal Kozubek, Masaryk University, Centre for Biomedical Image Analysis, Brno, Czech Republic
Mauricio Reyes, University of Bern, ARTORG Center for Biomedical Engineering Research, Bern, Switzerland
Michael A. Riegler, Simula Metropolitan Center for Digital Engineering, Oslo, Norway and UiT The Arctic University of Norway, Oslo, Norway
Manuel Wiesenfarth, German Cancer Research Center (DKFZ), Div. Biostatistics, Heidelberg, Germany
Michael Baumgartner, German Cancer Research Center (DKFZ), Div. Medical Image Computing, Heidelberg, Germany
Matthias Eisenmann, German Cancer Research Center (DKFZ), Div. Intelligent Medical Systems and HI Helmholtz Imaging, Heidelberg, Germany
Doreen Heckmann-Nötzel, German Cancer Research Center (DKFZ), Div. Intelligent Medical Systems, Heidelberg, Germany and National Center for Tumor Diseases (NCT), Heidelberg, Germany
A. Emre Kavur, German Cancer Research Center (DKFZ), HI Applied Computer Vision Lab, Div. of Medical Image Computing; Div. Intelligent Medical Systems, Heidelberg, Germany
Tim Rädsch, German Cancer Research Center (DKFZ), Div. Intelligent Medical Systems and HI Helmholtz Imaging, Heidelberg, Germany
Laura Acion, CONICET – Universidad de Buenos Aires, Instituto de Cálculo, Buenos Aires, Argentina and University of Iowa, Department of Psychiatry, Iowa City, Iowa, USA
Michela Antonelli, King's College London, School of Biomedical Engineering and Imaging Science, London, UK and University College London, Centre for Medical Image Computing, London, UK
Tal Arbel, McGill University, Centre for Intelligent Machines and MILA (Quebec Artificial Intelligence Institute), Montreal, Canada
Spyridon Bakas, University of Pennsylvania, Center for Biomedical Image Computing & Analytics, Philadelphia, Pennsylvania, USA and Perelman School of Medicine at the University of Pennsylvania, Department of Pathology & Laboratory Medicine and Department of Radiology, Philadelphia, Pennsylvania, USA
Peter Bankhead, University of Edinburgh, Institute of Genetics and Cancer, Edinburgh, UK
Arriel Benis, Holon Institute of Technology, Faculty of Industrial Engineering and Technology Management, Faculty of Digital Technologies in Medicine, Holon, Israel
M. Jorge Cardoso, King's College London, School of Biomedical Engineering and Imaging Science, London, UK and University College London, Department of Medical Physics and Biomedical Engineering, London, UK
Veronika Cheplygina, IT University of Copenhagen, Copenhagen, Denmark
Beth Cimini, Broad Institute of MIT and Harvard, Imaging Platform, Cambridge, Massachusetts, USA

Gary S. Collins, University of Oxford, Centre for Statistics in Medicine, Oxford, UK
Keyvan Farahani, National Cancer Institute, Center for Biomedical Informatics and Information Technology, USA
Luciana Ferrer, Instituto de Investigacion en Ciencias de la Computacion (ICC), CONICET-UBA, Argentina
Adrian Galdran, Universitat Pompeu Fabra, Barcelona, Spain and University of Adelaide, Adelaide, Australia
Bram van Ginneken, Fraunhofer MEVIS, Bremen, Germany and Radboud University Medical Center, Radboud Institute for Health Sciences, Nijmegen, The Netherlands
Robert Haase, DFG Cluster of Excellence „Physics of Life", Technische Universität (TU) Dresden, Dresden, Germany and Center for Systems Biology , Dresden, Germany
Daniel A. Hashimoto, Case Western Reserve University School of Medicine, University Hospitals Cleveland Medical Center, Cleveland, Ohio, USA
Michael M. Hoffman, University Health Network, Princess Margaret Cancer Centre, Toronto, Canada and University of Toronto, Department of Medical Biophysics, Department of Computer Science, Toronto, Canada and Vector Institute, Toronto, Canada
Merel Huisman, Radboud University Medical Center, Department of Radiology, Nijmegen, The Netherlands
Pierre Jannin, Université de Rennes 1, Inserm, Laboratoire Traitement du Signal et de l'Image – UMR_S 1099, Rennes, France
Charles E. Kahn, University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania, USA
Dagmar Kainmueller, Max-Delbrück Center for Molecular Medicine, Regensburg, Germany
Bernhard Kainz, Imperial College London, Department of Computing, Faculty of Engineering, London, UK
