Common Limitations of Image Processing Metrics: A Picture Story

ANNIKA REINKE*, German Cancer Research Center (DKFZ), Germany and Heidelberg University, Germany
MINU D. TIZABI, German Cancer Research Center (DKFZ), Germany
CAROLE H. SUDRE, University College London, UK and King's College London, UK
MATTHIAS EISENMANN, German Cancer Research Center (DKFZ), Germany
TIM RÄDSCH, German Cancer Research Center (DKFZ), Germany and understandAI GmbH, Germany
MICHAEL BAUMGARTNER, 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
BEN GLOCKER, Imperial College London, UK
PATRICK GODAU, German Cancer Research Center (DKFZ), Germany and Heidelberg University, Germany
FRED HAMPRECHT, Heidelberg University, Germany
DANIEL A. HASHIMOTO, Case Western Reserve University School of Medicine, USA
DOREEN HECKMANN-NÖTZEL, German Cancer Research Center (DKFZ), Germany
MICHAEL M. HOFFMAN, University Health Network, Canada, University of Toronto, Canada, and Vector Institute, Canada
MEREL HUISMAN, University Medical Center Utrecht, The Netherlands
FABIAN ISENSEE, German Cancer Research Center (DKFZ), Germany
PIERRE JANNIN, Université de Rennes 1, Inserm, France
CHARLES E. KAHN, University of Pennsylvania, USA
ALEXANDROS KARARGYRIS, IHU Strasbourg, France
ALAN KARTHIKESALINGAM, Google Health Deepmind, UK
BERNHARD KAINZ, Imperial College London, UK
EMRE KAVUR, German Cancer Research Center (DKFZ), Germany
HANNES KENNGOTT, Heidelberg University Hospital, Germany
JENS KLEESIEK, University Medicine Essen, Germany
THIJS KOOI, Lunit Inc, South Korea
MICHAL KOZUBEK, Masaryk University, Czech Republic
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
ANNE L. MARTEL, Sunnybrook Research Institute, Canada and University of Toronto, Canada
PETER MATTSON, Google, US
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
FELIX NICKEL, Heidelberg University Hospital, Germany
JENS PETERSEN, German Cancer Research Center (DKFZ), Germany
GORKEM POLAT, Middle East Technical University, Turkey
NASIR RAJPOOT, University of Warwick, UK
MAURICIO REYES, University of Bern, Switzerland
NICOLA RIEKE, NVIDIA GmbH, Germany
MICHAEL A. RIEGLER, Simula Metropolitan Center for Digital Engineering, Norway and UiT The Arctic University of Norway, Norway
HASSAN RIVAZ, Concordia University, Canada
JULIO SAEZ-RODRIGUEZ, Heidelberg University, Germany, Heidelberg University Hospital, Germany, and BioQuant, Germany
CLARISA SÁNCHEZ GUTIÉRREZ, University of Amsterdam, The Netherlands
JULIEN SCHROETER, McGill University, Canada
ANINDO SAHA, Radboud University Medical Center, The Netherlands
SHRAVYA SHETTY, Google, US
BRAM STIELTJES, University Hospital of Basel, Switzerland
RONALD M. SUMMERS, National Institutes of Health, USA
ABDEL A. TAHA,, Austria
SOTIRIOS A. TSAFTARIS, The University of Edinburgh, Scotland
BRAM VAN GINNEKEN, Fraunhofer MEVIS, Germany and Radboud University Medical Center, The Netherlands
GAËL VAROQUAUX, INRIA Saclay-Île de France, France
MANUEL WIESENFARTH, German Cancer Research Center (DKFZ), Germany
ZIV R. YANIV, National Institute of Allergy and Infectious Diseases, National Institutes of Health, USA
ANNETTE KOPP-SCHNEIDER, German Cancer Research Center (DKFZ), Germany
PAUL JÄGER, German Cancer Research Center (DKFZ), Germany
LENA MAIER-HEIN, German Cancer Research Center (DKFZ), Germany and Heidelberg University, Germany

Disclaimer

This is a continuous paper on limitations of commonly used metrics in image analysis. For missing references, use cases, other comments for improvements or questions, please contact `a.reinke@dkfz.de` or `l.maier-hein@dkfz.de`. Substantial contributions (best in form of a graphical draft) to this dynamic document will be acknowledged with a co-authorship.

Abstract: While the importance of automatic image analysis is continuously increasing, recent meta-research revealed major flaws with respect to algorithm validation. Performance metrics are particularly key for meaningful, objective, and transparent performance assessment and validation of the used automatic algorithms, but relatively little attention has been given to the practical pitfalls when using specific metrics for a given image analysis task. These are typically related to (1) the disregard of inherent metric properties, such as the behaviour in the presence of class imbalance or small target structures, (2) the disregard of inherent data set properties, such as the non-independence of the test cases, and (3) the disregard of the actual biomedical domain interest that the metrics should reflect. This living dynamically document has the purpose to illustrate important limitations of performance metrics commonly applied in the field of image analysis. In this context, it focuses on biomedical image analysis problems that can be phrased as image-level classification, semantic segmentation, instance segmentation, or object detection task. The current version is based on a Delphi process on metrics conducted by an international consortium of image analysis experts from more than 60 institutions worldwide.

[*] The complete list of affiliations and authors' addresses can be found in the Appendix A.

Keywords: Good Scientific Practice, Validation, Metrics, Image Processing, Image Analysis, Segmentation, Classification, Detection, Medical Imaging

Contents

1 ACRONYMS

AP Average Precision
ASSD Average Symmetric Surface Distance
AUC Area under the curve
AUROC Area under the Receiver Operating Characteristic curve
clDice Centerline Dice Similarity Coefficient
DSC Dice Similarity Coefficient
FN False Negative
FP False Positive
FPPI False Positives per Image
FPR False Positive Rate
FROC Free-Response Receiver Operating Characteristic
HD Hausdorff Distance
HD95 Hausdorff Distance 95% percentile
IoU Intersection over Union
MCC Matthews Correlation Coefficient
NPV Negative Predictive Value
NSD Normalized Surface Distance
PPV Positive Predictive Value
PR Precision-Recall
ROC Receiver Operating Characteristic
TN True Negative
TNR True Negative Rate
TP True Positive
TPR True Positive Rate

2 PURPOSE

Validation of biological and medical image analysis algorithms is of the utmost importance for making scientific progress and for translating methodological research into practice. Validation metrics[1], the measures according to which performance of algorithms is quantified, constitute a core component of validation design. While metrics can measure various quantities of interest, including speed, memory consumption or carbon footprint, most metrics applied today are *reference-based metrics*, which have the purpose of measuring the agreement of an algorithm prediction with a given reference. The reference, in turn, serves as an approximation of the (typically unknown) ground truth.

Knowing the properties of metrics in use and making educated choices is essential for meaningful and reliable validation in image analysis. Although several papers highlight specific strengths and weaknesses of common metrics [13, 25, 26, 31, 50], an international survey [30] revealed the choice of inappropriate metrics as one of the core problems related to performance assessment in medical image analysis. Similar problems are present in other imaging domains [10, 16]. Under the umbrella of the Helmholtz Imaging Platform (HIP)[2], three international initiatives have now joined forces to address these issues: the Biomedical Image Analysis Challenges (BIAS) initiative[3], the Medical Image Computing and Computer Assisted Interventions (MICCAI) Society's special interest group on challenges[4], as well as the benchmarking working group of the MONAI framework[5]. A core mission is to provide researchers with guidelines and tools to choose the performance metrics in a problem- and context-aware manner. This dynamically updated document aims to illustrate important pitfalls and drawbacks of metrics commonly applied in the field of image analysis. The current version is based on a Delphi process on metrics conducted with an international consortium of medical image analysis experts. A Delphi process is a multi-stage survey process designed to pool the knowledge of several experts to arrive at a consensus decision [4].

The Delphi consortium focused on problems reporting biomedical research that can be phrased as **image-level classification**, **semantic segmentation**, **instance segmentation** or **object detection** (Figure 1). Essentially, these can all be interpreted as a classification task at different scales and thus share many aspects in terms of validation (Figure 2). For example, an object detection task can be interpreted as an object-/instance-level classification task, while a segmentation task can be interpreted as a pixel-level classification task. We will refer to these four different task types as *problem categories*. Please note that we will use the term "pixel" even for three-dimensional (or n-dimensional) images for increased readability instead of referring to "pixels/voxels". Most of the examples are shown for two-dimensional images and can be translated to the n-dimensional case.

The manuscript is structured as follows. As a foundation, we first review the most commonly applied metrics for the problem categories addressed in this paper (Sec. 3). Since a common problem in the biomedical image analysis community is the selection of metrics from the wrong problem category, Sec. 4 highlights pitfalls relevant in this context. The following sections then present pitfalls for image-level classification (Sec. 5), image segmentation, including semantic and instance segmentation (Sec. 6), and object detection, including instance segmentation (Sec. 7). Finally, cross-topic pitfalls are highlighted (Sec. 8). An overview of all figures is presented in Table 1.

[1] not to be confused with distance metrics in the strict mathematical sense

[2] https://www.helmholtz-imaging.de/

[3] https://www.dkfz.de/en/cami/research/topics/biasInitiative.html

[4] https://miccai.org/index.php/special-interest-groups/challenges/

[5] https://monai.io/

Problem categories addressed by this paper

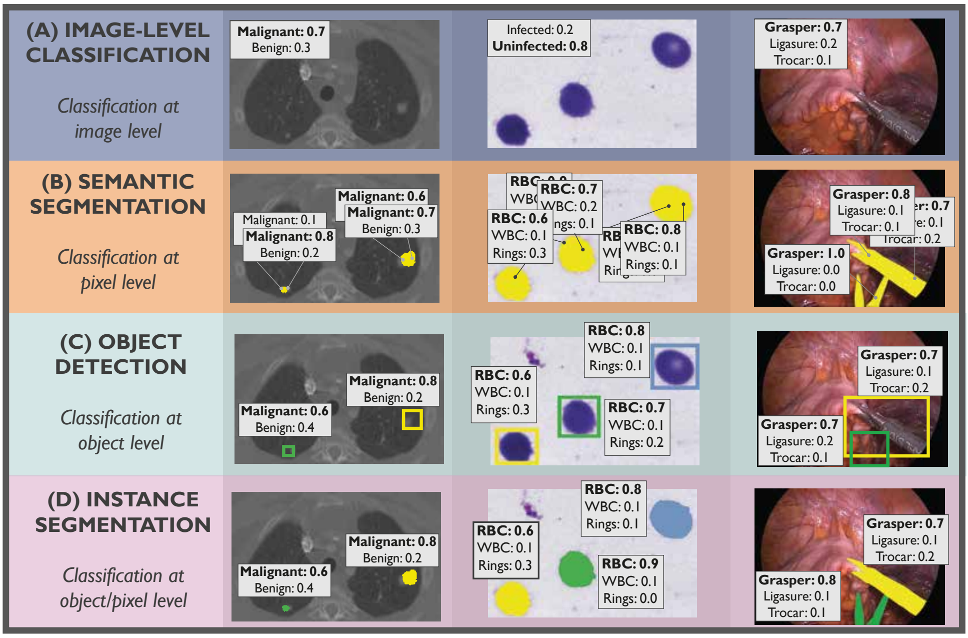


Fig. 1. Problem categories covered in this paper and illustrated for three different application domains: radiology (left), cell biology (middle), surgery (right). The common denominator of the underlying research problems is the fact that they can be interpreted as classification tasks. The classification occurs at various scales, image level, object level and/or pixel level. Each of these tasks assigns a class label to the image or (multiple) components of it. (A) The **image-level classification** task involves assigning a class label to the whole image. (B) The **semantic segmentation** task involves assigning a class label to each individual pixel. (C) The **object detection** task assigns a class label to identified objects. (D) The **instance segmentation** task assigns a label to identified objects made up of multiple pixels. Gray boxes show the predicted class probabilities on image level, pixel level or object level. The class with the highest probability is shown in bold. Further abbreviations: Red Blood Cell (RBC), White Blood Cell (WBC).

Table 1. Overview of figures on pitfalls related to metrics classified into (1) pitfalls due to category-metric mismatch, (2) category-specific pitfalls and (3) cross-topic pitfalls. For each illustration, the corresponding figure and page number is given. Please note that the pitfalls are typically illustrated for only one or two problem categories but often also apply to other problem categories, as indicated in the table.

Source of potential pitfall	**Figure(s)**
Pitfalls due to category-metric mismatch	
Mismatch: Semantic segmentation ↔ object detection	Fig. 14 (Page 28)
Mismatch: Semantic ↔ instance segmentation	Fig. 15 (Page 29)
Mismatch: Image-level classification ↔ object detection	Fig. 16 (Page 31)
No matching problem category	Fig. 17 (Page 32)
Pitfalls in image-level classification	
High class imbalance	Fig. 18 (Page 34)
More than two classes available	Fig. 19 (Page 35)
Unequal importance of classes	Fig. 20 (Page 37)
Interdependencies between classes	Fig. 21 (Page 38)
Stratification based on meta-information	Fig. 22 (Page 39)
Missing prevalence correction	Fig. 23 (Page 40)
Upper bound in *Cohen's* κ not equally obtainable	Fig. 24 (Page 42)
Multi-threshold metric-related properties *(here: pitfalls illustrated for object detection problems)*	Fig. 45 (Page 71)
	Fig. 46 (Page 72)
	Fig. 47 (Page 73)
Pitfalls in semantic segmentation	
Small size of structures relative to pixel size	Fig. 25 (Page 45)
(here: pitfall illustrated for object detection problems)	Fig. 41 (Page 66)
High variability of structure sizes	Fig. 26 (Page 46)
Complex shape of structures	Fig. 27 (Page 48)
Particular importance of structure volume	Fig. 28 (Page 49)
Particular importance of structure center	Fig. 29 (Page 50)
Particular importance of structure boundaries	Fig. 30 (Page 52)
(here: pitfall illustrated for object detection problems)	Fig. 41 (Page 66)
(here: pitfall illustrated for object detection problems)	Fig. 42 (Page 67)
Possibility of multiple labels per unit	Fig. 31 (Page 53)
Noisy reference standard	Fig. 32 (Page 54)
Possibility of outliers in reference annotation	Fig. 33 (Page 55)
Possibility of reference or prediction without target structure(s)	Fig. 34 (Page 57)
Dependency on image resolution	Fig. 35 (Page 58)
Over- *vs.* undersegmentation	Fig. 36 (Page 59)
Choice of global decision threshold	Fig. 37 (Page 60)
High class imbalance *(here: pitfall illustrated for image-level classification problems)*	Fig. 18 (Page 34)
More than two classes available *(here: pitfall illustrated for image-level classification problems)*	Fig. 19 (Page 35)
Unequal importance of classes *(here: pitfall illustrated for image-level classification problems)*	Fig. 20 (Page 37)
Interdependencies between classes *(here: pitfall illustrated for image-level classification problems)*	Fig. 21 (Page 38)
Pitfalls in object detection	
Mathematical implications of center-based localization criteria	Fig. 38 (Page 62)
Mathematical implications of *IoU*-based localization criterion	Fig. 39 (Page 64)
Type of the provided annotations	Fig. 40 (Page 65)
Effect of small structures on localization criterion	Fig. 41 (Page 66)
(here: pitfall illustrated for semantic segmentation problems)	Fig. 25 (Page 45)
Perfect *Boundary IoU* for imperfect prediction	Fig. 42 (Page 67)

Possibility of reference or prediction without target structure(s) *(here: pitfall illustrated for semantic segmentation problems)*	Fig. 43 (Page 69) Fig. 34 (Page 57)
Average Precision vs. Free-response ROC score	Fig. 44 (Page 70)
Multi-threshold metric-related properties	Fig. 45 (Page 71) Fig. 46 (Page 72) Fig. 47 (Page 73)
High class imbalance *(here: pitfall illustrated for image-level classification problems)*	Fig. 18 (Page 34)
More than two classes available *(here: pitfall illustrated for image-level classification problems)*	Fig. 19 (Page 35)
Unequal importance of classes *(here: pitfall illustrated for image-level classification problems)*	Fig. 20 (Page 37)
Interdependencies between classes *(here: pitfall illustrated for image-level classification problems)*	Fig. 21 (Page 38)
Particular importance of structure center *(here: pitfall illustrated for semantic segmentation problems)*	Fig. 29 (Page 50)
High variability of structure sizes *(here: pitfall illustrated for semantic segmentation problems)*	Fig. 26 (Page 46)
Complex shape of structures *(here: pitfall illustrated for semantic segmentation problems)*	Fig. 27 (Page 48)
Possibility of multiple labels per unit *(here: pitfall illustrated for semantic segmentation problems)*	Fig. 31 (Page 53)
Noisy reference standard *(here: pitfall illustrated for semantic segmentation problems)*	Fig. 32 (Page 54)
Possibility of outliers in reference annotation *(here: pitfall illustrated for semantic segmentation problems)*	Fig. 33 (Page 55)
Choice of global decision threshold *(here: pitfall illustrated for semantic segmentation problems)*	Fig. 37 (Page 60)
Pitfalls in instance segmentation	
Small size of structures relative to pixel size	Fig. 25 (Page 45) Fig. 41 (Page 66)
High variability of structure sizes	Fig. 26 (Page 46)
Complex shape of structures	Fig. 27 (Page 48)
Particular importance of structure volume	Fig. 28 (Page 49)
Particular importance of structure center	Fig. 29 (Page 50)
Particular importance of structure boundaries	Fig. 41 (Page 66) Fig. 42 (Page 67) Fig. 30 (Page 52)
Possibility of multiple labels per unit	Fig. 31 (Page 53)
Noisy reference standard	Fig. 32 (Page 54)
Possibility of outliers in reference annotation	Fig. 33 (Page 55)
Possibility of reference or prediction without target structure(s)	Fig. 34 (Page 57) Fig. 43 (Page 69)
Dependency on image resolution	Fig. 35 (Page 58)
Over- *vs.* undersegmentation	Fig. 36 (Page 59)
Choice of global decision threshold	Fig. 37 (Page 60)
Mathematical implications of center-based localization criteria	Fig. 38 (Page 62)
Mathematical implications of *IoU*-based localization criterion	Fig. 39 (Page 64)
Effect of small structures on localization criterion	Fig. 41 (Page 66) Fig. 25 (Page 45)
Perfect *Boundary IoU* for imperfect prediction	Fig. 42 (Page 67)
Average Precision vs. Free-response ROC score	Fig. 44 (Page 70)
Multi-threshold metric-related properties	Fig. 45 (Page 71) Fig. 46 (Page 72) Fig. 47 (Page 73)
High class imbalance *(here: pitfall illustrated for image-level classification problems)*	Fig. 18 (Page 34)
More than two classes available *(here: pitfall illustrated for image-level classification problems)*	Fig. 19 (Page 35)
Unequal importance of classes *(here: pitfall illustrated for image-level classification problems)*	Fig. 20 (Page 37)
Interdependencies between classes *(here: pitfall illustrated for image-level classification problems)*	Fig. 21 (Page 38)

Cross-topic pitfalls	
Uninformative visualization	Fig. 48 (Page 75)
Metric aggregation for invalid algorithm output (e.g. `NaN`s)	Fig. 49 (Page 76) Fig. 50 (Page 77)
Hierarchical data aggregation	Fig. 51 (Page 78)
Aggregation per class	Fig. 52 (Page 79)
Metric combination	Fig. 53 (Page 80)

3 FUNDAMENTALS

The present work focuses on biomedical image analysis problems that can be interpreted as a classification task at image, object or pixel level. The vast majority of metrics for these problem categories is directly or indirectly based on epidemiological principles of True Positive (TP), False Negative (FN), False Positive (FP), True Negative (TN), i.e. the *cardinalities* of the so-called confusion matrix, depicted in Figure 2. The TP/FN/FP/TN, from now on referred to as cardinalities, can occur at image (segment), object or pixel level. They are typically computed by comparing the prediction of the algorithm to a reference annotation. Modern neural network-based approaches typically require a threshold to be set in order to convert the algorithm output comprising predicted class probabilities[6] (also referred to as continuous class scores) to a confusion matrix, as illustrated in Figure 2. For the purpose of metric recommendation, the available metrics can be broadly classified as follows:

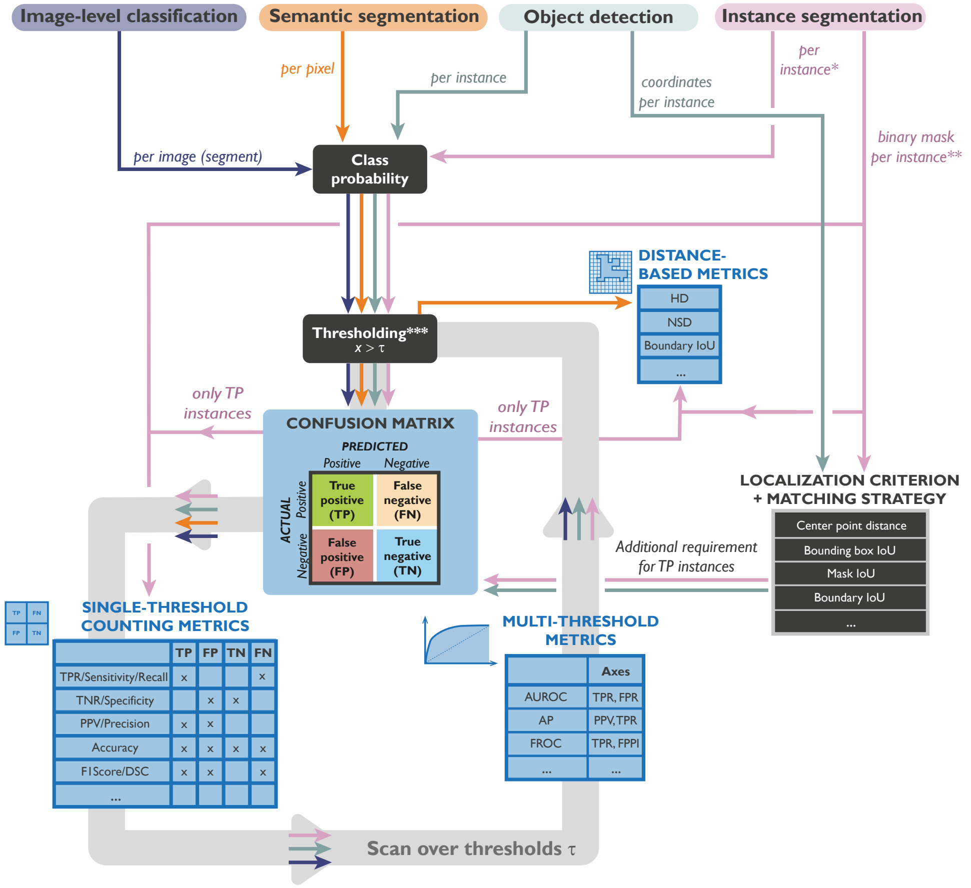
- **Single-threshold counting metrics** operate directly on the confusion matrix and express the metric value as a function of the cardinalities (see Figures 3, 4 and 6). In the context of segmentation, they have typically been referred to as **overlap-based** metrics [47]. Popular examples are *Sensitivity*, *Specificity*, *Precision*, *Accuracy*, *Dice Similarity Coefficient (DSC)* and *Intersection over Union (IoU)*.
- **Multi-threshold metrics** operate on a dynamic confusion matrix, reflecting the conflicting properties of interest, such as high *Sensitivity* and high *Specificity*. Popular examples include the *Area under the Receiver Operating Characteristic curve (AUROC)* (see Figure 5) and *Average Precision (AP)* (see Figure 11).
- **Distance-based metrics** have been designed for semantic and instance segmentation tasks. They operate exclusively on the TP and rely on the explicit definition of object boundaries (see Figures 7 and 8). Popular examples are the *Hausdorff Distance (HD)* and the *Normalized Surface Distance (NSD)* (see Figure 8).

Depending on the context (e.g. image-level classification *vs.* semantic segmentation task) and the community (e.g. medical imaging community *vs.* computer vision community), identical metrics are referred to with different terminology. For example, *Sensitivity*, *True Positive Rate (TPR)* and *Recall* refer to the same concept. The same holds true for the *DSC* and the *F1 score*. The most relevant metrics for the problem categories in the scope of this paper are introduced in the following.

Most metrics are recommended to be applied per class, meaning that a potential multi-class problem is converted to multiple binary classification problems, such that each relevant class serves as the positive class once. This results in different confusion matrices depending on which class is used as the positive class.

[6] Please note that we refer to pseudo-probabilities.

Relationships between metric families



CONFUSION MATRIX

		PREDICTED Positive	PREDICTED Negative
ACTUAL	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

DISTANCE-BASED METRICS

HD
NSD
Boundary IoU
...

LOCALIZATION CRITERION + MATCHING STRATEGY

Center point distance
Bounding box IoU
Mask IoU
Boundary IoU
...

TP	FN
FP	TN

SINGLE-THRESHOLD COUNTING METRICS

	TP	FP	TN	FN
TPR/Sensitivity/Recall	x			x
TNR/Specificity		x	x	
PPV/Precision	x	x		
Accuracy	x	x	x	x
F1Score/DSC	x	x		x
...				

MULTI-THRESHOLD METRICS

	Axes
AUROC	TPR, FPR
AP	PPV,TPR
FROC	TPR, FPPI
...	...

*This visualization assumes that potential ad hoc identification of instances (e.g. via connected components) has already been performed. Not assigning class probabilities per instance during post-processing amounts to assigning scores of 1 to all instances.

** If no overlapping objects are expected, masks are often collapsed to numbered instances in one integer map for efficiency.

*** In multi-class scenarios, thresholding is often implicitly performed via an argmax operation implying a fixed threshold $\tau = 1/n_{classes}$.

Fig. 2. Most popular metric families and their relationships relevant for the problem categories addressed by this paper. The vast majority of metrics is directly or indirectly based on the cardinalities of the confusion matrix. The available metrics can be broadly classified into **single-threshold counting metrics** that operate directly on the confusion matrix generated for a fixed threshold, **multi-threshold metrics** that operate on a dynamic confusion matrix (depending on threshold) and **distance-based metrics** that take into account the structure contour(s) or other spatial information, such as the structure center. Further abbreviations: Area under the Receiver Operating Characteristic curve (AUROC), Average Precision (AP), Dice Similarity Coefficient (DSC), False Positives per Image (FPPI), False Positive Rate (FPR), Hausdorff Distance (HD), Intersection over Union (IoU), Normalized Surface Distance (NSD), Positive Predictive Value (PPV), True Negative Rate (TNR), True Positive Rate (TPR).

3.1 Image-level Classification

Image-level classification refers to the process of assigning one or multiple labels, or *classes* to an image. If there is only one class of interest (e.g. cancer *vs.* no cancer), we speak of *binary classification*, otherwise of *categorical classification.* Modern algorithms usually output **predicted class probabilities** (or continuous class scores) between 0 and 1 for every image and class, indicating the probability of the image belonging to a specific class. By introducing a threshold (e.g. 0.5), predictions are considered as positive (e.g. cancer = true) if they are above the threshold or negative if they are below the threshold. Afterwards, predictions are assigned to the cardinalities (e.g. a cancer patient with prediction cancer = true is considered as TP) [11]. The most popular classification metrics are single-threshold counting metrics, operating on a confusion matrix with fixed threshold on the class probabilities, and multi-threshold metrics, as detailed in the following.

Single-threshold counting metrics

Figures 3 and 4 present the most common binary **single-threshold counting metrics** used for image-level classification. Please note that these metrics are also commonly used in segmentation and object detection tasks. For segmentation tasks, they are often referred to as overlap-based metrics. Each of the presented metrics covers specific properties.

The *Sensitivity* (also referred to as *Recall*, *TPR* or *Hit rate*) focuses on the actual positives (TP and FN) and represents the fraction of positives that were correctly detected as such. In contrast, *Precision* (or *PPV*) divides the TP by the total number of predicted positive cases, thus aiming to represent the probability of a positive prediction corresponding to an actual positive. A value of 1 would imply that all positive predicted cases are actually positives, but it might still be the case that positive cases were missed. Please note that the term *Precision* has multiple meanings. In the context of computer assisted interventions, for example, it typically refers to the measured variance. Hence, the usage of its synonym *PPV* may be preferred.

In analogy to the *Sensitivity* for positives, *Specificity* (also referred to as *Selectivity* or *TNR*) focuses on the negative cases by computing the fraction of negatives that were correctly detected as such. Similarly to the *Precision*, the *Negative Predictive Value (NPV)* divides the TN by the total number of predicted negative cases and measures how many of the predicted negative samples were actually negative. *Specificity* and *NPV* require the definition of TN cases, which is not always possible. In object detection tasks, for example (see Sec. 3.3), TN are typically ill-defined and not provided. Therefore, these measures can not be computed in those cases.

As illustrated in Figure 18, reporting of a single metric like *Sensitivity*, *Precision* or *Specificity* can be highly misleading because, for example, non-informative classifiers can achieve high values on imbalanced classes. The *F1 score* (also known as *DSC* in the context of segmentation), overcomes this issue by representing the harmonic mean of *Precision* and *Sensitivity* and therefore penalizing extreme values of either metric [15], while being relatively robust against imbalanced data sets [46]. The *F1 score* is a specification of the *Fβ score*, which adds a weighting between *Precision* and *Sensitivity*, or more specifically a weighting between FP and FN samples. All of the metrics presented so far are bounded between 0 and 1 with 1 representing a perfect value and 0 the worst possible prediction of this metric. However, all of them rely on the definition of the positive class, which may be straightforward in some cases but can be based on a rather arbitrary choice in others. Notably, metric values may be completely different depending on the choice of positive class.

To overcome the need for selecting one class as the positive class, other metrics have been suggested that can be based on all entries of a multi-class confusion matrix, in which each class is assigned a row and a column of the matrix. The *Accuracy* is one of the most commonly used metrics and measures the ratio between all correct predictions (TP and TN) and the total number of samples. *Accuracy* is not robust against imbalanced data sets (see Figure 18), and is therefore often replaced by the more robust *Balanced Accuracy* that averages the *Sensitivity* over all classes [14]. The *Matthews Correlation Coefficient (MCC)*, also known as *Phi Coefficient*, measures the correlation between the actual and predicted class. The metric is bounded between -1 and 1, with high positive values referring to a good prediction which can only be achieved when all cardinalities are good, i.e. with a low number of FP/FN and high TP/TN. Another popular metric is *Cohen's Kappa* κ, which calculates the agreement between the reference and prediction while incorporating information on the agreement by chance. It is therefore a form of chance-corrected *Accuracy*. Similarly to *MCC*, it incorporates all values of the confusion matrix and is bounded between -1 and 1. In contrast to *MCC*, negative values do not indicate anti-correlation, but less agreement than expected by chance. *Cohen's Kappa* κ can be generalized by introducing a weighting scheme for the cardinalities in the *Weighted Cohen's Kappa* κ metric. For those three metrics, a value of 0 refers to a prediction which is not better than random guessing. All of the presented binary single-threshold counting metrics can be transferred to the multi-class case [14, 17], where *MCC* and *(Weighted) Cohen's Kappa* κ have explicit definitions, whereas the others become the implicit result of an aggregation across a rotating one-versus-the-rest binary perspective for each of the classes.

Multi-threshold metrics

The classical single-threshold counting metrics presented above rely on fixed thresholds to be set on the predicted class probabilities (if available), resulting in them being based on the cardinalities of the confusion matrix. **Multi-threshold metrics** overcome this limitation by calculating metric scores based on multiple thresholds. For instance, to emphasize how well a prediction distinguishes between the positive and negative class, the *AUROC* can be utilized. The *Receiver Operating Characteristic (ROC)* curve plots the *FPR*, which is equal to $1 - Specificity$, against the *Sensitivity* for multiple thresholds of the predicted class probabilities, contrarily to just choosing one fixed threshold. For computation of the ROC curve, the class scores can be ordered in descending order and each score regarded as a potential threshold. For each threshold, the resulting *Sensitivity* and *Specificity* are computed, and the resulting tuple is added to the *ROC* curve as one point (cf. Figure 5); note that the lower the threshold, the higher the *Sensitivity* but the lower (potentially) the *Specificity*. This leads to a monotonic increase of the curve. To interpolate between all points, meaning to approximate the values between the calculated *Sensitivity* and *Specificity* tuples, a simple linear interpolation can be employed by drawing a line between each pair of points [11]. An optimal classifier would lead to *Sensitivity* and *Specificity* of 1 (1-*Specificity* of 0), therefore corresponding to a single point $(0, 1)$ on the *ROC* curve. In contrast, a classifier with no skill level (random guessing) would result in a diagonal line from $(0, 0)$ to $(1, 1)$ (dashed line in Figure 5). The area under the *ROC* curve is referred to as *AUROC*, also called *AUC ROC* or simply *AUC*.

AUROC comes with two advantages: threshold and scale invariance. *AUROC* measures the quality of the predictions regardless of the threshold, as it is calculated over a number of thresholds. Furthermore, *AUROC* does not focus on the absolute values of predictions, but rather on how well they are ranked. However, those properties are not always desired. If a specific penalization of FP or FN is desired (cf. Figure 20), *AUROC* is not the best metric choice as it is invariant to the threshold. If the predicted class probabilities are desired to be well calibrated, the scale invariance feature will prevent from doing so.

Per definition, *AUROC* measures the complete area under the *ROC* curve. If only a specific range is of interest, a partial or ranged *AUROC* can also be computed [29]. Similarly, metrics can be assessed at a certain point of the *ROC* curve, for example the *Sensitivity* value at a specific score of the *Specificity* (e.g. 0.9), also referred to as *Sensitivity@Specificity*. This approach can similarly be used for other curve measures, e.g. the *Precision-Recall (PR)* curve, introduced in Sec. 3.3.

Common binary classification metrics: Single-threshold counting-based (part 1)

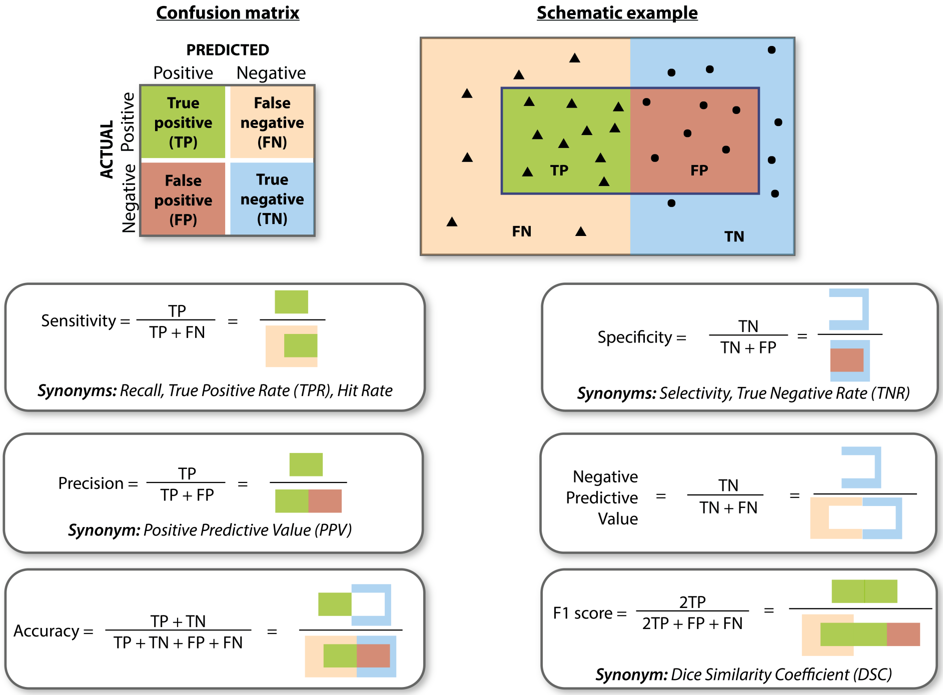
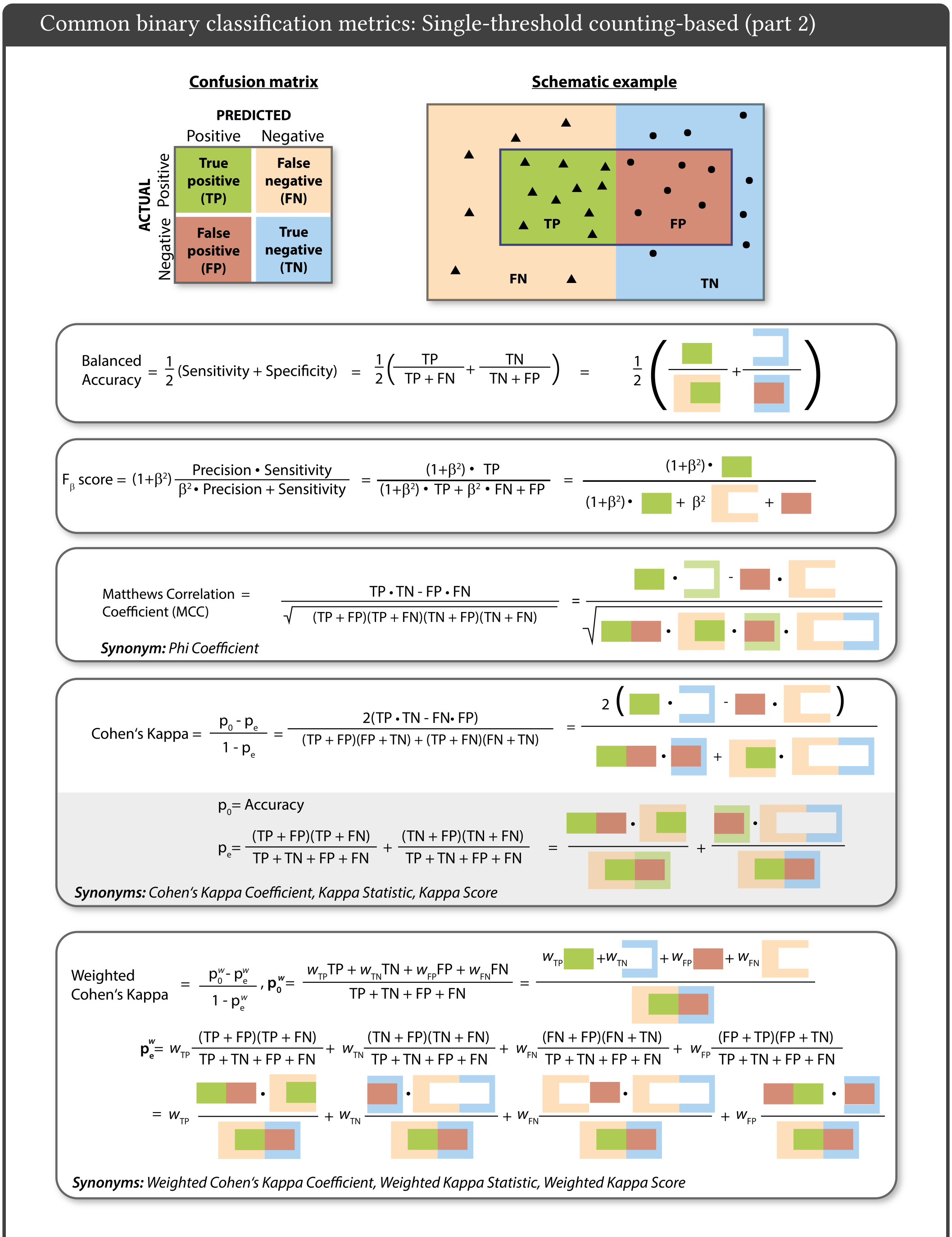

Confusion matrix
PREDICTED
Positive
Negative
ACTUAL
Positive
Negative
True positive (TP)
False negative (FN)
False positive (FP)
True negative (TN)
Schematic example
TP
FP
FN
TN
Sensitivity = TP / TP + FN =
Synonyms: Recall, True Positive Rate (TPR), Hit Rate
Specificity = TN / TN + FP =
Synonyms: Selectivity, True Negative Rate (TNR)
Precision = TP / TP + FP =
Synonym: Positive Predictive Value (PPV)
Negative Predictive Value = TN / TN + FN =
Accuracy = TP + TN / TP + TN + FP + FN =
F1 score = 2TP / 2TP + FP + FN =
Synonym: Dice Similarity Coefficient (DSC)

Fig. 3. Overview of the most commonly used binary classification measures that are based on the cardinalities of the confusion matrix, i.e. the true (T)/false (F) positives (P)/negatives (N).



Balanced Accuracy

$$\text{Balanced Accuracy} = \frac{1}{2}(\text{Sensitivity} + \text{Specificity}) = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

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

$$\text{Matthews Correlation Coefficient (MCC)} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Synonym: *Phi Coefficient*

$$\text{Cohen's Kappa} = \frac{p_0 - p_e}{1 - p_e} = \frac{2(TP \cdot TN - FN \cdot FP)}{(TP + FP)(FP + TN) + (TP + FN)(FN + TN)}$$

$$p_0 = \text{Accuracy}$$

$$p_e = \frac{(TP + FP)(TP + FN)}{TP + TN + FP + FN} + \frac{(TN + FP)(TN + FN)}{TP + TN + FP + FN}$$

Synonyms: *Cohen's Kappa Coefficient, Kappa Statistic, Kappa Score*

$$\text{Weighted Cohen's Kappa} = \frac{p_0^w - p_e^w}{1 - p_e^w},\ p_0^w = \frac{w_{TP}TP + w_{TN}TN + w_{FP}FP + w_{FN}FN}{TP + TN + FP + FN}$$

$$p_e^w = w_{TP}\frac{(TP + FP)(TP + FN)}{TP + TN + FP + FN} + w_{TN}\frac{(TN + FP)(TN + FN)}{TP + TN + FP + FN} + w_{FN}\frac{(FN + FP)(FN + TN)}{TP + TN + FP + FN} + w_{FP}\frac{(FP + TP)(FP + TN)}{TP + TN + FP + FN}$$

Synonyms: *Weighted Cohen's Kappa Coefficient, Weighted Kappa Statistic, Weighted Kappa Score*

Fig. 4. Overview of the most commonly used binary classification measures that are based on the cardinalities of the confusion matrix, i.e. the true (T)/false (F) positives (P)/negatives (N).

Common classification metrics: Multi-threshold-based

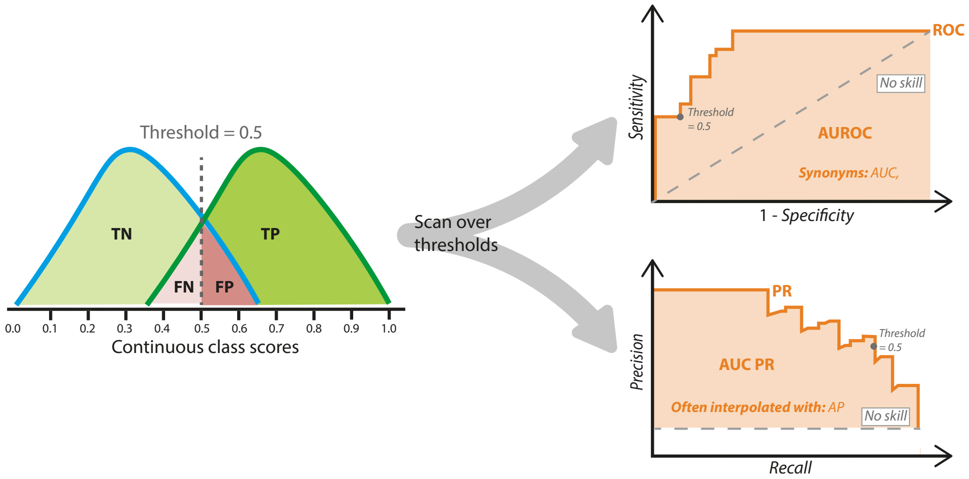


Fig. 5. Principle of multi-threshold metrics. Rather than being based on a static threshold (e.g. for generating the confusion matrix), multi-threshold-based metrics integrate over a range of thresholds. Prominent examples are the *Area under the Receiver Operating Characteristic curve (AUROC)* (also known as *Area under the curve (AUC)* or *AUC Receiver Operating Characteristic (ROC)*) and the Area under the *Precision-Recall (PR)* curve (*AUC PR*). Cardinalities, i.e. the true (T)/false (F) positives (P)/negatives (N), are computed based on a threshold (e.g. 0.5) of predicted class probabilities (left). Based on those values, *Sensitivity* and *1 - Specificity*/*Precision* are calculated and plotted against each other (right). The procedure is repeated for several thresholds, resulting in the *ROC*/*PR* curve. The area under the *ROC*/*PR* curve is referred to as *AUROC*/*AUC PR*. The latter is often interpolated by the *Average Precision (AP)* metric as detailed in Figure 11. The dashed gray lines refer to a classifier with no skill level (random guessing).

3.2 Semantic Segmentation

Semantic segmentation Semantic segmentation is commonly defined as the process of partitioning an image into multiple segments/regions. To this end, one or multiple labels are assigned to every pixel such that pixels with the same label share certain characteristics. Semantic segmentation can therefore also be regarded as pixel-level classification. As in image-classification problems, predicted class probabilities are typically calculated for each pixel deciding on the class affiliation based on a threshold over the class scores [2]. In semantic segmentation problems, the pixel-level classification is typically followed by a post-processing step, in which connected components are defined as objects, and object boundaries are created accordingly. Semantic segmentation metrics can roughly be classified into three classes: (1) single-threshold counting metrics or overlap-based metrics, for measuring the overlap between the reference annotation and the prediction of the algorithm, (2) distance-based metrics, for measuring the distance between object boundaries, and (3) problem-specific metrics, measuring, for example, the volume of objects.