Alexandros Karargyris, IHU Strasbourg, Strasbourg, France
Alan Karthikesalingam, Google Health Deepmind, London, UK
Hannes Kenngott, Heidelberg University Hospital, Department of General, Visceral and Transplantation Surgery, Heidelberg, Germany
Florian Kofler, Helmholtz AI, München, Germany
Annette Kopp-Schneider, German Cancer Research Center (DKFZ), Div. Biostatistics, Heidelberg, Germany
Anna Kreshuk, European Molecular Biology Laboratory (EMBL), Cell Biology and Biophysics Unit, Heidelberg, Germany
Tahsin Kurc, Stony Brook University, Stony Brook Cancer Center, Stony Brook, New York, USA
Bennett A. Landman, Vanderbilt University, Electrical Engineering, Nashville, Tennessee, USA
Geert Litjens, Radboud University Medical Center, Department of Pathology and Radboud Institute for Health Sciences, Nijmegen, The Netherlands
Amin Madani, University Health Network, Department of Surgery, Toronto, Canada
Klaus Maier-Hein, German Cancer Research Center (DKFZ), Div. Medical Image Computing and HI Helmholtz Imaging, Heidelberg, Germany and Heidelberg University Hospital, Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg, Germany
Anne L. Martel, Sunnybrook Research Institute, Physical Sciences, Toronto, Canada and University of Toronto, Department of Medical Biophysics, Toronto, Canada
Peter Mattson, Google, Mountain View, USA
Erik Meijering, University of New South Wales, School of Computer Science and Engineering, New South Wales, Australia
Bjoern Menze, University of Zurich, Department of Quantitative Biomedicine, Zurich, Switzerland
David Moher, Ottawa Hospital Research Institute, Centre for Journalology, Clinical Epidemiology Program, Ottawa, Canada and University of Ottawa, School of Epidemiology and Public Health, Faculty of Medicine, Ottawa, Canada
Karel G.M. Moons, UMC Utrecht, University Utrecht , Julius Center for Health Sciences and Primary Care, Utrecht, The Netherlands
Henning Müller, University of Applied Sciences Western Switzerland (HES-SO), Information Systems Institute, Sierre, Switzerland and University of Geneva, Medical Faculty, Geneva, Switzerland
Brennan Nichyporuk, MILA (Quebec Artificial Intelligence Institute), Montreal, Canada
Felix Nickel, Heidelberg University Hospital, Department of General, Visceral and Transplantation Surgery, Heidelberg, Germany

Jens Petersen, German Cancer Research Center (DKFZ), Div. Medical Image Computing, Heidelberg, Germany
Nasir Rajpoot, University of Warwick, Tissue Image Analytics Laboratory, Department of Computer Science, Coventry, West Midlands, UK
Nicola Rieke, NVIDIA GmbH, Munich, Germany
Julio Saez-Rodriguez, Heidelberg University, Institute for Computational Biomedicine, Heidelberg, Germany, Faculty of Medicine and Heidelberg University Hospital, Heidelberg, Germany and BioQuant, Heidelberg, Germany
Clarisa Sánchez Gutiérrez, University of Amsterdam, Informatics Institute, Faculaty of Science, Amsterdam, The Netherlands
Shravya Shetty, Google, Google Health, Palo Alto, USA
Maarten van Smeden, University Medical Center Utrecht, Julius Center for Health Sciences and Primary Care, Utrecht, The Netherlands
Carole H. Sudre, University College London, Centre for Medical Image Computing and Medical Research Council Unit for Lifelong Health and Ageing at UCL, London, UK and King's College London, School of Biomedical Engineering and Imaging Science, London, UK
Ronald M. Summers, National Institutes of Health, Radiology and Imaging Sciences, Clinical Center, Bethesda, Maryland, USA
Abdel A. Taha, Scigility International GmbH, , Vienna, Austria
Sotirios A. Tsaftaris, The University of Edinburgh, School of Engineering, Edinburgh, Scotland
Ben Van Calster, Katholieke Universiteit (KU) Leuven, Department of Development and Regeneration and EPI-centre, Leuven, Belgium and Leiden University Medical Center, Department of Biomedical Data Sciences, Leiden, The Netherlands
Gaël Varoquaux, INRIA Saclay-Île de France, Parietal Project-team, Palaiseau, France
Paul F. Jäger, German Cancer Research Center (DKFZ), Interactive Machine Learning Group and HI Helmholtz Imaging, Heidelberg, Germany.