Single-threshold counting metrics

The most frequently used segmentation metrics are **single-threshold counting metrics**. In the context of segmentation they are also referred to as **overlap metrics**, as they essentially measure the overlap between a reference mask and the algorithm prediction. According to a comprehensive analysis of biomedical image analysis challenges [30], the *DSC* [12] is the by far most widely used metric in the field of medical image analysis. As illustrated in Figure 6, it yields a value between 0 (no overlap) and 1 (full overlap). The *DSC* is identical to the *F1 score* and closely related to the *IoU*, which is identical to the *Jaccard Index*:

$$IoU = \frac{DSC}{2 - DSC} \tag{1}$$

$$DSC = \frac{2IoU}{1 + IoU} \tag{2}$$

Distance-based metrics

Overlap-based metrics are often complemented by **distance-based metrics** that operate exclusively on the TP and compute one or several distances between the reference and the prediction. Apart from a few exceptions, distance-based metrics are often **boundary-based metrics** which focus on assessing the accuracy of object boundaries. According to [30], the *HD* and its 95% percentile variant (*Hausdorff Distance 95% percentile (HD95)*) [18] are the most commonly used boundary-based metrics. 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* [53]. The *HD95* calculates the 95% percentile instead of the maximum, therefore disregarding outliers. Another popular metric is the *Average Symmetric Surface Distance (ASSD)*, measuring the average of all distances for every point from one object to the other and vice versa [51, 53] (see Figure 7). For the *HD(95)* and *ASSD* metrics, a value of 0 refers to a perfect prediction (distance of 0 to the reference boundary), while there exists no fixed upper bound.

A major problem related to boundary-based metrics are the error-prone reference annotations (see Figures 32 and 33). In fact, domain experts often disagree on the definition and annotation of objects and their boundaries [22]. While the *HD(95)* and *ASSD* are not robust with respect to uncertain reference annotations, the *NSD* was explicitly designed for this purpose as a hybrid metric between boundary-based and counting-based approaches. Known uncertainties in the reference as well as acceptable deviations of the predicted boundary from the reference are captured by a threshold τ [38], as shown in Figure 8. Only boundary parts within the border regions defined by τ are counted as TP. The metric is bounded between 0 (no boundary overlap) and 1 (full boundary overlap), so that it can be interpreted similarly to the classical *DSC* (though restricted to the boundary). Please note that τ is another important hyperparameter which should be chosen wisely, based on inter-rater agreement, for example. Another option for addressing inter-rater variability is the *Boundary IoU* (cf. Figure 9). It measures the overlap of boundaries while capturing tolerable uncertainties with a distance parameter d (see Sec. 3.3 for details) and is similarly bounded between 0 and 1.

Problem-specific segmentation metrics

While overlap-based metrics and distance-based metrics are the standard metrics used by the general computer vision community, biomedical applications often have special domain-specific requirements. In medical imaging, for example, the actual volume of an object may be of particular interest (for example tumor volume). In this case, **volume metrics** like the *Absolute* or *Relative Volume Error* and the *Symmetric Relative Volume Difference* can be computed [35]. However, they are less common than overlap metrics, as the location of objects is not considered at all (see Figure 29). If the structure center or center line is of particular interest (e.g. in cells or vessels) **connectivity metrics** come into play, which measure the agreement of the center line between two objects. This is of special interest if linear or tube-like objects are present in a data set. For this purpose, the center line *Centerline Dice Similarity Coefficient (clDice)* [44] has been designed.

Common segmentation metrics: Overlap-based

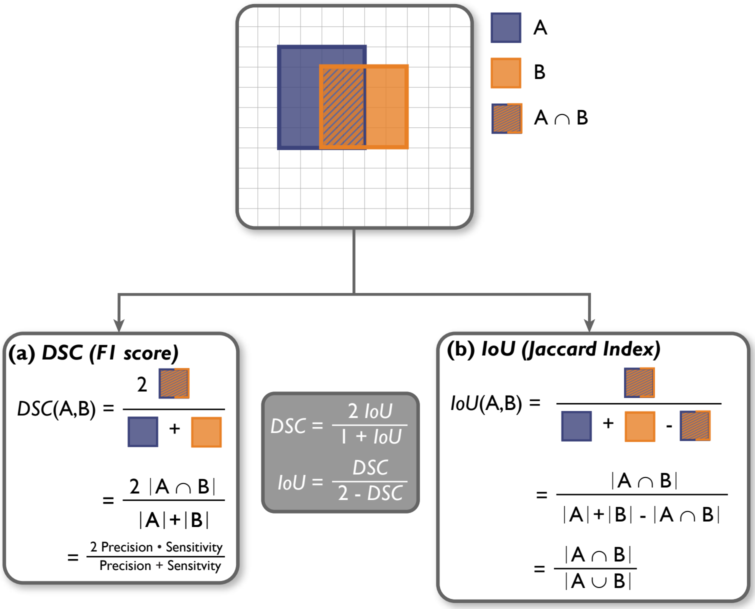


Fig. 6. Most commonly used overlap-based segmentation metrics: **(a)** the *Dice Similarity Coefficient (DSC)* and **(b)** the *Intersection over Union (IoU)*, with $|A|$ denoting the cardinality of set A, $A \cap B$ the intersection between sets A and B, and $A \cup B$ the union of sets A and B. The *DSC* can furthermore be computed as the harmonic mean of *Precision* and *Sensitivity*. Note that the *DSC* is equivalent to the *F1 score* and the *IoU* is equivalent to the *Jaccard index.*

Common segmentation metrics: Boundary-based

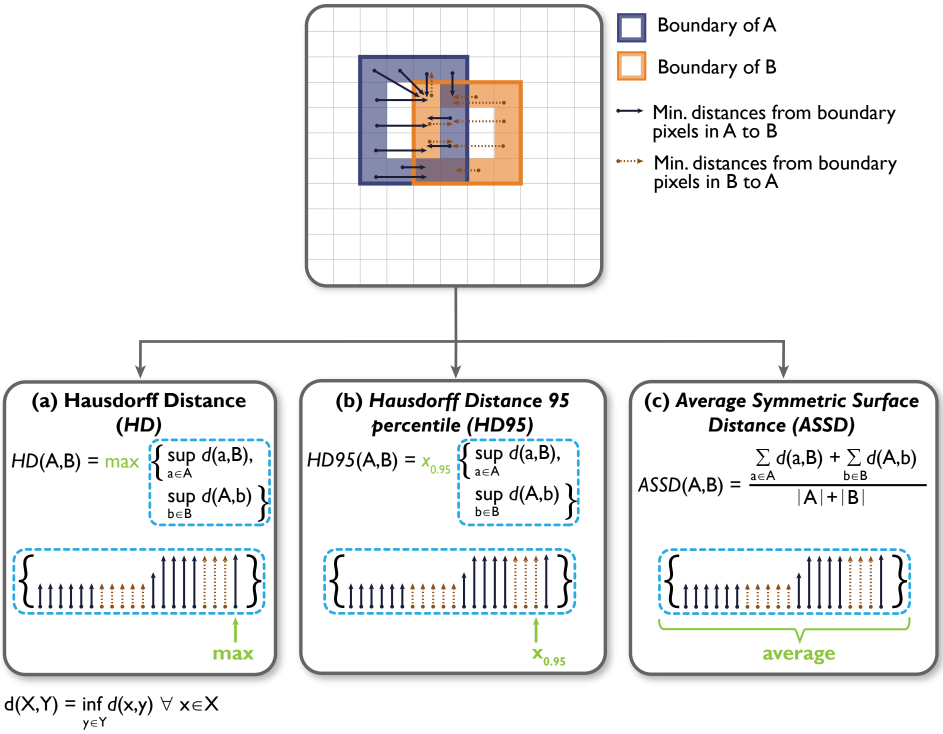

Boundary of A
Boundary of B
Min. distances from boundary pixels in A to B
Min. distances from boundary pixels in B to A
(a) Hausdorff Distance (HD)
$HD(A,B) = \max \left\{ \sup_{a \in A} d(a,B), \sup_{b \in B} d(A,b) \right\}$
max
(b) Hausdorff Distance 95 percentile (HD95)
$HD95(A,B) = x_{0.95} \left\{ \sup_{a \in A} d(a,B), \sup_{b \in B} d(A,b) \right\}$
$x_{0.95}$
(c) Average Symmetric Surface Distance (ASSD)
$ASSD(A,B) = \frac{\sum_{a \in A} d(a,B) + \sum_{b \in B} d(A,b)}{|A|+|B|}$
average
$d(X,Y) = \inf_{y \in Y} d(x,y) \ \forall \ x \in X$

Fig. 7. Most commonly used distance-based segmentation metrics: **(a)** the *Hausdorff Distance (HD)*, **(b)** the 95% percentile (denoted as x_{95}) of the *HD*, *Hausdorff Distance 95% percentile (HD95)* and **(c)** the *Average Symmetric Surface Distance (ASSD)*, with $d(x, y)$ denoting the Euclidean distance between boundary pixels x and y. Only True Positive (TP) are considered.

Common segmentation metrics: Normalized Surface Distance (NSD)

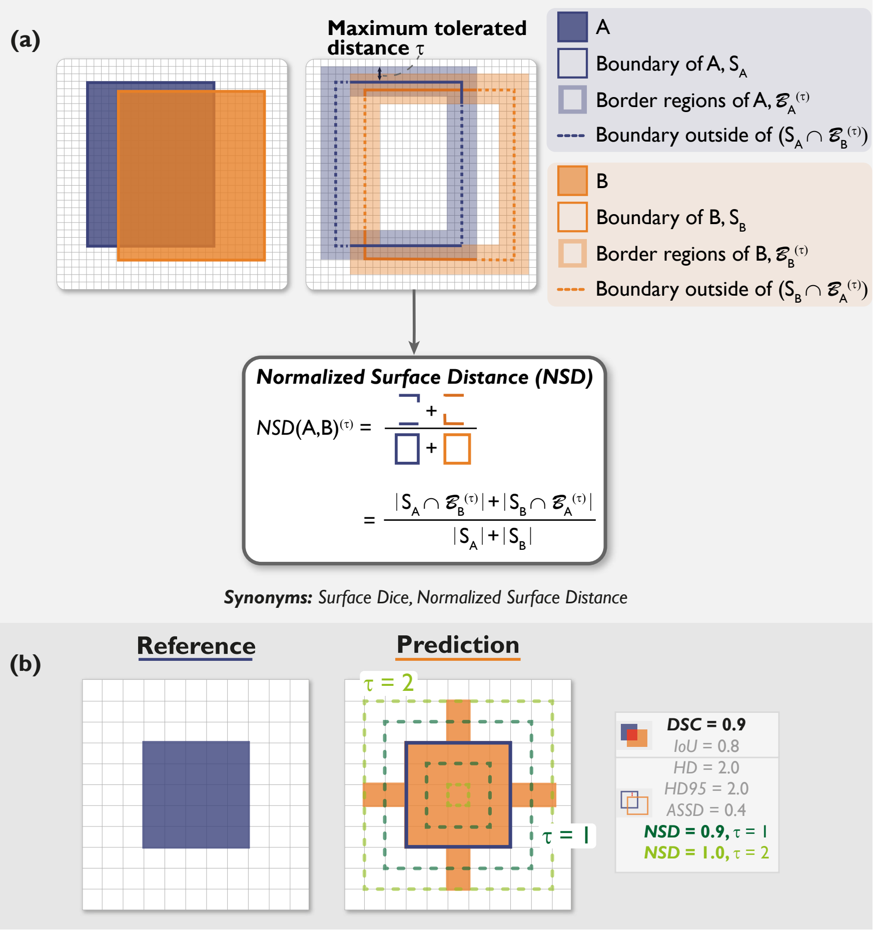

(a)
Maximum tolerated distance τ
A
Boundary of A, S_A
Border regions of A, $\mathcal{B}_A^{(\tau)}$
Boundary outside of ($S_A \cap \mathcal{B}_B^{(\tau)}$)
B
Boundary of B, S_B
Border regions of B, $\mathcal{B}_B^{(\tau)}$
Boundary outside of ($S_B \cap \mathcal{B}_A^{(\tau)}$)
Normalized Surface Distance (NSD)
NSD(A,B)$^{(\tau)}$ =
+
+
$= \frac{|S_A \cap \mathcal{B}_B^{(\tau)}| + |S_B \cap \mathcal{B}_A^{(\tau)}|}{|S_A| + |S_B|}$
Synonyms: Surface Dice, Normalized Surface Distance
(b)
Reference
Prediction
τ = 2
τ = 1
DSC = 0.9
IoU = 0.8
HD = 2.0
HD95 = 2.0
ASSD = 0.4
NSD = 0.9, τ = 1
NSD = 1.0, τ = 2

Fig. 8. **(a)** The *Normalized Surface Distance (NSD)* is an **uncertainty-aware** segmentation metric that measures the overlap between two boundaries. The parameter τ represents the tolerated difference between the prediction and the reference boundary S and defines the border regions $\mathcal{B}^{(\tau)}$ for each structure, i.e. the pixels within the range of τ from the boundary. They are defined as all pixels within distance τ from the boundary S. The threshold can be based on the domain-related requirements and/or the inter-rater variability, for example. **(b)** Example showing how the *NSD* can handle outliers in the prediction by adjusting the tolerance value to $\tau = 2$ pixels. The *Dice Similarity Coefficient (DSC)* and other metrics (*Intersection over Union (IoU)*, *Hausdorff Distance (HD)(95)*, *Average Symmetric Surface Distance (ASSD)*), in contrast, penalize these tolerated errors.

3.3 Object Detection

Object detection refers to the detection of one or multiple objects (or: instances) of a particular class (e.g. lesion) in an image [28]. The following description assumes single-class problems, but translation to multi-class problems is straightforward, as validation for multiple classes on object-level is performed individually per class. Notably, as multiple predictions and reference instances may be present in one image, the predictions need to include localization information, such that a matching between reference and predicted objects can be performed. Important design choices with respect to the validation of object detection methods include:

(1) *How to represent an object?* Representation is typically composed of location information and a class affiliation. The former may take the form of a bounding box (i.e. a list of coordinates), a pixel mask, or the object's center point. Additionally, modern algorithms typically assign a confidence value to each object, representing the probability of a prediction corresponding to an actual object of the respective class. Note that a confusion matrix is later computed for a fixed threshold on the predicted class probabilities.[7]
(2) *How to decide whether a reference instance was correctly detected?* This step is achieved by applying the *localization criterion*. This may, for example, be based on comparing the object centers of the reference and prediction or computing their overlap (Figures 38 and 39).
(3) *How to resolve assignment ambiguities?* The above step might lead to ambiguous matchings, such as two predictions being assigned to the same reference object. Several strategies exist for resolving such cases.

The following sections provide details on (1) applying the localization criterion, (2) applying the assignment strategy and (3) computing the actual performance metrics.

Localization criterion
As one image may contain multiple objects or no object at all, the **localization criterion** or **hit criterion** measures the (spatial) similarity between a prediction (represented by a bounding box, pixel mask, center point or similar) and a reference object. It defines whether the prediction *hit/detected* (TP) or *missed* (FP) the reference. Any reference object not detected by the algorithm is defined as FN. Please note that TN are not defined for object detection tasks, which has several implications on the applicable metrics, as detailed below.

There are multiple ways to define the localization or hit criterion (see Figures 6, 9 and 38). Popular center-based localization criteria are (a) *the center-cover criterion*, for which the reference object is considered hit if the center of the reference object is inside the predicted detection, (b) the *distance-based hit criterion*, which considers a TP if the distance d between the center of the reference and the detected object is smaller than a certain threshold τ and (c) the *center-hit criterion*, which holds true if the center of the predicted object is inside the reference bounding box or mask.

The most commonly used overlap-based hit criterion is determined by computing the *IoU* [19] (cf. Figure 6b). The prediction is considered as TP if the overlap is larger than a certain threshold (e.g. 0.3 or 0.5) and as FP otherwise. If bounding boxes are considered, the *IoU* is computed between the reference and predicted bounding boxes (*Box IoU*). For more fine-grained annotations in form of a pixel mask, the *IoU* may be computed for the complete mask (*Mask IoU*). The *Mask IoU* is less sensitive to structure boundary quality in larger objects (cf. Section 6). This is due to the fact that boundary pixels will increase linearly while pixels inside the structure will increase quadratically

[7] Please note that we will use the term confidence scores analogously to predicted class probabilities in the context of object detection and instance segmentation.

with an increase in structure size. The *Boundary IoU* measures the *IoU* of two structures for mask pixels within a certain distance d from the structure boundaries [8], as illustrated in Figure 9.

Common localization criterion: Boundary Intersection over Union (Boundary IoU)

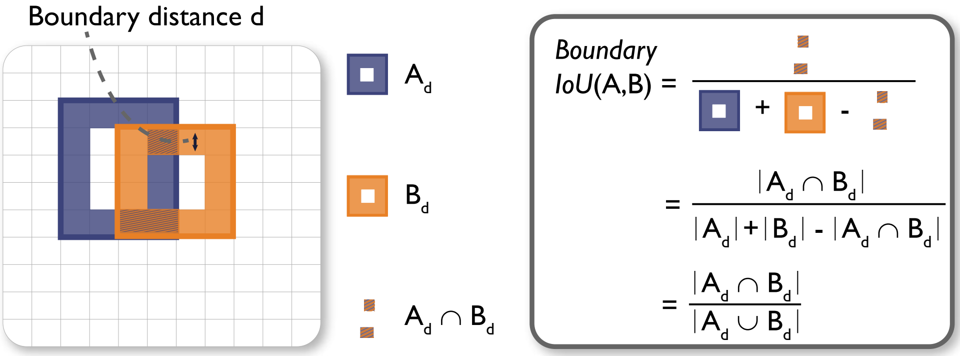

Boundary distance d
A_d
B_d
$A_d \cap B_d$
Boundary IoU(A,B) =
+
-
$= \frac{|A_d \cap B_d|}{|A_d| + |B_d| - |A_d \cap B_d|}$
$= \frac{|A_d \cap B_d|}{|A_d \cup B_d|}$

Fig. 9. The *Boundary Intersection over Union (Boundary IoU)* is an **uncertainty-aware** metric that measures the overlap between two boundaries. The overlap is computed for the structures A and B for mask pixels within a certain distance d from the structure boundaries (A_d and B_d), by computing the *IoU* between A_d and B_d.

The localization criterion should be carefully chosen according to the underlying motivation and research question and depending on the available coarseness of annotations. However, it should be noted that annotations of a lower resolution will result in an information loss, as illustrated in Figure 10. For example, the *Box IoU* is sometimes used although pixel-mask annotations are available because algorithms are expected to output rough localization in the shape of boxes. Such a simplification might cause problems if structures are not well-approximated by a box shape, or if structures can overlap causing multi-component masks (cf. Section 7, Figure 40). Lastly, it should be noted that the decision for a cutoff value on the localization criterion leads to instabilities in the validation (e.g. see Figure 39). For this reason, it is common practice in the computer vision community to average metrics over multiple cutoff values (default for *IoU* criteria: 0.5, 0.6, 0.7, 0.8, 0.9). Generally speaking, the cutoff values should be chosen according to the driving biomedical question. For example, if a particular interest lays on the exact outlines, higher thresholds should be chosen. On the other hand, for noisy reference standards, a low cutoff value is preferable.

Discarding information of provided annotations

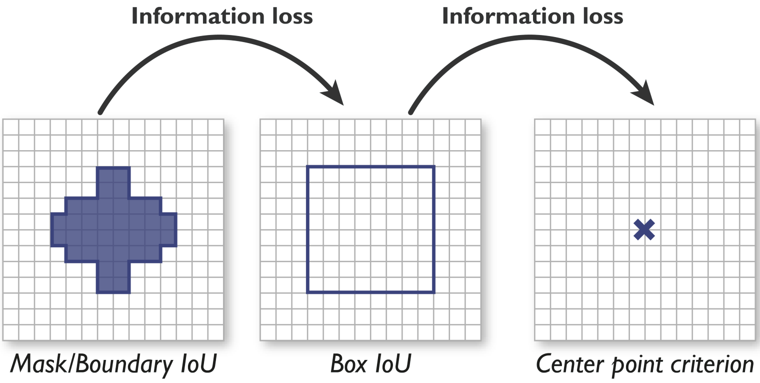


Fig. 10. Selection of a localization criterion that discards spatial information should be well motivated by the given task.

Assignment strategy
The localization criterion alone is not sufficient to extract the final confusion matrix based on a fixed threshold for the predicted class probabilities (confidence scores), as ambiguities can occur. For example, two predictions may have been assigned to the same reference object in the localization step, or vice versa. These ambiguities need to be resolved in a further **assignment step**.

This assignment and thus the resolving of potential assignment ambiguities can be done via different strategies. The most common strategy in the computer vision community is the *Greedy by Score* strategy. All predictions in an image are ranked by their predicted class probability and iteratively (starting with the highest probability) assigned to the reference object with the highest localization criterion for this prediction. The selected reference object is subsequently removed from the process since it can not be matched to any other prediction (unless double assignments are allowed). The *Hungarian Matching* [27] is associated with a cost function, usually depending on the localization criterion, which is minimized to find the optimal assignment of predictions and reference. In the biomedical domain, more sophisticated matching strategies are often avoided by setting the localization criterion threshold to $IoU > 0.5$ and only allowing non-overlapping object predictions (which inherently avoids matching conflicts). In the case of a high ratio of touching reference objects and common non-split errors, meaning that one prediction overlaps with multiple reference objects, the Intersection over Reference (IoR) [32] might be considered as an alternative to IoU [32].

Metric computation

Similar to image-level classification and semantic segmentation algorithms, object detection algorithms are commonly assessed with single-threshold counting metrics, assuming a fixed confusion matrix, (cf. Figures 3 and 4). However, one of the most popular object detection metrics is the multi-threshold metric *Average Precision (AP)* [28], which is the area under the *Precision-Recall (PR)* curve for a certain interpolation scheme. The *PR* curve is computed similarly to the *ROC* curve by scanning over confidence thresholds and computing the *Precision* and *Recall (Sensitivity)* for every threshold (cf. Figure 5). Note in this context that the popular *ROC* curve is not applicable in object detection tasks because TN are not available. Also, while the ROC curve is monotonically rising, this behaviour may not be expected from the *PR* curve, which typically features a zigzag shape, as illustrated in Figure 11. Specifically "as the level of *Recall* varies, the *Precision* does not necessarily change linearly due to the fact that FP replaces FN in the denominator of the *Precision* metric." [11]. A linear interpolation would therefore be overly optimistic, which is why more complex interpolation is needed, as detailed in [11].

The area under the *PR* curve is typically calculated as the *AP* implying a conservative simplification of curve interpolation,

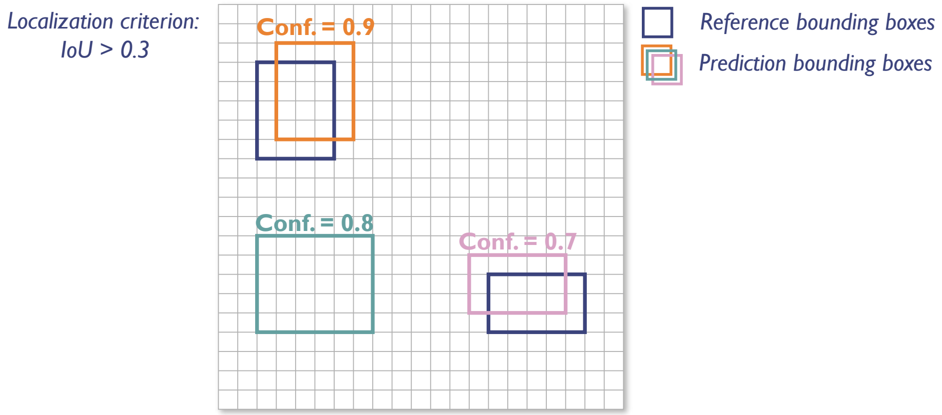
$$AP = \sum_{n} (R_n - R_{n-1}) P_n, \tag{3}$$

with R_n and P_n denoting the *Recall* and *Precision* at the nth threshold[8] (cf. dashed gray line in Figure 11). For the *PR* curve, an optimal model would lead to *Recall (Sensitivity)* and *Precision* of 1, therefore being the point $(1, 1)$ on the *PR* curve. Conversely, a model with no skill level (random guessing) would result in a horizontal line with a precision proportional to the portion of positive samples (dashed line in Figure 5). For computation of the metric (for a given class), the predictions are sorted in descending order of the confidence for each prediction (for that class). For each possible confidence threshold, the cardinalities are computed and the resulting tuple of *Recall (Sensitivity)* and *Precision* is added to the curve (cf. Figure 11).

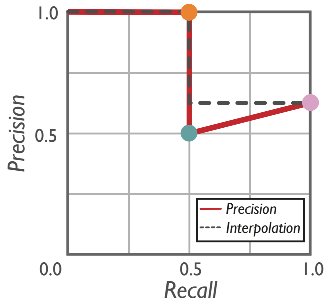
In contrast to drawing the *PR* curve and computing the *AP*, *Free-Response Receiver Operating Characteristic (FROC)* curve is often favoured in the clinical context due to its easier interpretability. It operates at object level and plots the average number of *FPPI* (in contrast to the *FPR*) against the *Sensitivity* [7]. The area under the *FROC* curve, however, is not bounded between 0 and 1 and the employed *FPPI* scores vary across studies, such that there exists no standardized definition of an area under the respective curve. Overall, 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.

[8] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html

Common object detection metrics: Average Precision (AP)



Rank	Conf.	IoU	TP / FP?	Cumulative TP	Cumulative FP	Recall	Precision
1	0.9	0.43	TP	1	0	0.50	1.00
2	0.8	0.00	FP	1	1	0.50	0.50
3	0.7	0.36	TP	2	1	1.00	0.67



$$AP = \sum_n (R_n - R_{n-1}) P_n \approx 0.83$$

Fig. 11. Exemplary computation of the *Average Precision (AP)* in object detection tasks. Predictions are ranked according to their predicted class probabilities, represented by the confidence (conf.). Based on the *Intersection over Union (IoU)* or a similar localization/hit criterion, it is determined whether the prediction is a True Positive (TP) or False Positive (FP) (here: *IoU* > 0.3). For the creation of the *Precision-Recall (PR)* curve, *Precision* and *Recall* are computed for the accumulated TP and FP for every confidence score (conf.). The *AP* interpolates the points of the *PR* curve as shown by the dashed gray line.

In image-level classification problems, validation is naturally performed on the entire data set, while segmentation typically relies on computing metrics for each image and then aggregating metric values. This latter approach is not applicable in object detection in a straightforward manner because of the relatively small amount of samples per image (typically a few objects rather than thousands of pixels). Figure 12 illustrates the per-image and the per-data set validation of objects. In the per-image aggregation approach, special care needs to be taken in the case of an empty reference or prediction, as detailed in Sec. 7 (Figure 43).

Validating objects per data set *vs.* validating objects per image

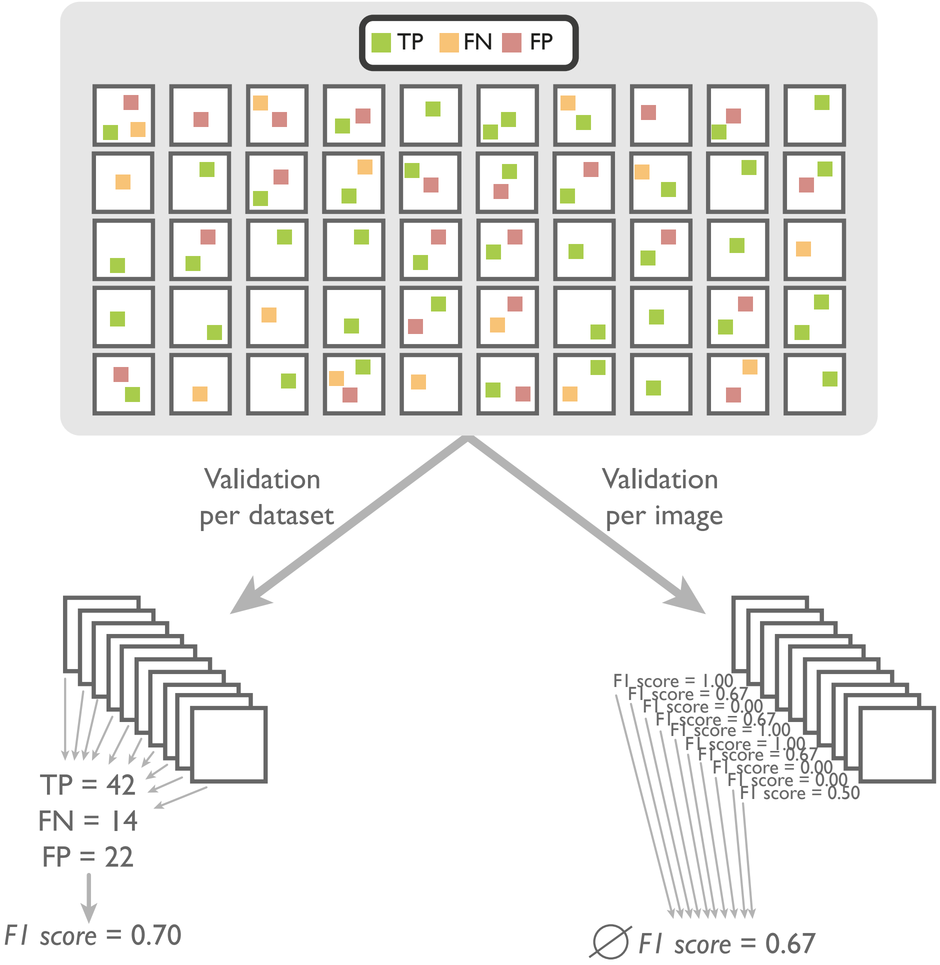


Fig. 12. Validation on object-level can be performed per data set (left) or per image (right). For the per-data set validation of objects, the cardinalities are calculated over the whole data set. For the per-image validation of objects, metric scores are computed per image and aggregated afterwards. ∅ refers to the average *F1 score*.

3.4 Instance Segmentation

In contrast to semantic segmentation, instance segmentation problems distinguish different instances of the same class (e.g. different lesions). Similar as in object detection problems, the task is to detect individual instances of the same class, but detection performance is measured by pixel-level correspondences (as in semantic segmentation problems). Optionally, instances can be applied to one of multiple classes. Validation metrics in instance segmentation problems often combine common detection metrics with segmentation metrics applied per instance. It should be noted that instance segmentation problems are often phrased as semantic segmentation problems with an additional post-processing step, such as connected component analysis [41]. In practice, predicted class probabilities, yielded by modern segmentation algorithms, are often discarded in the post-processing step and are thus not available for subsequent validation. Figure 13 illustrates how to overcome this potential problem.

Retaining class scores when extracting instances from semantic segmentation output

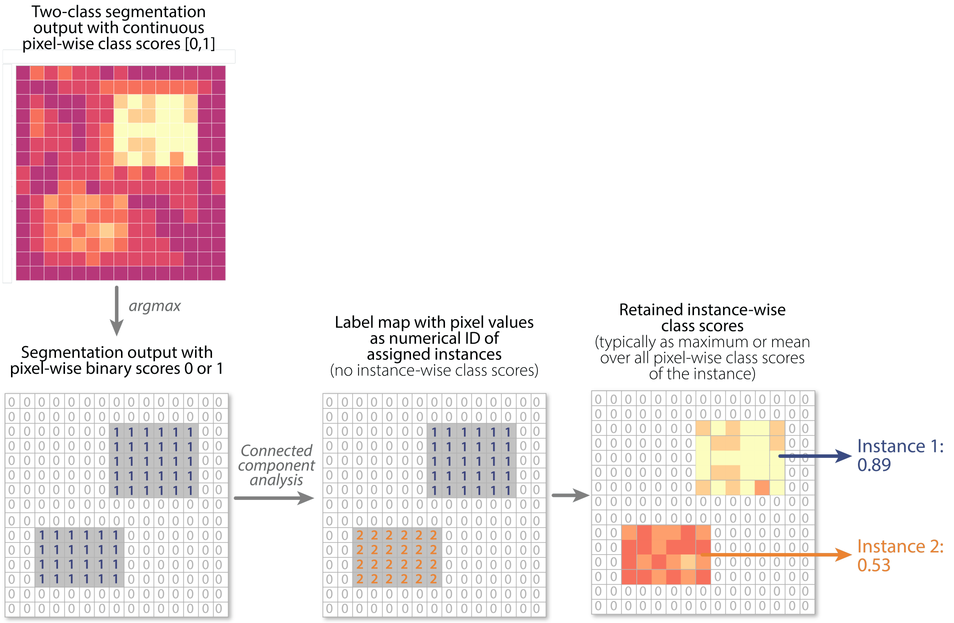

Two-class segmentation output with continuous pixel-wise class scores [0,1]
argmax
Segmentation output with pixel-wise binary scores 0 or 1
Connected component analysis
Label map with pixel values as numerical ID of assigned instances
(no instance-wise class scores)
Retained instance-wise class scores
(typically as maximum or mean over all pixel-wise class scores of the instance)
Instance 1: 0.89
Instance 2: 0.53

Fig. 13. Retaining class scores when extracting instances from semantic segmentation output. Often, semantic segmentation methods are used for instance segmentation problems by post-processing the segmentation output (e.g. via connected component analysis). Predicted class probabilities are often lost during this process and therefore not available for further validation processes. However, instance-wise class scores can be retained by taking the maximum or average over all pixel-wise class scores of every instance.

4 PITFALLS DUE TO CATEGORY-METRIC MISMATCH

Performance metrics are typically expected to reflect a domain-specific validation goal (e.g. clinical goal). Previous research, however, suggests, that this is often not the case [43]. Before choosing validation metrics, the correct problem category needs to be defined. In the following, we will describe pitfalls related to metrics not being applied to the appropriate problem category.

Mismatch semantic segmentation ↔ object detection. A common problem is that segmentation metrics, such as the *DSC*, are applied to *object detection* tasks [6, 21], as illustrated in Figure 14. From a clinical perspective, for example, the algorithm producing *Prediction 2* and covering all three structures of interest (e.g. tumors) would be clinically much more valuable compared to the one producing a highly accurate segmentation for one structure but missing the other two in *Prediction 1*. This is not reflected in the metric values, which are substantially higher for *Prediction 1*. In general, the *DSC* is strongly biased against single objects, therefore not appropriate for the detection of multiple structures [24, 53].

Pitfall: Mismatch semantic segmentation ↔ object detection

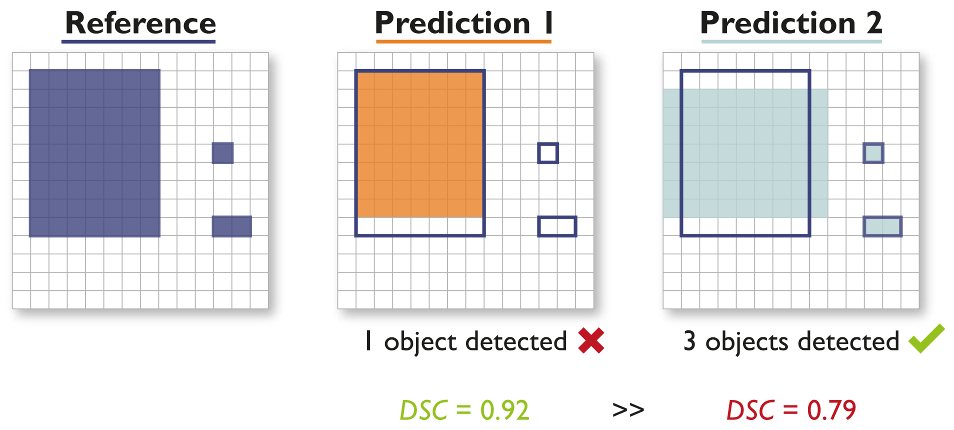


Fig. 14. Effect of using a segmentation metric for object detection. In this example, the prediction of an algorithm only detecting one of three structures (*Prediction 1*) leads to a higher *Dice Similarity Coefficient (DSC)* compared to that of another algorithm (*Prediction 2*) detecting all structures.

Mismatch semantic ↔ instance segmentation. In segmentation problems, the driving research question should decide whether semantic or instance segmentation should be chosen for validation. This is particularly relevant when multiple objects within one image overlap or touch, as often occurring in cell images. For semantic segmentation problems, overlapping or touching objects may end up merged into a single object without clear boundaries or distinction between the single objects. Instance segmentation problems, on the other hand, ensure that the borders of touching or overlapping structures can be accurately assigned and that objects can be differentiated. If instance segmentation is preferred, the labels need to be chosen accordingly. An example is shown in Figure 15: The desired annotation consists of two different instances, but only semantic labels are available (middle). A prediction will only be as accurate as the reference, hence detecting only one instance but yielding a perfect metric score although the desired task is not solved.

Problem: Mismatch semantic ↔ instance segmentation

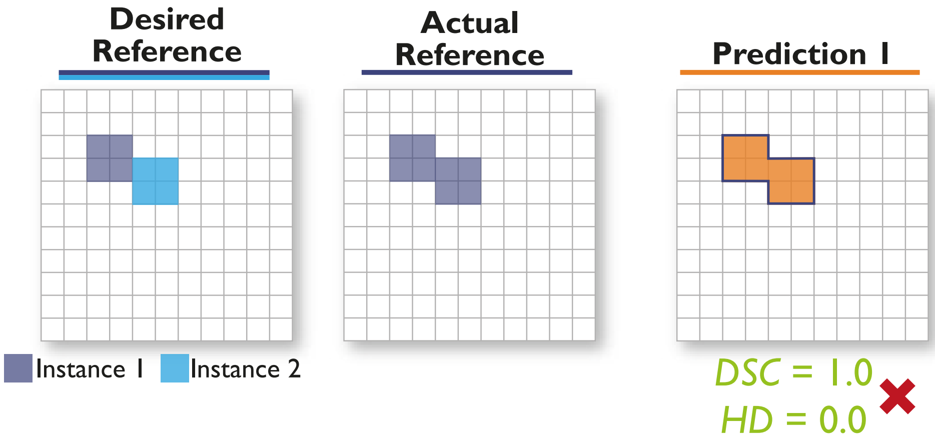


Fig. 15. Effect of using a semantic segmentation metric for instance segmentation. When only the presence of a certain class (e.g. tumor) is labeled, and instance information is not provided in the reference, a common means for assessing performance is to regard each connected component as one instance. This may yield misleading results.

Mismatch image-level classification ↔ object detection. Tasks that should be validated at image level are sometimes erroneously approached with object detection models instead of image-level classification models [20]. Object detection models are designed to handle different objects in an image rather than the complete image and will naturally introduce problems in a validation setting on image level. Object detection tasks are dependent on choosing a proper localization criterion, which is not needed for an image-level classification problem. For example, a *ROC* curve, typically used for assessing the performance of image-level classification algorithms, does not consider the localization step needed in object detection tasks, as it was designed to validate at image rather than object level. It therefore does not take into account whether a detected object is at the correct location in the image. Moreover, when validation on image level is conducted by using an object detection model, the detection with the largest class probability (confidence score) of all detections in one image is usually taken, neglecting all other predictions. This does not capture the performance of the model accurately. Figure 16 illustrates some of the resulting problems:

(1) The image-level *ROC* curve **does not measure the localization performance**. As can be seen from Figure 16a, the validation is done per image, not per object, therefore not considering whether an object is actually hit (see *Prediction 2*).
(2) The image-level *ROC* curve is **invariant to the number of annotated objects**. As can be seen from Figure 16b, the curve can not discriminate between a model detecting all objects in an image (*Prediction 1*) or just detecting one object (*Prediction 2*), as long as the largest score is the same across predictions.
(3) The image-level *ROC* curve is **invariant to the number of detected objects**. As can be seen from Figure 16c, the curve can not discriminate between a model detecting many FP objects in an image (*Prediction 2*) or only detecting one FP (*Prediction 1*), as long as the largest score is the same.

Pitfall: Mismatch image-level classification ↔ object detection

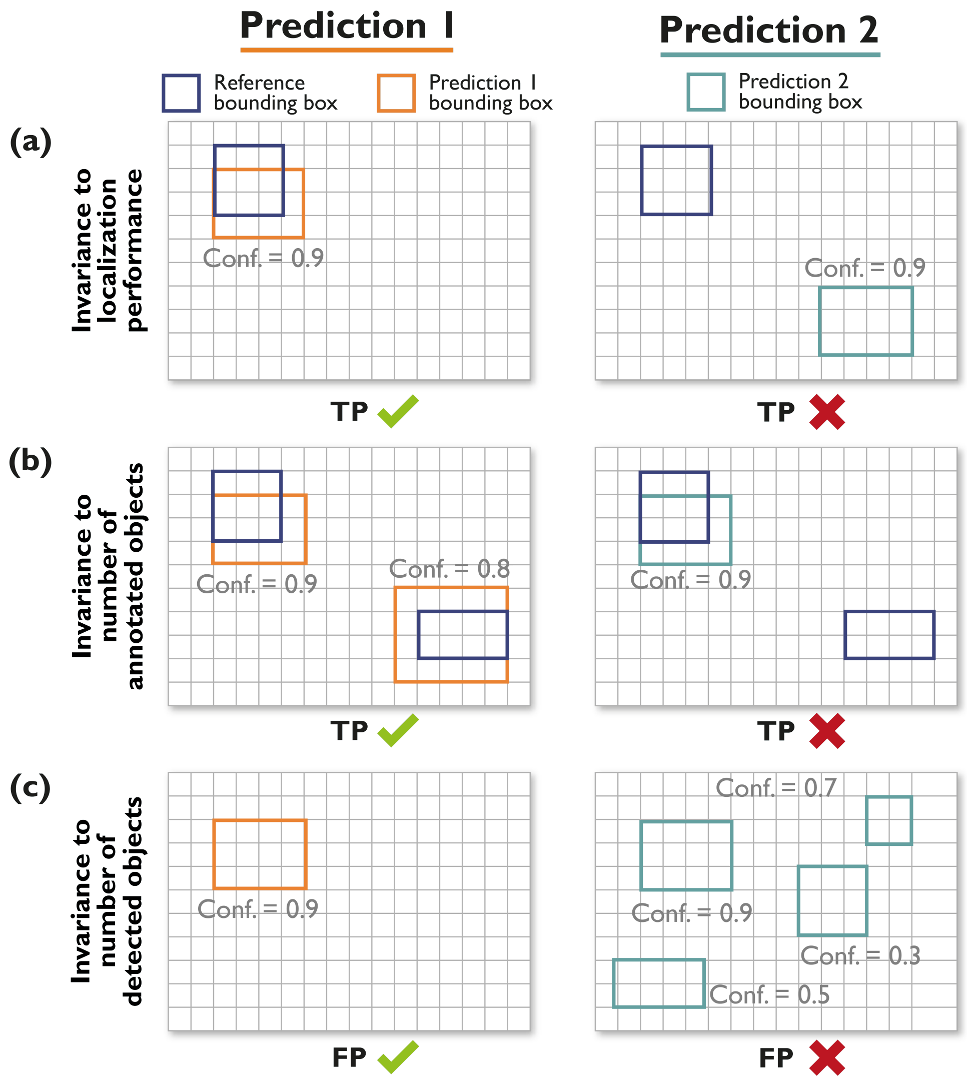


Fig. 16. When image-level classification metrics like the area under the *Receiver Operating Characteristic (ROC)* curve are used to validate object detection models, (1) information on the object matching (localization, number of objects etc.) is lost and typically (2) only one detection per image (the one with the highest confidence) is considered. This leads to several problems: **(a)** The image-level *ROC* curve does not measure the localization performance. Both *Prediction 1* and *2* are considered as True Positive (TP) due to their score being very high, although *Prediction 2* is not hitting the annotated object. **(b)** The image-level *ROC* is invariant to the number of annotated objects in an image. The curve does not discriminate between a model detecting all positives (*Prediction 1*) and a model detecting only one of the positives (*Prediction 2*), as long as the maximum score is the same. **(c)** The image-level *ROC* is invariant to the number of detections in an image. The curve does not discriminate between a model with many False Positive (FP) (*Prediction 2*) and a model with just one FP (*Prediction 1*), as long as the maximum score is the same. The class probabilities are represented by confidence scores (conf.).

No matching problem category. Metrics should reflect a domain-specific validation goal. This goal may not align with commonly used technical measures like the *DSC*. Figure 17 shows an example with the property of interest being the accuracy of the ratio between two structure volumes, indicating, for example, the percentage of blood volume ejected in each cardiac cycle [3]. Both predictions will result in similar averaged *DSC* scores, although the ratio of the volumes vastly differs. A common segmentation metric thus does not reflect the actual research question in this case.

Pitfall: Metrics may be poor proxies for computing properties of interest

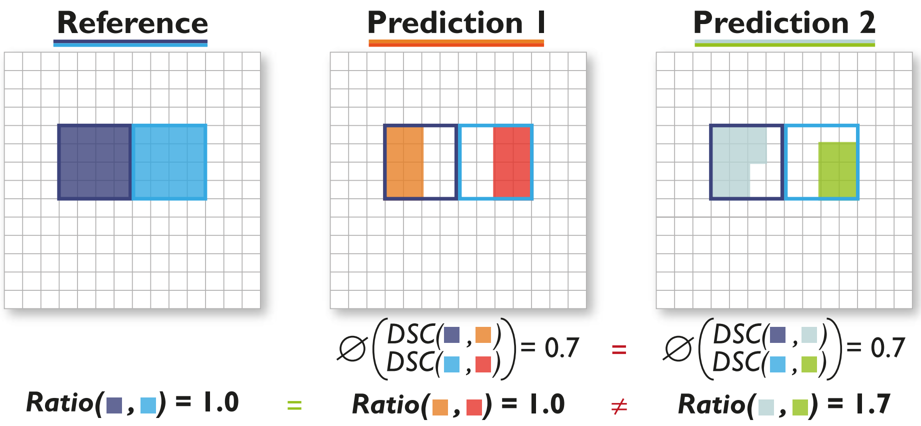


Fig. 17. It may not always be possible to find a common metric that ideally captures the domain interest. In this example, accuracy of the ratio between two volumes is the property of interest. *Predictions 1* and *2* result in similar averaged *Dice Similarity Coefficient (DSC)* metric values although both predictions result in a different ratio between structure volumes, which is the parameter of interest.

5 PITFALLS RELATED TO IMAGE-LEVEL CLASSIFICATION

Most issues related to classification metrics are related to one of the following properties of the underlying biomedical problem:

- High class imbalance (Figure 18)
- Presence of more than two classes (Figure 19)
- Unequal importance of classes (Figure 20)
- Interdependencies between classes (Figure 21)
- Lack of stratification (Figure 22)
- Missing prevalence correction (Figure 23)

Furthermore, metric-specific limitations may arise (Figure 24) Please note that all of these also apply to semantic/instance segmentation or object detection problems, as summarized in Table 1. The discourse focuses on the most commonly used image-level classification metrics, as presented in Figures 3, 4 and 5. For most of the problems, it focuses on the *Accuracy*, *Sensitivity* or *Precision*, because those are the most common metrics for image-level classification in biomedical image analysis. Please note that we do not recommend their indiscriminate use, as they come with limitations (discussed in the following paragraphs), but rather wish to spotlight the problems and pitfalls of those most commonly used metrics.

To preserve the clarity of the illustrations, the most important of the presented metric values may be highlighted with color. Green metric values correspond to a "good" metric value (e.g. a high *Sensitivity* score), whereas red values correspond to a "bad" value (e.g. a low *Sensitivity*). Green check marks indicate desirable behaviour of metrics, red crosses indicate undesirable behaviour. Please note that a low metric value is not automatically a "bad" score. A metric value should always be put into perspective and compared to inter-rater variability. For simplicity, we still use the terms "good" and "bad/poor" throughout the section. Finally, our illustrations do not provide the concrete class probabilities of the presented classifiers.

High class imbalance. *Accuracy* is one of the commonly applied metrics in classification problems, presumably because it is particularly straightforward to interpret. However, the metric is not designed to handle imbalanced data sets, which often occur across all domains. Figure 18 provides an example in which the positive class (orange circle) is heavily underrepresented. While *Prediction 1* gives a reasonable separation of the classes, *Prediction 2* results in the same *Accuracy* value (0.97) although the algorithm only provides the majority vote as a result. In this specific example, *Sensitivity*, *Precision* and *F1 score* reveal the issue, as does *Matthews correlation coefficient (MCC)*, a metric designed to handle class imbalance which reflects that *Prediction 2* is not better than a random guess (0.00) [9]. As many classification measures are easily computable using the number of TP, TN, FP and FN samples, it is highly recommended to report these TP, TN, FP and FN values explicitly and then compute multiple metrics [15].

Plotting the *ROC* and *PR* curves (Figure 18b and c) also reveal the limitations of *Prediction 2*, which yields an *AUROC* of 0.52 and *AP* of 0.04, indicating that the prediction is not better than random guessing.

Pitfall: Class imbalance

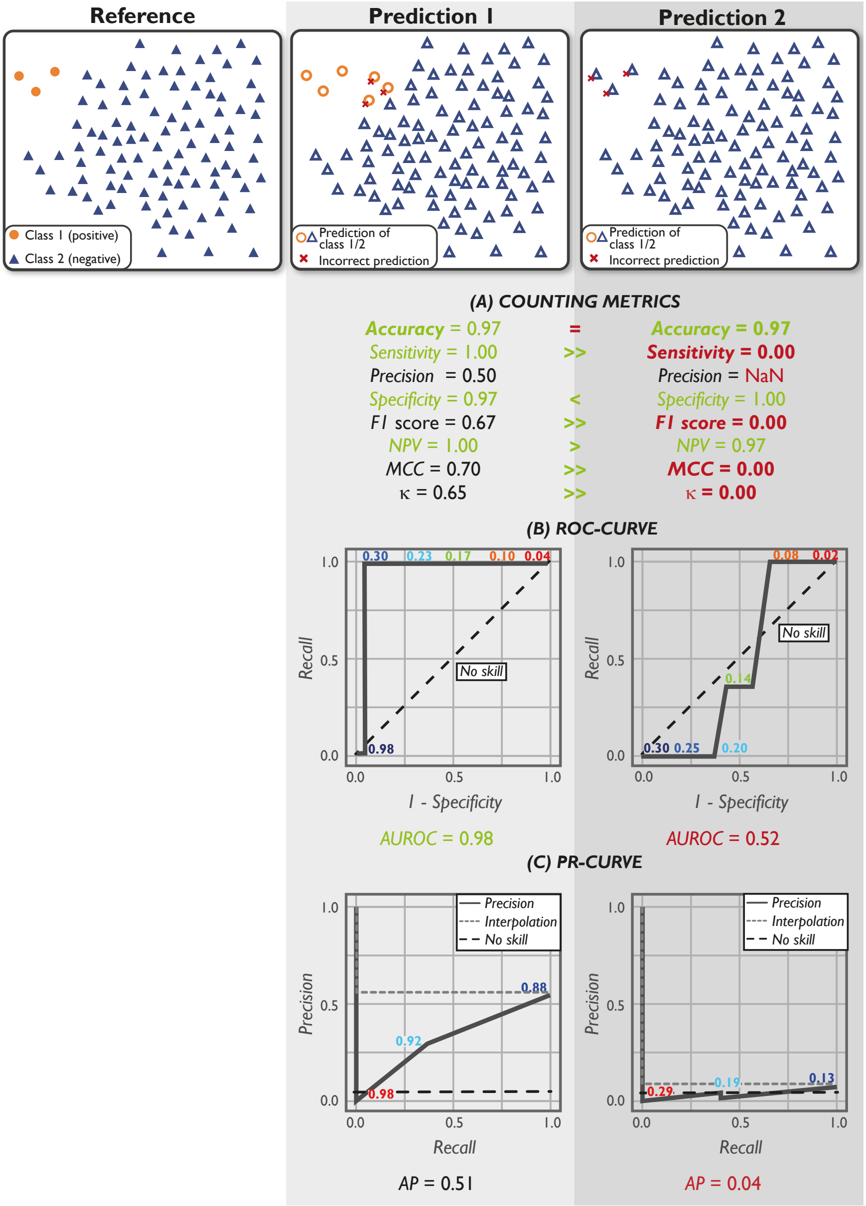
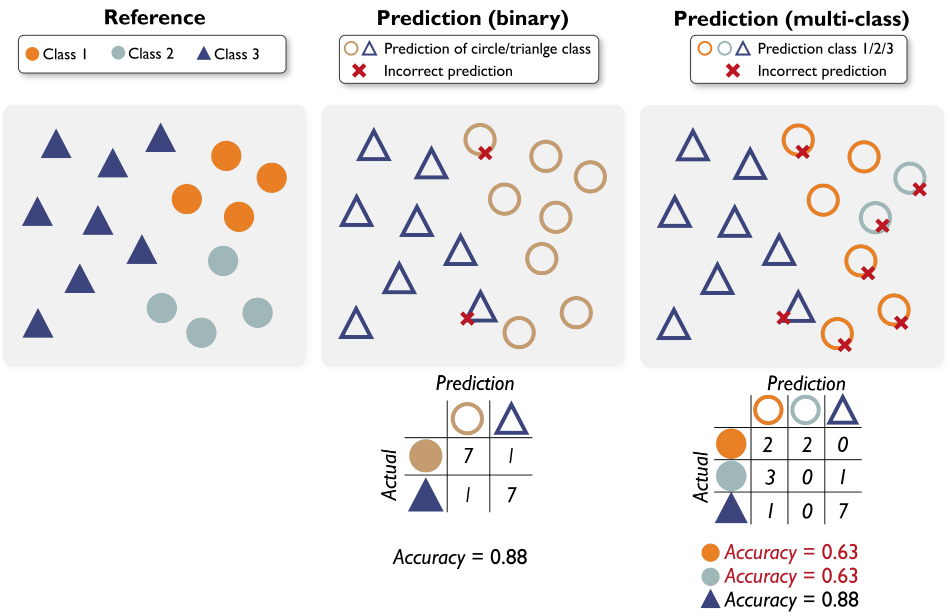


Fig. 18. Effect of class imbalance. Not every metric is designed to reflect class imbalance (e.g. *Accuracy*). In the case of underrepresented classes, such a metric yields a high value even if the classifier performs very poorly for one of the classes (here: *Prediction 2*). Multi-threshold metrics, such as the *Area under the Receiver Operating Characteristic curve (AUROC)* and the *Average Precision (AP)*, reveal the weakness, indicating that *Prediction 2* is not better than random guessing. For comparison, a no skill classifier (random guessing) is shown as a black dashed line. For the *Precision-Recall (PR) curves*, the interpolation applied to compute the *AP* metric is shown by a dashed grey line. Thresholds used for curve generation are provided as small numbers in the curve. Further abbreviations: *Negative Predictive Value (NPV)*, *Matthews Correlation Coefficient (MCC)*, *Cohen's Kappa κ*.

More than two classes available. Many binary metrics can directly be translated to the multi-class case by expanding the confusion matrix to all classes. These classes are often hierarchically structured, for example in the shape of one negative class (e.g. no pathology) and multiple positive classes (e.g. different types of pathologies). Figure 19 shows an example of a classification into triangles and circles, for which the circle class is further separated into two distinct classes (green and orange). The binary performance into triangle *vs.* circle, shown in the middle, is good (*Accuracy* of 0.88). But when considering the three classes separately, the prediction struggles to identify the color of the circles, causing their per-class accuracy scores to drop significantly (0.63 each).

Pitfall: Hierarchical structure of classes



	○	△
●	7	1
▲	1	7

Accuracy = 0.88

	○	○	△
●	2	2	0
●	3	0	1
▲	1	0	7

● *Accuracy* = 0.63
● *Accuracy* = 0.63
▲ *Accuracy* = 0.88

Fig. 19. Classes in categorical classification may be hierarchically structured, for example in the shape of multiple positive classes and one negative class. The phrasing of the problem as binary *vs.* multi-class hugely affects the validation result. Binary classification (middle), differentiating triangles from circles, yields a good *Accuracy*, while per-class validation yields a poor score because the two circle classes cannot be distinguished well. Incorrect predictions are indicated by a red cross.

Unequal importance of classes. In biomedical applications, classes are often not equally important. Consider the task of colon polyp detection in the gastrointestinal tract, for example. To provide the patient with the best care, it is crucial to detect all of these precancerous lesions. This requires a particular penalization of those samples containing a polyp which have been marked as 'no polyp' (FN), and the metrics need to be chosen accordingly. The *Precision*, for example, does not include the FN in its definition, hence would not be appropriate for this research question. *Sensitivity*, on the other hand, would show the desired poor performance in the presence of many FN predictions, as seen in the top row of Figure 20a.

For image retrieval, the task of finding images for a specific content, it is not important to find every single existing image, but the images found should be correct. In this setting, the FP (assigning an incorrect image as correct) need to be penalized. Since it includes the computation of FP, in this case, the *Precision* would be a good metric. In contrast, *Sensitivity* does not consider FP, therefore being inappropriate in this context (see bottom row of Figure 20a). Penalization in both cases is especially important in cases of imbalanced data sets (see Figure 18).

When it comes to multi-class problems, different approaches may be chosen to compute the metric values. One possibility is to first compute the metric values per class and aggregate them subsequently. Special care has to be taken in the case of unequal importance of the different classes. For example, identifying whether a patient harbors a pathology in general might be more important than identifying the specific type of pathology. In this case, one should not just average over all class metric scores, but instead apply a sufficient weighting scheme. In the example of Figure 20b, the triangle class is the most important class but also the one with the lowest per-class *Accuracy*. Simple averaging, so-called macro-averaging, would ignore that property and thus result in a higher aggregated *Accuracy* than merited. This effect can be compensated with the *Weighted Accuracy*.

Pitfall: Unequal importance of classes or misclassifications

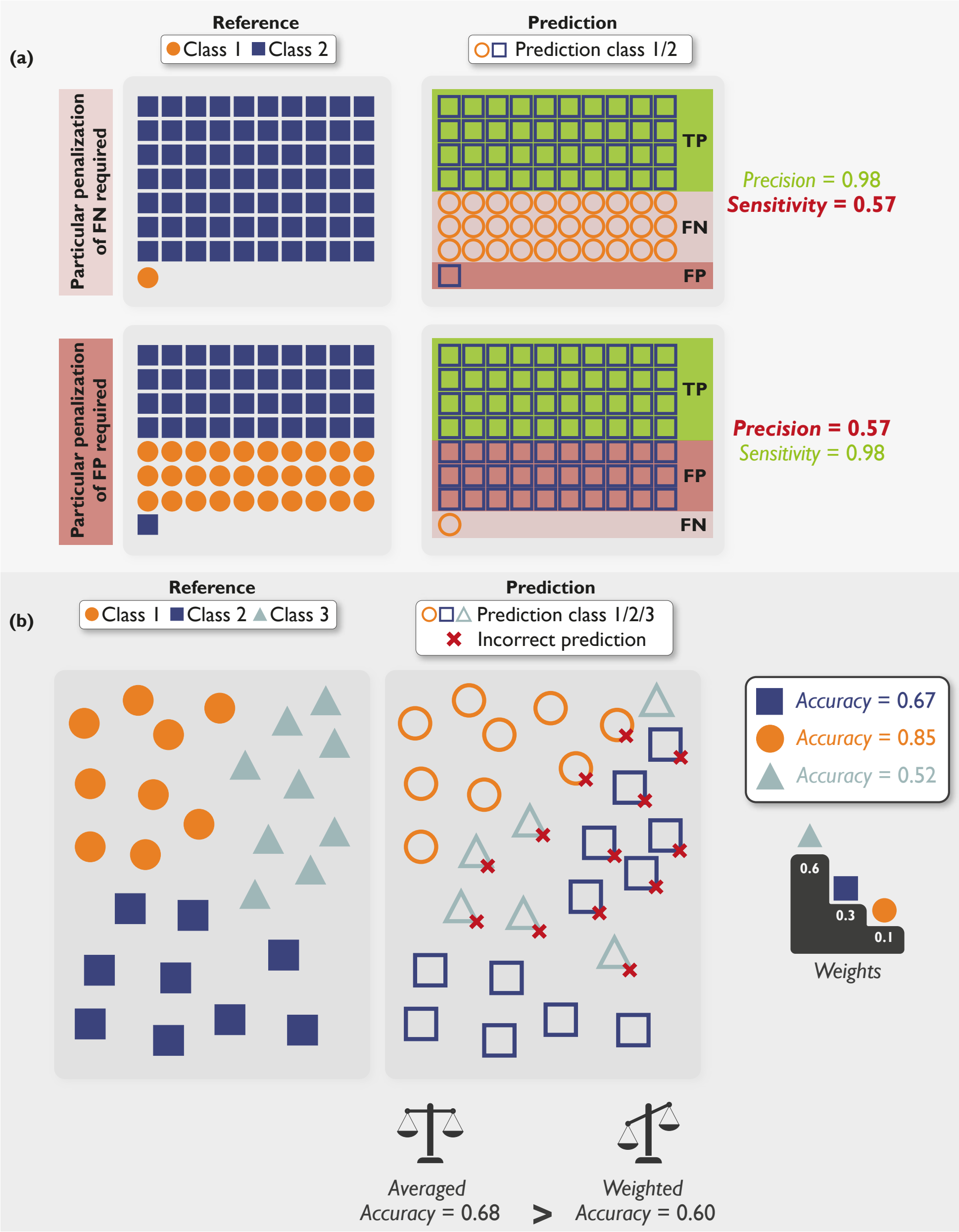


Fig. 20. Effect of unequal importance of classes. **(a)** Effect of using metrics that are not suitable for penalizing False Negative (FN)/False Positive (FP). The definition of the *Precision* metric does not incorporate FN and is therefore not well-suited for penalizing FN, as required in cancer screening tasks, for example. Analogously, the *Sensitivity* is not well-suited for penalizing FP, as required in many retrieval tasks, for example. **(b)** Simple averaging (macro-averaging) of the *Accuracy* ignores the unequal importance of classes, given by pre-defined weights of classes. Incorrect predictions are indicated by a red square.

Interdependencies between classes. If multiple classes are visible in the data set, one should carefully account for interdependencies between the classes. Interdependencies can happen in cases of multi-colinearities in which two classes are correlated, either inherently, such as for the body mass index (BMI) and the body fat percentage, or in the case of dependent data settings, for example multiple images per patient or the presence of confounders. An algorithm aiming to classify the dark blue triangle class in Figure 21 may result in a nearly perfect *Accuracy* of 0.94, but only because the dark blue triangle almost always appears in conjunction with the orange square. Computing the *Accuracy* for those images individually without the square class would lead to a much lower performance.

Pitfall: Interdependencies between classes

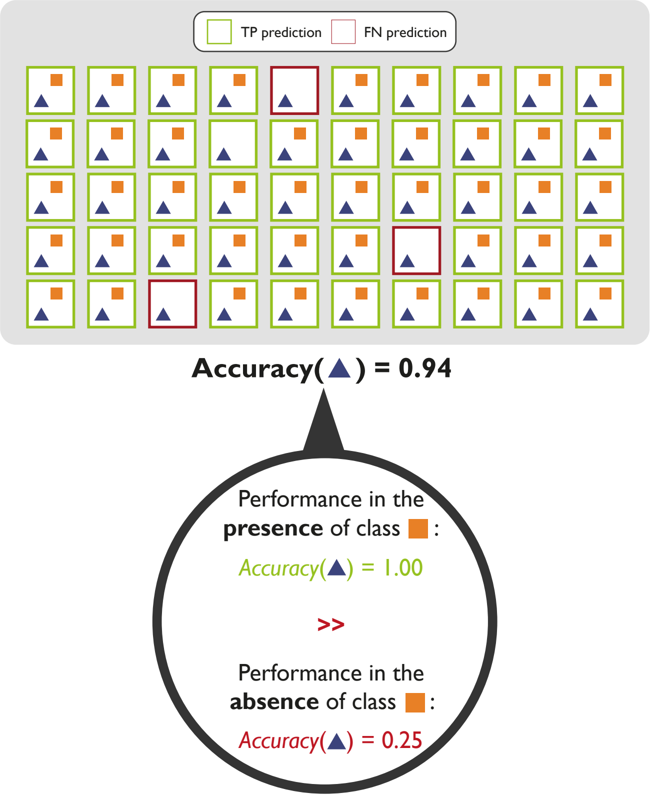


Fig. 21. Effect of interdependencies between classes. A prediction may show a near-perfect *Accuracy* score of 0.94 for the dark blue triangle as it frequently appears in conjunction with the orange square. By calculating the *Accuracy* in the *presence* and *absence* of the square class, it can be seen that the algorithm only works well in the presence of the orange class. Incorrect predictions are indicated by a red cross.

Stratification based on meta-information. Different kinds of meta-information may be available for a data set, including the presence and relevance of artifacts or artificial structures (e.g. metal artifacts in CT images or text overlay in endoscopic data) as well as specifics of acquisition protocols (e.g. acquisition angle or viewpoint) or grid size (cf. Figure 35). Another typical example is the gender of a patient, as shown in Figure 22. In this case, the *Accuracy* is computed over twelve cases, disregarding the available meta-information (gender). Stratification based on gender will reveal that the prediction performs much worse for women compared to men.

Pitfall: Meta-information ignored

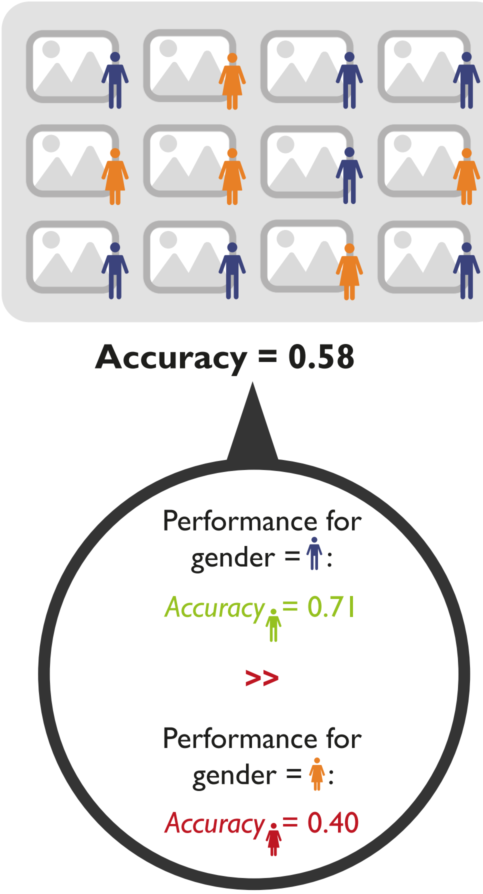


Fig. 22. Effect of disregarding relevant meta-information (here: gender). Ignoring the available meta-information of the patient's gender per image, the *Accuracy* does not reveal that the algorithm performs much better for men compared to women.

Missing prevalence correction. The *Positive Predictive Value (PPV)*, also known as *Precision*, and the *Negative Predictive Value (NPV)* are common measures to validate classification performances. In many cases, binary classification is considered, for example presence or absence of a disease. In contrast to *Sensitivity* and *Specificity*, in case-control studies, *PPV* and *NPV* should be seen as the conditional probability of a disease being present based on a test result and the prevalence in a general population.

$$PPV = \frac{Sensitivity \cdot Prevalence}{Sensitivity \cdot Prevalence + (1 - Specificity) \cdot (1 - Prevalence)} \tag{4}$$

$$NPV = \frac{Specificity \cdot (1 - Prevalence)}{Specificity \cdot (1 - Prevalence) + (1 - Sensitivity) \cdot Prevalence} \tag{5}$$

However, *PPV* and *NPV* are frequently used incorrectly. This is due to the fact that many practitioners assume the same prevalence in an analysed case-control study group as in the general population. However, a study group is often heavily biased, either due to the study design or due to the observation of patient groups from specialized clinics. Thus, the assumed disease prevalence in scientific literature is often higher than that found in the general population (cg. Figure 23). This problem is amplified by default implementations (e.g. in `scipy` [52]) which disregard wider population prevalence and calculate prevalence from the study group. Without prevalence correction, this can lead to misleading results, confusion among patients and ill-informed policy-making.

Pitfall: Missing prevalence correction

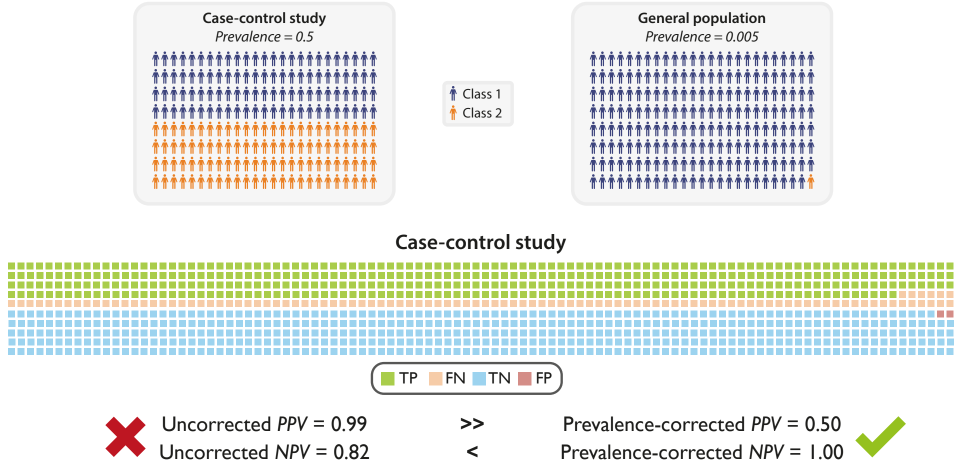


Fig. 23. Effect of missing prevalence correction. For case-control studies, the *Positive Predictive Value (PPV)* and *Negative Predictive Value (NPV)* are required to apply prevalence correction based on the general population. Falsely used case-control prevalence leads to incorrect metric scores compared to using the general population prevalence (which is often significantly lower).

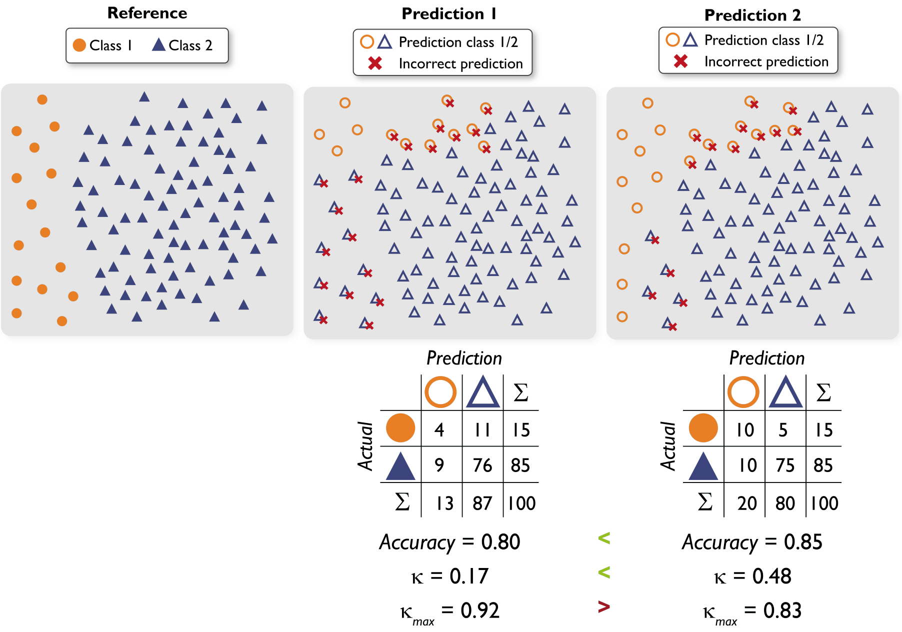
Upper bound in Cohen's κ ***calculation.*** *Cohen's* κ measures the agreement between ratings while incorporating information on the *Accuracy* by chance. It therefore investigates how well a prediction follows the distribution of the actual class. The maximum *Cohen's* κ helps interpreting the calculated κ score by symbolizing the corner case in which either the FP or FN are equal to 0 [49]:

$$\begin{aligned}\kappa_{max} &= \frac{p_{max} - p_e}{1 - p_e},\\ p_{max} &= \min\left(\frac{TP + FN}{TP + TN + FP + FN}, \frac{TP + FP}{TP + TN + FP + FN}\right)\\ &+ \min\left(\frac{TN + FN}{TP + TN + FP + FN}, \frac{TN + FP}{TP + TN + FP + FN}\right)\end{aligned} \tag{6}$$

The maximum *Cohen's* κ score will be lower as the distribution of the prediction and the actual classes diverge. This is shown in Figure 24 with two predictions. *Prediction 1* achieves lower *Accuracy* and *Cohen's* κ scores compared to *Prediction 2*, as it only predicts a very low number of TP. However, the predicted distribution in *Prediction 1* is more similar to the actual distribution (13 circle predictions *vs.* 15 actual circles and 87 triangle predictions *vs.* 85 actual triangles). The distribution of *Prediction 2* differs more from the actual distribution, yielding a lower *Cohen's* κ_{max} value[9].

[9] https://www.knime.com/blog/cohens-kappa-an-overview

Pitfall: Upper bound not equally obtainable in Cohen's κ



Prediction 1:

Actual	○	△	Σ
●	4	11	15
▲	9	76	85
Σ	13	87	100

Prediction 2:

Actual	○	△	Σ
●	10	5	15
▲	10	75	85
Σ	20	80	100

Prediction 1		Prediction 2
Accuracy = 0.80	<	*Accuracy* = 0.85
κ = 0.17	<	κ = 0.48
κ_{max} = 0.92	>	κ_{max} = 0.83

Fig. 24. Effect of different prediction distributions compared to target distribution. A prediction with a distribution similar to the actual distribution (*Prediction 1*) reaches higher maximum *Cohen's κ* values compared to a prediction with a dissimilar distribution (*Prediction 2*), although the overall *Accuracy* and *Cohen's κ* is lower. Incorrect predictions are indicated by a red cross.

6 PITFALLS RELATED TO SEGMENTATION

All pitfalls compiled for this work and relevant for semantic or instance segmentation are summarized in Table 1. This section focuses on limitations for semantic segmentation, but some of them are also transferable to other problem categories, as indicated in the table. Limitations of metrics are typically related to the following properties:

- Small size of structures relative to pixel size (Figure 25)
- High variability of structure sizes (Figure 26)
- Complex shapes of structures (Figure 27)
- Particular importance of structure volume (Figure 28)
- Particular importance of structure center (Figure 29)
- Particular importance of structure boundaries (Figure 30)
- Possibility of multiple labels per unit (Figure 31)
- High inter-rater variability (Figure 32)
- Possibility of outliers in reference annotation (Figure 33)
- Possibility of reference or prediction without the target structure (Figure 34)
- Preference for over- *vs.* undersegmentation (Figure 36)

Further pitfalls are related to technical peculiarities, such as the choice of global decision threshold for creating the confusion matrix (Figure 37) and the image resolution (Figure 35).

The limitations are presented for the most commonly used overlap segmentation metrics, namely *DSC*, *IoU*, and the most common boundary-based metrics, namely *HD*, *HD95*, *ASSD* and *NSD*. The *NSD* calculation is based on a user-defined threshold (cf. Figure 8). Results differ for different thresholds. Unless stated otherwise, we set the threshold to $\tau = 1$.

To preserve clarity of the illustrations, specific values may only be highlighted for one metric from each metric family, if the other metrics share similar properties (e.g. *DSC* and *IoU* share the same properties). Green metric values correspond to a "good" value (e.g. a high *DSC* or a low *HD* score), whereas red values correspond to a "bad" value (e.g. a low *DSC* or a high *HD* score). Green check marks indicate metric scores reflecting the research question, red crosses show those that do not. Please note that a low *DSC* value (or similar) is not automatically a "bad" score. A metric value should always be put into perspective and compared to inter-rater variability. We only use the terms "good" and "bad/poor" for simplicity.

Small size of structures relative to pixel size. Segmentation of small structures, such as brain lesions or cells imaged at low magnification, is essential for many image processing applications. In these cases, the *DSC* or *IoU* may not be appropriate metrics, as illustrated in Figure 25 (cf. [8]). In fact, a single-pixel difference between two predictions can have a large impact on the metric values. Given that the correct outlines (e.g. of pathologies) are often unknown and taking into account the potentially high inter-observer variability related to generating reference annotations [22], it is typically not desirable for few pixels to influence the metrics as much. This problem is particularly amplified in cases of large variability of structure sizes (cf. Figure 26). This pitfall also applies to object detection tasks. It should be noted that once a data set exclusively contains only very tiny structures, one may consider it an object detection rather than a segmentation problem.

Pitfall: Small structures

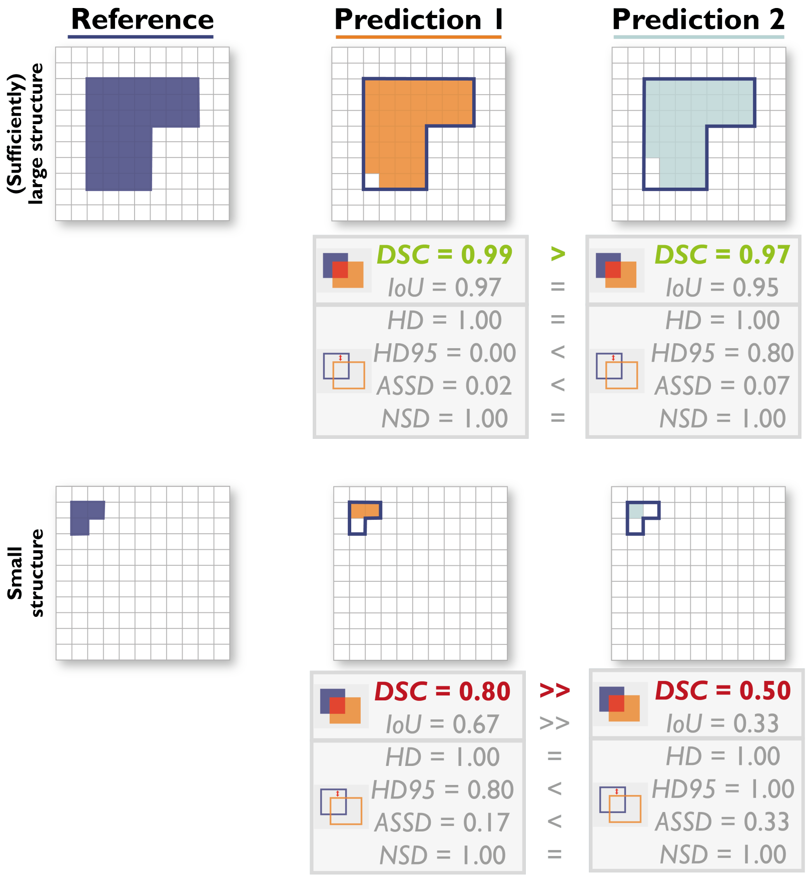


Fig. 25. Effect of structure size on the *Dice Similarity Coefficient (DSC)*. The predictions of two algorithms (*Prediction 1/2*) differ in only a single pixel. In the case of the small structure (bottom row), this has a substantial effect on the corresponding metric value (similar for the *Intersection over Union (IoU)*). The effects are considerably lower for the boundary-based metrics (*Hausdorff Distance (HD)*, *Hausdorff Distance 95% percentile (HD95)*, *Average Symmetric Surface Distance (ASSD)* and *Normalized Surface Distance (NSD)*.

High variability of structure sizes. The size of target structures may vary substantially, both within an image and across images. For example, in medical instrument segmentation in laparoscopic video data, an image frame may contain full-sized instruments as well as only the tip of an instrument just entering the scene [42]. In these cases, metrics need to be chosen carefully. As shown in the example above (Figure 25), metrics like the *DSC* or *IoU* are typically not well-suited for very small structures. Furthermore, size stratification – the aggregation of metric values for objects of similar sizes to uncover differences between them – should be employed. Figure 26 shows an exemplary data set of four images, containing three large structures and one small structure. When aggregating over all *DSC* values, the average *DSC* is 0.82. Computing the average for large and small structures separately, however, shows that the performance is much lower for the small structures compared to the large ones, demonstrating the large influence of the low metric values of small objects. This pitfall also applies to object detection tasks and other metrics.

Pitfall: High variability of structure sizes

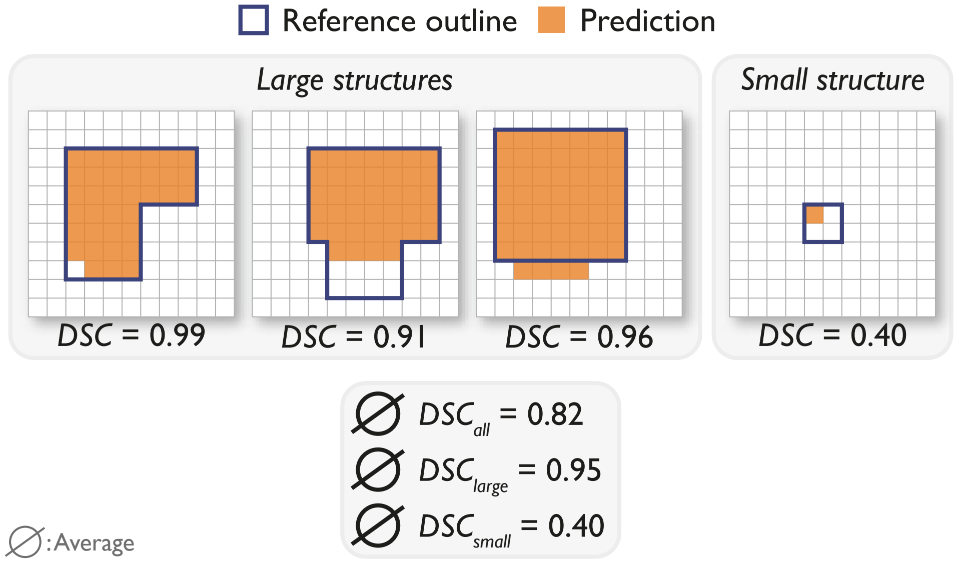
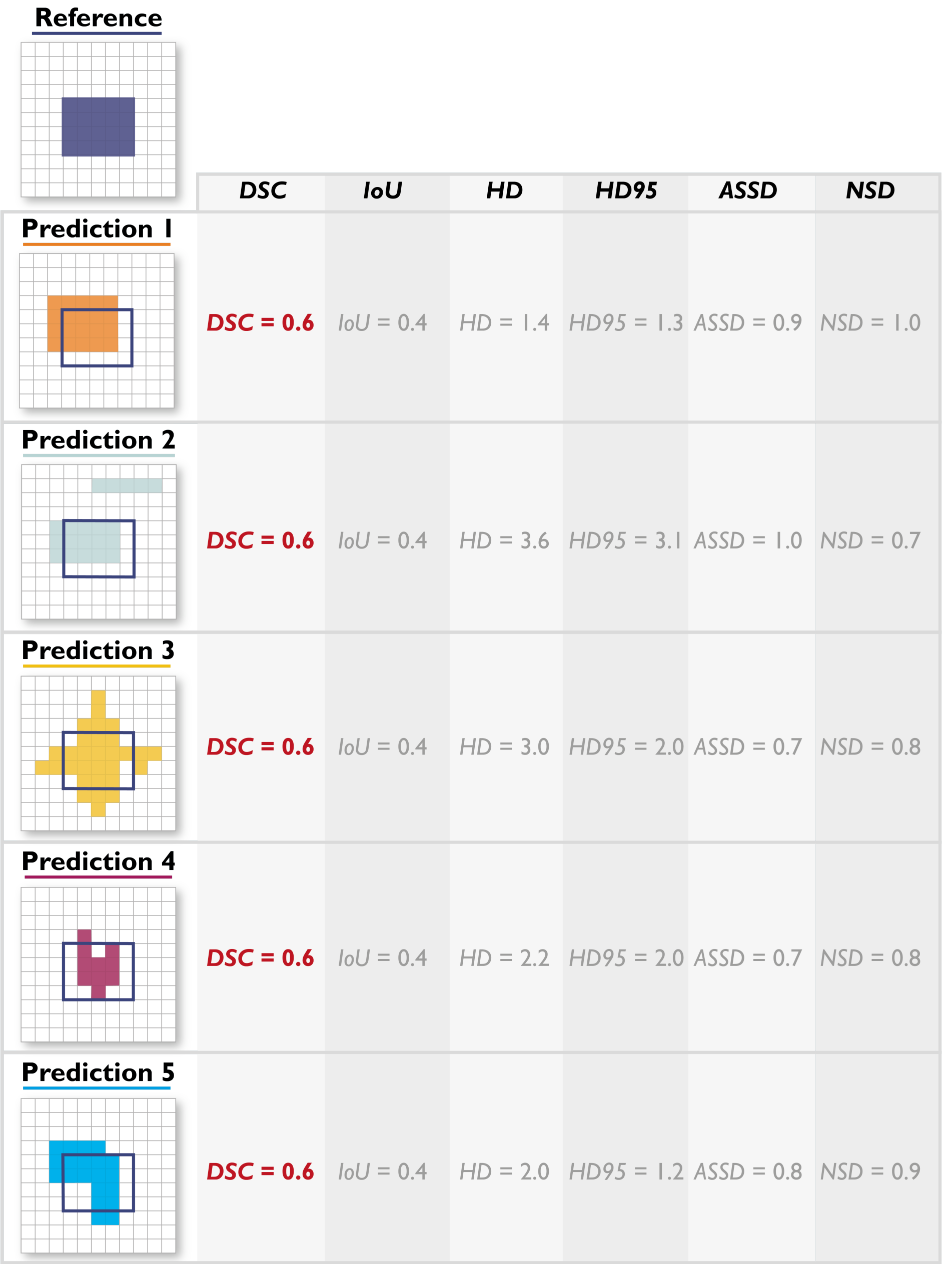


Fig. 26. Effect of high variability of structure sizes across images. As shown in Figure 25, Dice Similarity Coefficient (DSC) scores penalize errors in small structures much more, leading to large and small structures influencing the overall averaged performance to different extents. Only by computing the average for small and large structures separately can it be seen that the prediction in the present example is much poorer for the small structure. ∅ refers to the average *DSC* values.

Complex shapes of structures. Metrics measuring the overlap between objects are not designed to uncover differences in shapes. This is an important problem in many applications such as radiotherapy, for which identifying and treating all parts of the tumor is essential to avoid recurrence [5]. Figure 27 illustrates that completely different object shapes may lead to the exact same *DSC* and *IoU* values. Boundary-based measures are able to detect the changes in shapes [47]. Note that this pitfall also applies to object detection tasks.

Pitfall: Shape unawareness



	DSC	IoU	HD	HD95	ASSD	NSD
Prediction 1	DSC = 0.6	IoU = 0.4	HD = 1.4	HD95 = 1.3	ASSD = 0.9	NSD = 1.0
Prediction 2	DSC = 0.6	IoU = 0.4	HD = 3.6	HD95 = 3.1	ASSD = 1.0	NSD = 0.7
Prediction 3	DSC = 0.6	IoU = 0.4	HD = 3.0	HD95 = 2.0	ASSD = 0.7	NSD = 0.8
Prediction 4	DSC = 0.6	IoU = 0.4	HD = 2.2	HD95 = 2.0	ASSD = 0.7	NSD = 0.8
Prediction 5	DSC = 0.6	IoU = 0.4	HD = 2.0	HD95 = 1.2	ASSD = 0.8	NSD = 0.9

Fig. 27. Effect of different shapes. The shapes of the predictions of five algorithms (*Predictions 1-5*) differ substantially, but lead to the exact same *Dice Similarity Coefficient (DSC)* and *Intersection over Union (IoU)*, while boundary-based metrics (*Hausdorff Distance (HD)*, *Hausdorff Distance 95% percentile (HD95)*, *Average Symmetric Surface Distance (ASSD)* and *Normalized Surface Distance (NSD)*) consider the shape differences.

Particular importance of structure volume. Depending on the domain focus, a surgeon, radiologist or similar may be especially interested in the volume of a segmented structure. The most commonly used metrics may, however, result in predictions at entirely wrong locations if boundary or overlap are not considered. Figure 28 shows two predictions of a 3x3 square structure, both of them being at the wrong position. While the volume difference is correct for both predictions, the overlap is zero. Only boundary-based metrics will indicate the magnitude of mislocalization of the predicted objects.

Pitfall: Ignoring structure location when focusing on volume

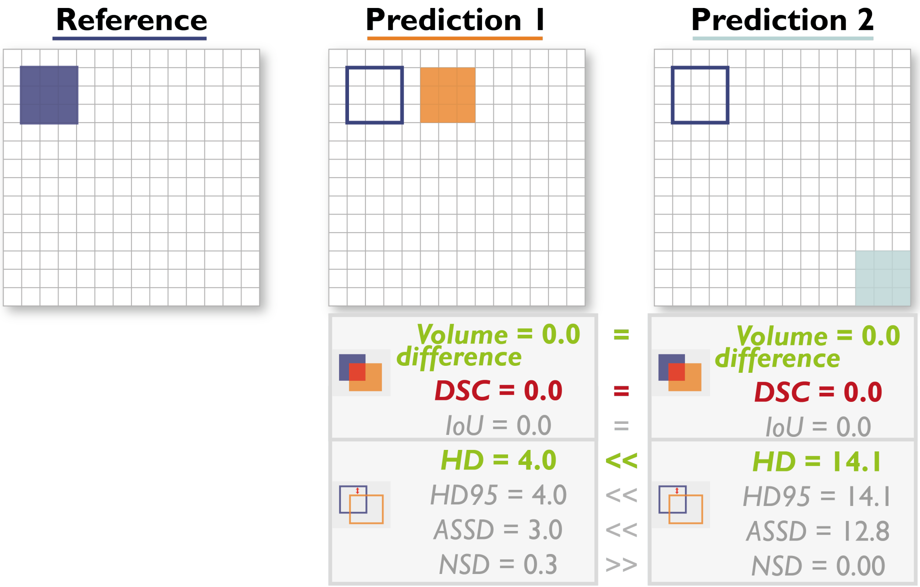


Fig. 28. Effect of only focusing on the volume of an object. Both *Predictions 1* and *2* result in the correct volume difference of 0, but do not overlap with the reference (*Dice Similarity Coefficient (DSC)* and *Intersection over Union (IoU)* of 0). Only the boundary-based measures (*Hausdorff Distance (HD)*, *Hausdorff Distance 95% percentile (HD95)*, *Average Symmetric Surface Distance (ASSD)* and *Normalized Surface Distance (NSD)*) recognize the mislocalization.

Particular importance of structure center. The structure center point or center line may be more important than an accurate boundary or overlap of the structure, as for example in nerve segmentation [34]. In these cases, the accuracy of the center point or line should be examined via an additional metric to make sure the center is correct for the prediction. Figure 29 shows two predictions yielding the same *DSC* values, as they have the same overlap to the reference annotation. However, only *Prediction 1* is centered around the same point as the reference, while *Prediction 2* is shifted slightly towards the upper left corner and thus centered incorrectly. This pitfall also applies to object detection tasks. It should be noted that once the center location is of particular importance to the task, one may consider it an object detection rather than a segmentation problem.

Pitfall: Center unawareness of overlap metrics

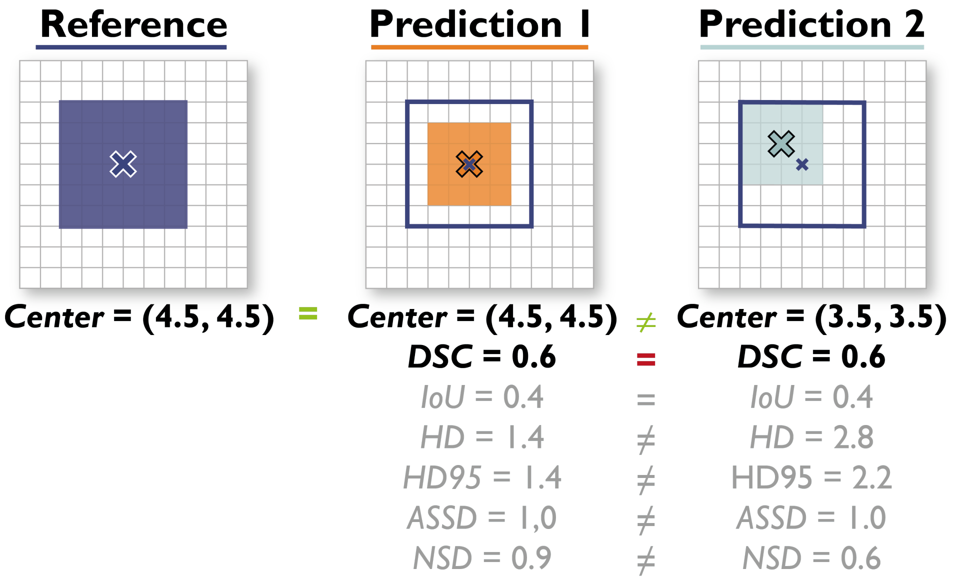


Fig. 29. The most common counting-based metrics are poor proxies for the center point alignment. Here, *Predictions 1* and *2* yield the same *Dice Similarity Coefficient (DSC)* value although *Prediction 1* approximates the location of the object much better.

Particular importance of structure boundaries. While boundary-based metrics like the *HD(95)*, *ASSD* and others can help to detect shape differences between the reference and the predicted object, they do not focus on the object itself. As shown in Figure 30(top), the boundary-based metrics do not recognize a prediction with a large hole inside as poor (*Prediction 2*). Furthermore, in Figure 30(bottom) [47], those metrics do not punish the spotted pattern within the object. It should be noted that this behaviour may also be desirable. For example, it may be highly difficult to decide whether a necrotic core (hole) is present in a tumor or not. A boundary-based metric would not punish errors resulting from such annotation uncertainties.

Pitfall: Holes in the segmentation ignored by boundary-based metrics

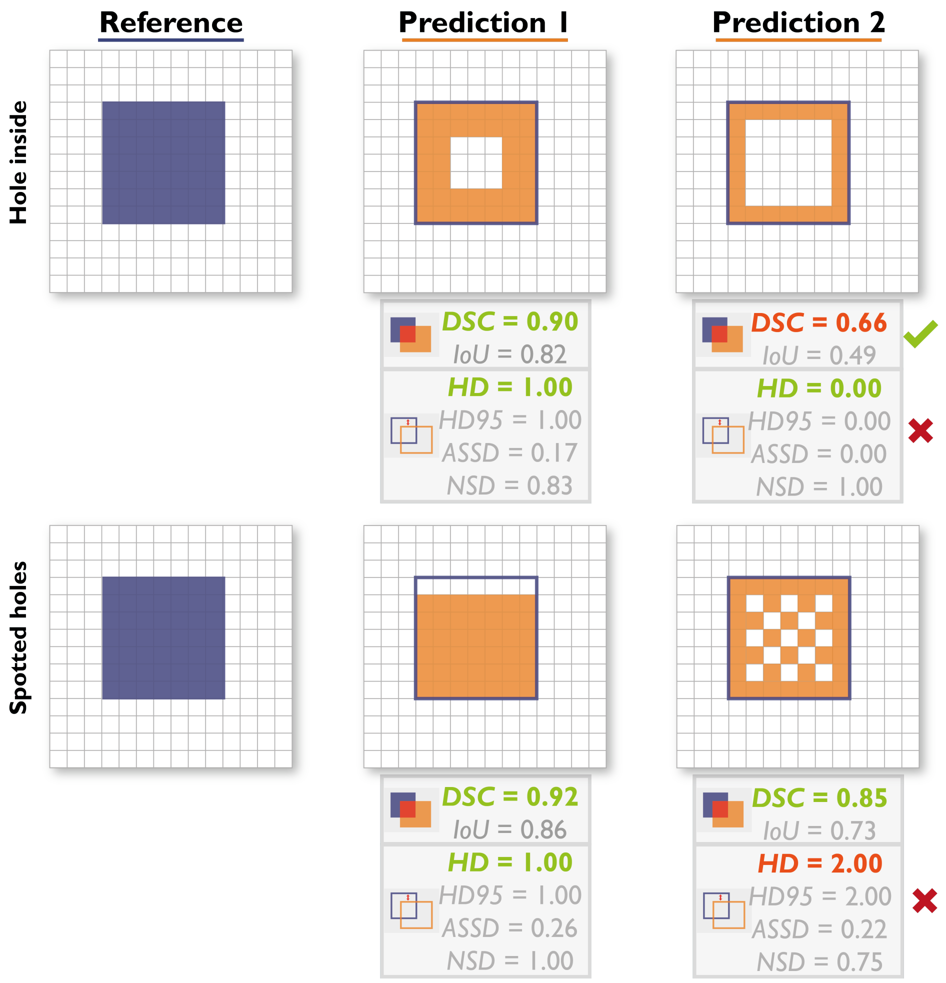


Fig. 30. Boundary-based metrics commonly ignore the overlap between structures and are thus insensitive to holes in structures. **Upper part:** Here, *Predictions 1* and *2* feature holes within the object. The boundary-based metrics (*Hausdorff Distance (HD)*, *Hausdorff Distance 95% percentile (HD95)*, *Average Symmetric Surface Distance (ASSD)*, *Normalized Surface Distance (NSD)*) do not recognize this problem, yielding very good or even perfect metric scores of 0.00 for the *HD(95)*/*ASSD* and 1.00 for the *NSD* (*Prediction 2*), whereas the overlap-based metrics (*Dice Similarity Coefficient (DSC)*, *Intersection over Union (IoU)*) reflect the fact that the inner area is missed by the predictions. **Lower part:** Here, *Predictions 1* and *2* feature a spotted pattern within the object. Although the boundary of *Prediction 2* is perfect, the holes are penalized by the boundary-based metrics compared to *Prediction 1*. *Prediction 1* shows an imperfect boundary. Depending on the surface-based metric used, slight deviations in the boundary (here in *Prediction 1*) may be tolerated, reflected by calculating the *NSD* for $\tau = 1$.

Possibility of multiple labels per unit. In several biomedical imaging scenarios, multiple labels per pixel may be possible. A prominent example would be the tumor core inside the tumor [33]. Often, however, prior knowledge related to such scenarios (e.g. a tumor core cannot lie outside the tumor) is not reflected by common metrics, which simply calculate the agreement of the reference and prediction per class. Figure 31 shows two predictions for a multi-label example. The *DSC* value of *Label 2*, which is required to be inside of *Label 1*, is higher for *Prediction 2* although *Label 2* is also found outside the *Label 1* area. For simplicity, we only show the results for the *DSC* metric. This pitfall also applies to object detection tasks.

Pitfall: Multiple labels per pixel

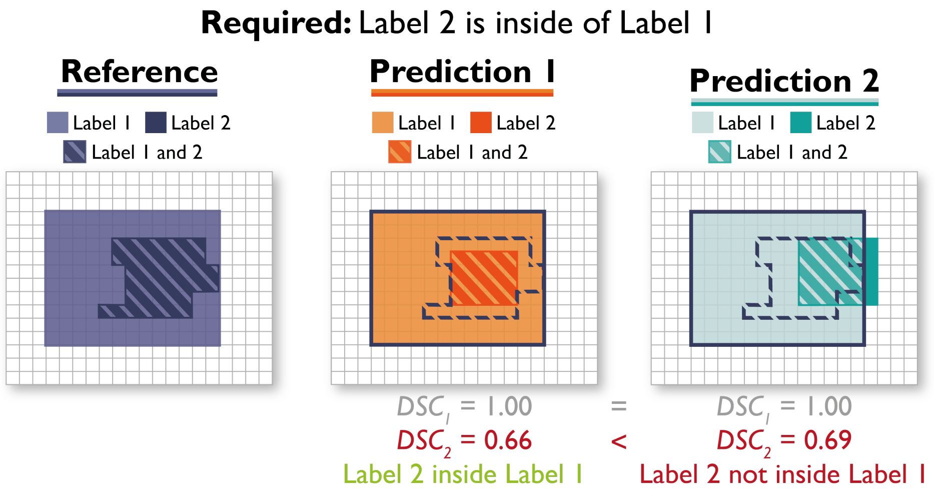


Fig. 31. Effect of multiple labels per pixel. The requirement of *Label 2* being inside of *Label 1* is violated by *Prediction 2*. Nevertheless, *Prediction 2* shows a higher *Dice Similarity Coefficient (DSC)* score compared to *Prediction 1*, which adheres to the requirement.

Noisy reference standard. A high quality reference annotation is crucial to determine the performance of a supervised learning algorithm. A prediction can be almost perfect, but low quality reference images will still result in a bad metric score. Especially in the medical domain, the inter-rater variability is often very high as domain knowledge is required and experts themselves often disagree [22]. Figure 32 shows two masks from different annotators approximating the same structure. Although the annotations differ only slightly at the boundary, the *DSC* score is 0.7. With such inter-rater variability, a *DSC* score of 1 would not be achievable in practice. To address this issue, the *NSD* metric can be applied as an alternative or additional metric, as it is designed to allow a certain tolerance of outline pixels based on the threshold τ. This pitfall can also be translated to object detection and image-level classification tasks.

Pitfall: Inter-rater variability

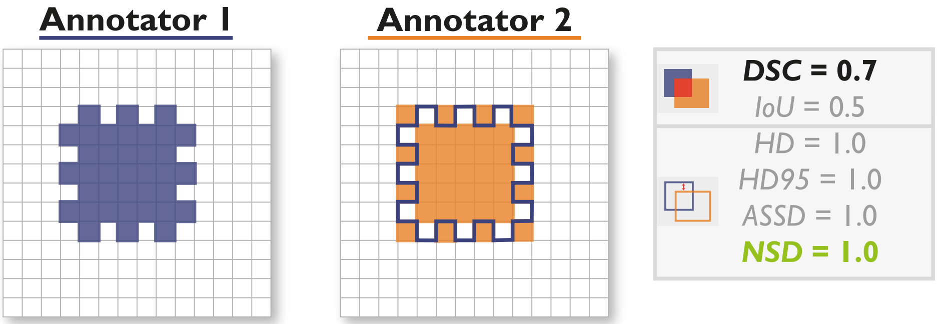


Fig. 32. Effect of inter-rater variability between two annotators. Assessing the performance of *Annotator 2* while using *Annotator 1* for creating the reference annotation leads to a low *Dice Similarity Coefficient (DSC)* score because inter-rater variability is not taken into account by common overlap-based metrics. In contrast, the *Normalized Surface Distance (NSD)*, applied with a threshold of $\tau = 1$, captures this variability. It should be noted, however, that this effect occurs primarily in small structures as overlap-based metrics tend to be robust to variations in the object boundaries in large structures. Further abbreviations: *Intersection over Union (IoU)*, *Hausdorff Distance (HD)*, *Hausdorff Distance 95% percentile (HD95)*, *Average Symmetric Surface Distance (ASSD)*.

Possibility of outliers in reference annotation. The presence of spatial outliers, such as noise or reference annotation artifacts, may severely impact performance metric values. Figure 33 demonstrates how a single erroneous pixel in the reference annotation (or the prediction) leads to a substantial decrease in the measured performance, especially in the case of the *HD*. Using the 95% percentile instead of the maximum (*HD95*) to compute the distance significantly improves the metric score as it can handle outliers. Please note that the presented example may also be seen vice versa, with a prediction including single pixel errors. It should further be noted that whether or not outliers should be considered depends on the respective research question.

Pitfall: Noise and artifacts

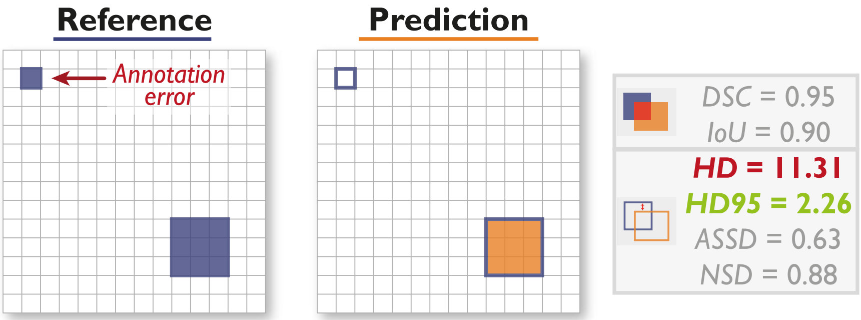


Fig. 33. Effect of annotation errors/noise. A single erroneously annotated pixel may lead to a large decrease in performance, especially in the case of the *Hausdorff Distance (HD)* when applied to small structures. The *Hausdorff Distance 95% percentile (HD95)*, on the other hand, was designed to deal with spatial outliers. Further abbreviations: *Dice Similarity Coefficient (DSC)*, *Intersection over Union (IoU)*, *Average Symmetric Surface Distance (ASSD)*, *Normalized Surface Distance (NSD)*.

Possibility of reference/prediction without target structure(s). A given data set may contain reference annotations without the target structure(s). For example, the data set may consist of healthy and sick patients. A healthy patient will not have a tumor in the image, yielding an empty reference if the tumor is the targeted structure. An algorithm should be careful not to classify a healthy patient as tumourous as this may lead to unnecessary medical interventions. Similarly, a patient with a tumor should not be classified as healthy (empty prediction). These cases require special care to be taken in the validation, because some metrics may be undefined due to division by zero errors or similar. It is necessary to either choose appropriate metrics that consider empty references (or predictions) or account for it in the metric implementation. For example, boundary-based metrics such as the *HD(95)* and *ASSD* will be `NaN` if one of the structures is empty. Figure 34 shows three examples, for which several counting- and boundary-based metrics were computed. The top row depicts the case of an empty reference and a prediction of an object. Given the number of TP and FN being 0, this will result in a division by zero in the *Sensitivity* calculation, yielding a `NaN` score. A similar case is given in the second row, showing an empty prediction for a given target structure in the reference annotation, yielding an undefined *Precision*. When both reference and prediction are empty (bottom row), all scores will be undefined. Please note that this example is shown for a validation per image, as done for segmentation tasks. For classification and object detection tasks, the validation is typically performed over the whole data set, which would possibly preclude this problem. The presented pitfall also applies to object detection tasks.

Pitfall: Reference or prediction without target structure(s)

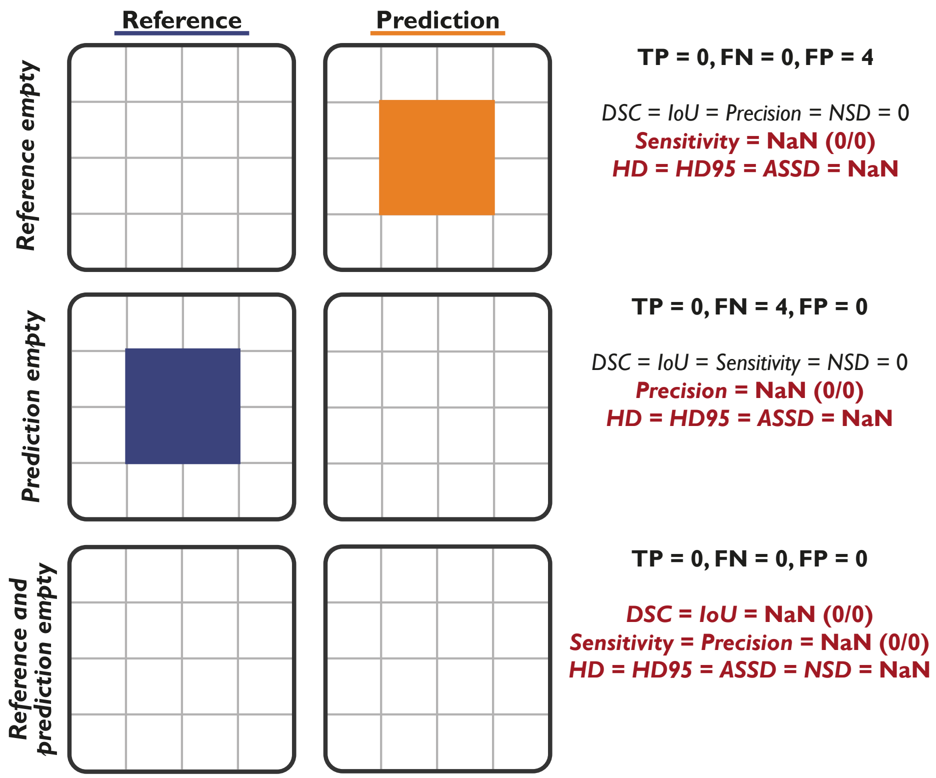


Fig. 34. Effect of empty references or predictions when applying common metrics per image (here for semantic segmentation). Empty images lead to division by zero for many common metrics as the numbers of the true (T)/false (F) positives (P)/negatives (N) turn zero.

Technical peculiarities. Several technical peculiarities also have an impact on metric behaviour. For example, the image resolution and pixel sizes highly influence the reference annotation and the predicted shapes in image processing tasks. Figure 35 illustrates how the reference annotation differs between a low resolution image (top) and a high resolution image (bottom) compared to a circle. The latter is more exact. A prediction of the same size will therefore lead to different corresponding metric values, independent of the type of the metric. This pitfall also applies to object detection tasks.

Pitfall: Grid size resolution

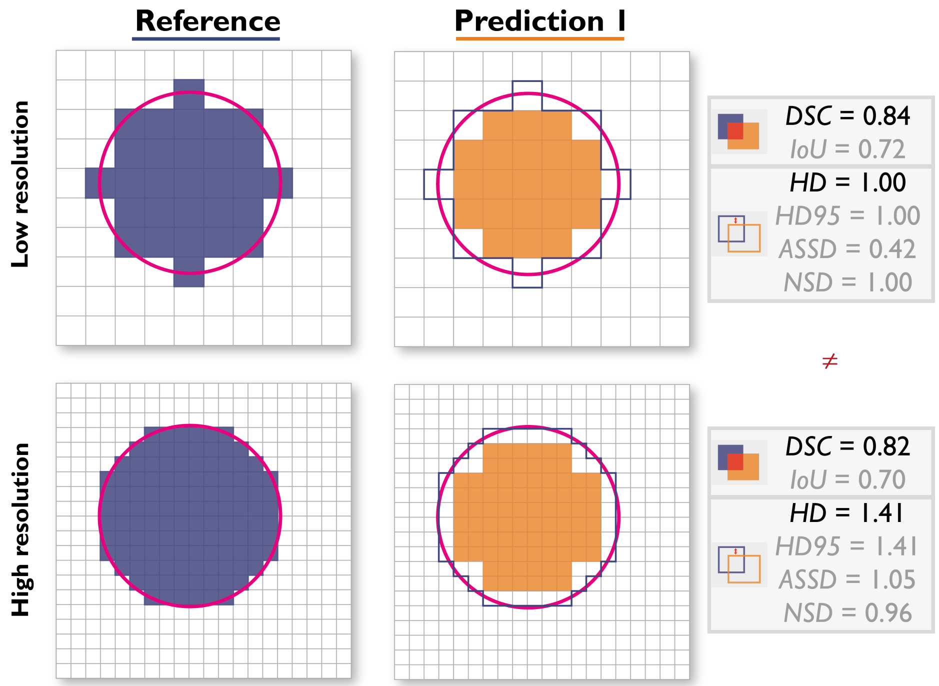


Fig. 35. Effect of different grid sizes. Differences in the grid size (resolution) of an image highly influence the image and the reference annotation (dark blue shape (reference) *vs.* pink outline (desired circle shape)). A prediction of the exact same shape (*Prediction 1*) leads to different metric scores due to the different resolution. Abbreviations: *Dice Similarity Coefficient (DSC)*, *Intersection over Union (IoU)*, *Hausdorff Distance (HD)*, *Hausdorff Distance 95% percentile (HD95)*, *Average Symmetric Surface Distance (ASSD)*, *Normalized Surface Distance (NSD).*

In some applications such as radiotherapy, it may be highly relevant whether an algorithm tends to over- or undersegment the target structure. The *DSC* metric, however, does not represent over- and undersegmentation equally [53]. As depicted in Figure 36, a difference of a single layer of pixels in the outline yields different *DSC* scores (oversegmentation preferred) [47]. Other boundary-based performance values such as the *HD* are invariant to these properties.

Pitfall: Oversegmentation *vs.* undersegmentation

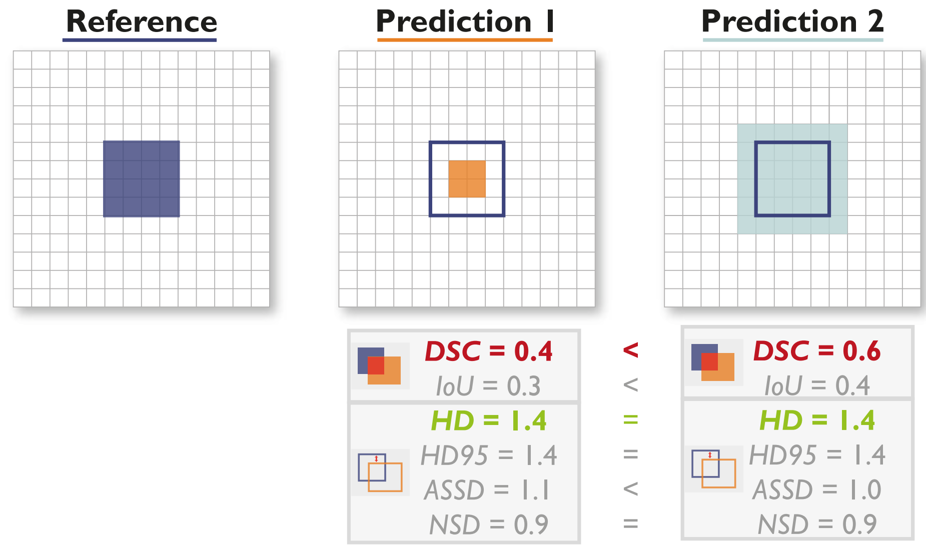
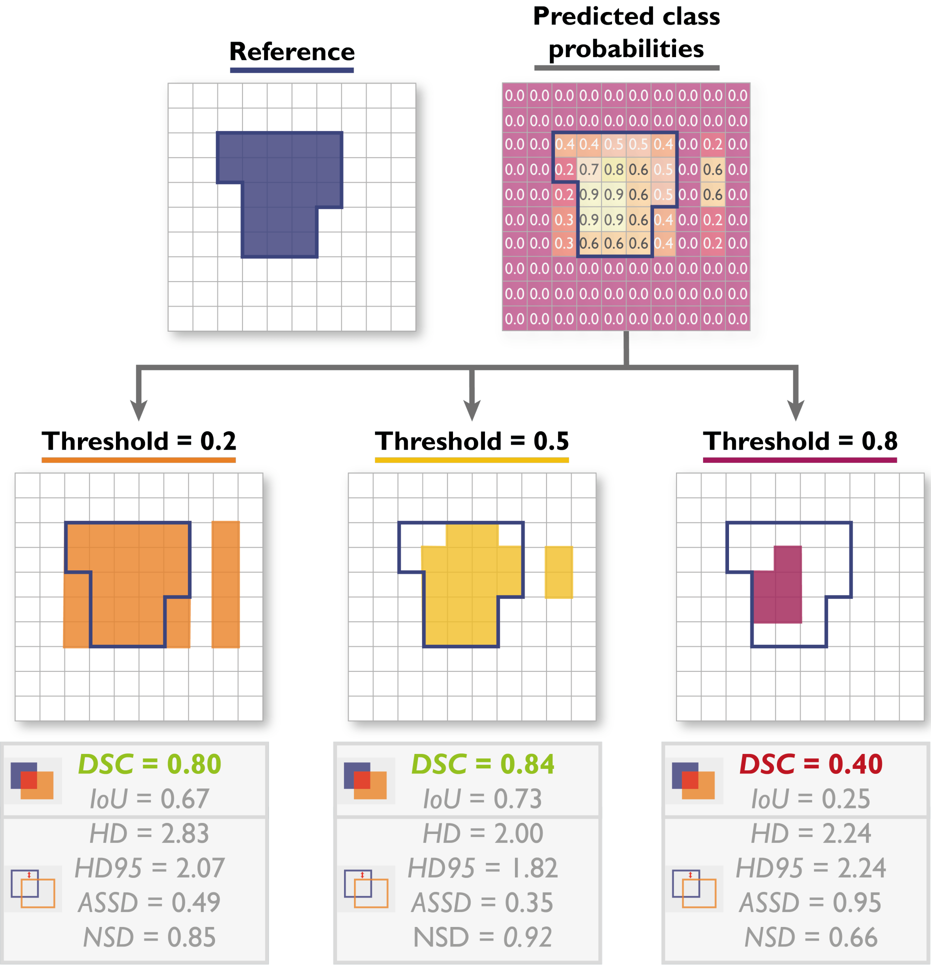


Fig. 36. Effect of undersegmentation *vs.* oversegmentation. The outlines of the predictions of two algorithms (*Prediction 1/2*) differ in only a single layer of pixels (*Prediction 1*: undersegmentation, *Prediction 2*: oversegmentation). This has no (or only a minor) effect on the *Hausdorff Distance (HD)/(95%)*, the *Normalized Surface Distance (NSD)* and the *Average Symmetric Surface Distance (ASSD)*, but yields a substantially different *Dice Similarity Coefficient (DSC)* or *Intersection over Union (IoU)* score.

Another technical peculiarity is the choice of global decision threshold. Most methods in modern image analysis output continuous class scores. While it is quite common to provide those scores in image-level classification and object detection tasks, segmentation architectures often do not output class probabilities per pixel. However, fuzzy segmentation masks are getting more and more common (for instance, see [1, 23, 37]) and the choice of a global decision threshold τ is very important for the algorithm's result. Figure 37 (cf. [36]) shows the predicted class probabilities for a reference annotation. For a binarization typically required for segmentation outputs, a threshold needs to be defined based on which a pixel is assigned to a class (here: a pixel with class probability $< \tau$ corresponds to the background class and to the foreground class otherwise). The resulting segmentation masks are shown for the thresholds 0.2, 0.5 and 0.8. It can be seen that the respective masks completely differ across the thresholds. Consequently, metric values will also vastly change.

Pitfall: Choice of global decision threshold



0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.4	0.4	0.5	0.5	0.4	0.0	0.2	0.0
0.0	0.0	0.2	0.7	0.8	0.6	0.5	0.0	0.6	0.0
0.0	0.0	0.2	0.9	0.9	0.6	0.5	0.0	0.6	0.0
0.0	0.0	0.3	0.9	0.9	0.6	0.4	0.0	0.2	0.0
0.0	0.0	0.3	0.6	0.6	0.6	0.4	0.0	0.2	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Fig. 37. Effect of the choice of a global decision threshold for binarization in segmentation problems. The results and metric values for three decision thresholds (0.2, 0.5 and 0.8), from which predicted class probabilities per pixel are translated into a segmentation mask, are provided. Based on the threshold chosen, the predicted objects and resulting metric values change dramatically.

7 PITFALLS RELATED TO OBJECT DETECTION

All pitfalls compiled for this work and relevant for object detection are summarized in Table 1. Note that these pitfalls equally apply to instance segmentation problems. While most issues related to the actual metric selection have already been mentioned in the previous paragraphs, this section is primarily dedicated to technical peculiarities related to the localization and assignment criteria. These include:

- Mathematical implications of center-based localization criteria (Figure 38)
- Mathematical implications of *IoU*-based localization criteria (Figure 39)
- Effect of the provided annotations (Figure 40)
- Effect of small structures on localization criterion (Figure 41)
- Perfect *Boundary IoU* for imperfect prediction (Figure 42)
- Possibility of reference or prediction without the target structure and `NaN` handling (Figure 43)
- *Average Precision vs. Free-response ROC* score (Figure 44)
- Effect of predicted class probabilities on multi-threshold metrics (Figures 45, 46 and 47)

Mathematical implications of center-based localization criteria. Before calculating metrics for object detection tasks, it is necessary to define what qualifies a detection as a *hit* (TP) or *miss* (FP). There are multiple ways to define a hit, all of which come with their specific limitations. Below, the most commonly used center-based localization criteria are presented[10].

- For the **center-cover criterion**, the reference object is considered a hit if the center of the reference object is inside the predicted detection. Figure 38a shows how this criterion can be fooled by a model outputting very large boxes to maximize the chance of a correct detection.
- In the case of the **distance-based hit criterion**, a prediction is considered a hit if the distance d between the center of the reference and the detected object is smaller than a certain threshold τ. In Figure 38b, both predictions have the same distance to their corresponding reference object centers. However, the prediction on the top right shows no overlap with the reference and should therefore not be counted as a hit.
- The **center-hit criterion** holds true if the center of the predicted object is inside the reference bounding box or contour. Given this definition, large reference objects are more likely to be hit, as shown in Figure 38c. The left prediction is defined as a missed object (FP), the right detection as a hit because of its larger size.

[10] For more details, please refer to the blogpost "Evaluation curves for object detection algorithms in medical images": https://medium.com/lunit/evaluation-curves-for-object-detection-algorithms-in-medical-images-4b083fddce6e.

Pitfall: Mathematical implications of center-based localization criteria

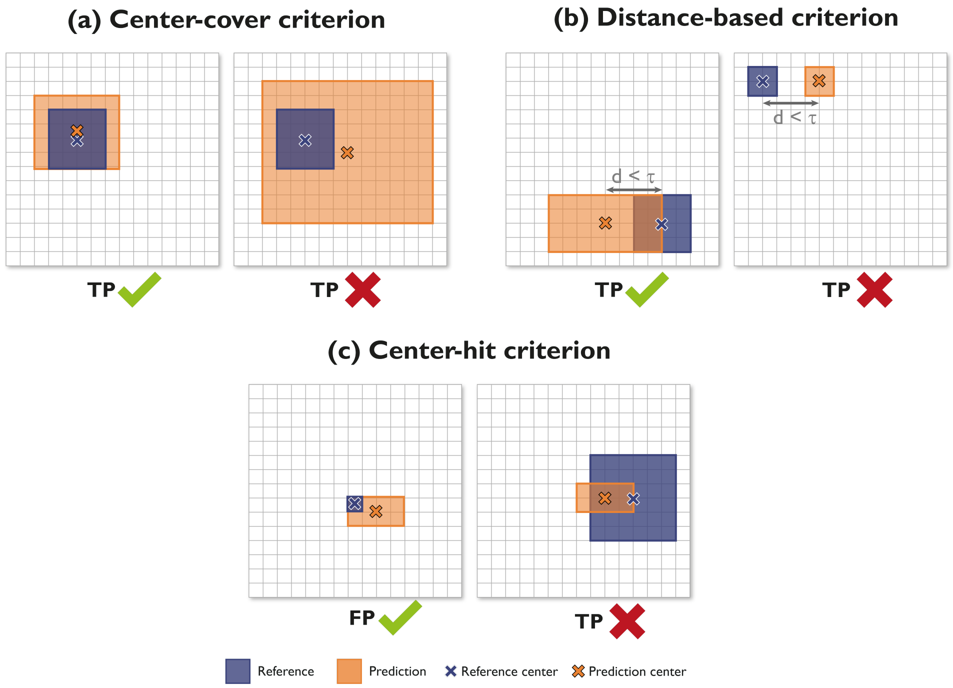


Fig. 38. Pitfalls for several center-based hit criteria in object detection. Reference objects are shown in dark blue, predictions in orange. The object centers are shown as blue/orange crosses. **(a)** The **center-cover criterion**, which requires the center of the reference to be inside the detection, can be fooled easily by predicting a very large bounding box/object. **(b)** Both predictions have the same distance to their corresponding reference center. The **distance-based criterion**, which requires the distance 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). **(c)** The **center-hit criterion**, which requires the center of the prediction to be located inside the reference, favors large reference objects, as they are easier to detect. The left prediction is considered a False Positive (FP), as the reference center was not hit. The right prediction is considered a TP because of the larger size of the object.

Mathematical implications of IoU-based localization criteria. The most commonly used hit criterion is determined by computing the *IoU* between the predicted and the reference mask/bounding box/boundary. Pitfalls related to *IoU*-based criteria are mainly related to the setting of the threshold (Figure 39). Many biomedical applications involve 3D rather than 2D images. When working with a higher dimension, it should be kept in mind that metrics may be affected. The additional dimension will lead to overlap errors being punished even more. Figure 39a shows a comparison of the *IoU* for two rectangles (or bounding boxes) in 2D and 3D. Being mistaken by one voxel in the z-dimension will lead to a much lower *IoU* score in 3D compared to the 2D case.

As *IoU*-based criteria take the overlap between regions into account, it is only possible to cheat with very large boxes if the *IoU* threshold is set to a very small value (here: 0), as shown in Figure 39b. However, special care should be taken when applying the *Box IoU* in the presence of highly concave or elongated structures, as illustrated in Figure 39c. This is because bounding boxes may quickly grow for narrow and diagonally placed objects, such as medical instruments, and result in FP although visual inspection would indicate a correct prediction.

Pitfall: Mathematical implications of *IoU*-based localization criteria

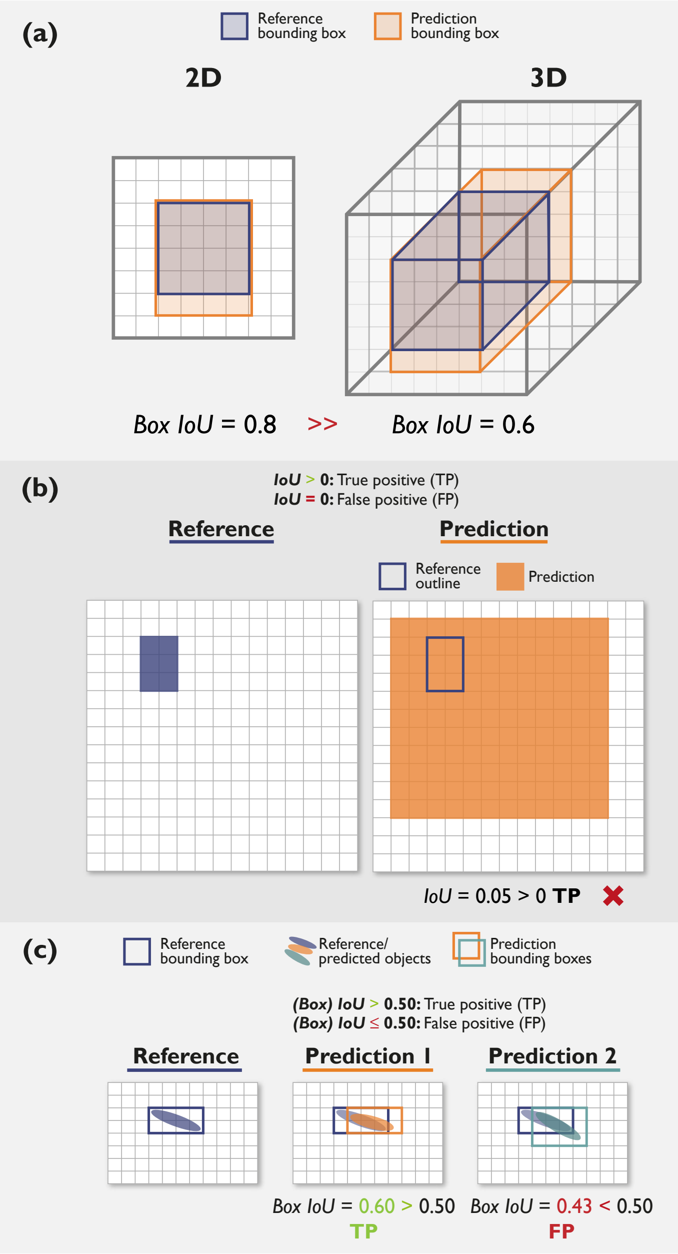


Fig. 39. The *Intersection over Union (IoU)* is a commonly used localization criterion in object detection. It comes with several limitations: **(a)** The image dimension should be considered when setting the *IoU* (here: *Box IoU*) threshold for object detection. In 3D settings, the additional z-dimension results in a cubical increase in erroneous pixels. **(b)** Effect of a loose *IoU* criterion for object detection. When defining a True Positive (TP) by an *IoU* > 0, the resulting localizations may be fooled by very large predictions. **(c)** Effect of defining TP based on the *IoU* (here: *Box IoU*) threshold of the reference and predicted bounding boxes. Especially for diagonal, narrow objects, the number of bounding box pixels may change quadratically. Although *Predictions 1* and *2* are very similar, their bounding boxes diverge and lead to one of them being defined as TP, the other as False Positive (FP).

Possibility of disconnected structures. The *Box IoU* is sometimes employed despite access to pixel-mask annotations. A possible explanation is that researchers want to phrase their problem as an object detection problem and then apply the most commonly used validation methods. Such simplification might cause problems, if structures are not well approximated by a box shape, or if structures yield multi-component masks, appearing to be disconnected. This may occur in the case of a tubular structure shown in a 2D tomographic image or a medical instrument occluded by tissue in an endoscopic image, for example. Figure 40 provides examples of a complex diagonal (top) and a disconnected structure (bottom). Both box predictions yield a *Box IoU* larger than 0.3, and are thus counted as TP because of the chosen localization threshold. Nevertheless, *Prediction 1* is not hitting the actual object at all. This is due to the fact that the target structures are not well approximated by the bounding box, leaving many empty pixels in the boxes.

Pitfall: Effect of annotation type

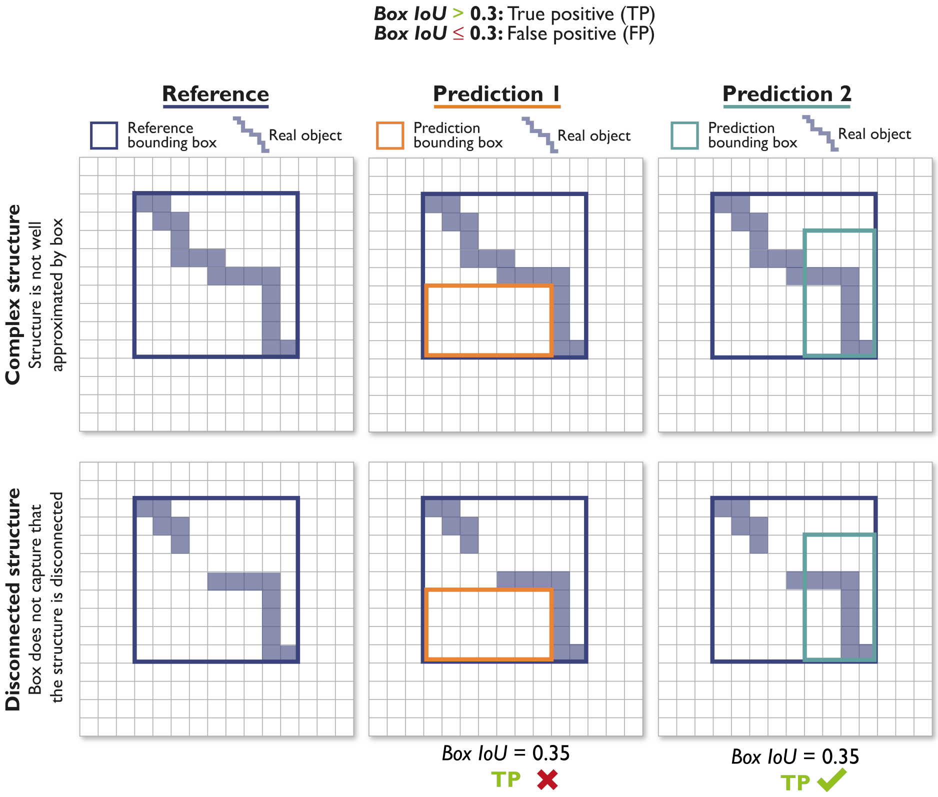


Fig. 40. Bounding boxes are not well-suited for representing complex (top) and disconnected (bottom) shapes. Specifically, they are not well-suited for capturing multi-component structures. *Predictions 1* and *2* would both end up in a True Positive (TP) detection, as the *Box Intersection over Union (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.

Effect of small structures on localization criterion. *Box IoU* and *Mask IoU* are not sensitive to structure boundary quality in larger objects (cf. Section 6). This is due to the fact that boundary pixels will increase linearly (quadratically in 3D) while pixels inside the structure will increase quadratically (cubic in 3D) with an increase in structure size. In consequence, the *IoU*-scores tend to be higher for large objects compared to small objects. For this reason, localization criteria such as the ***Boundary IoU*** were designed.

Figure 41 shows an example of *Mask IoU* and *Boundary IoU* for a large (top) and a rather small structure (bottom). In the case of the *Mask IoU*, the score drops substantially for the small structure, while the scores are more consistent for the *Boundary IoU* when comparing small and large structures. This pitfall also applies to segmentation problems in which the *(Mask) IoU* and *Boundary IoU* are applied as overlap-based metrics.

Pitfall: Effect of small structures on localization criterion

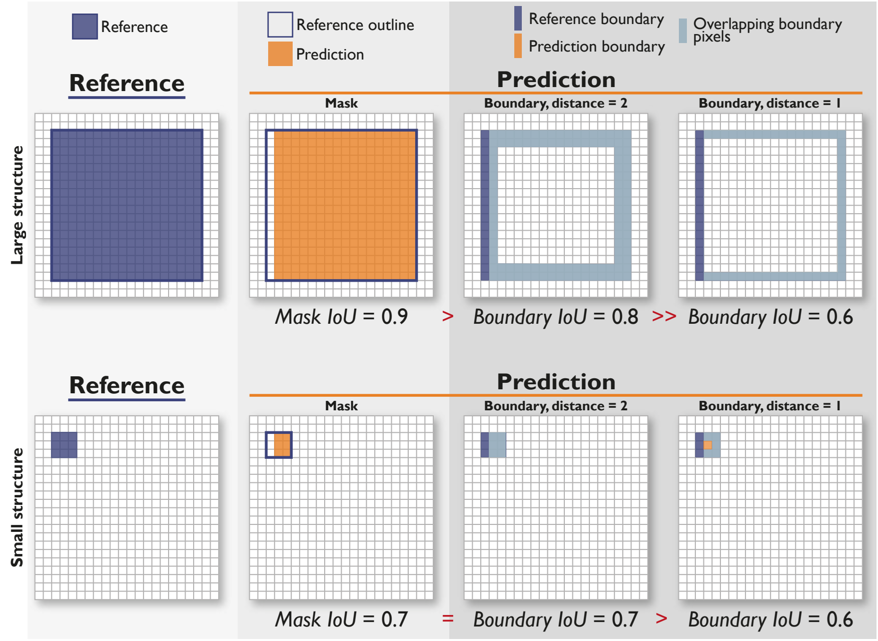


Fig. 41. Comparison of *Mask* and *Boundary Intersection over Union (Boundary IoU)* localization criteria in the case of particular importance of structure boundaries. Overlapping pixels from the reference and prediction are shown in light blue. The *Mask IoU* (second column) is less sensitive to boundary errors for large objects. The *Boundary IoU* (third and fourth column) especially considers contours and (1) leads to smaller metric scores, thus penalizing errors in the boundaries and (2) is more invariant to structure sizes, as it leads to very similar values for large and small structures (fourth column).

It should further be noted that the *Boundary IoU* is highly dependent on the chosen distance *d*, as illustrated in Figure 41 (third vs fourth column). Similarly to the example provided in Figure 30, the *Boundary IoU* can be fooled to result in a perfect value of 1.0. A prediction with a hole in the middle of the structure may result in a perfect metric score if the distance is chosen in a way that it incorporates all pixels of the predicted mask, as shown in Figure 42 [8]. The *Mask IoU*, however, will be able to recognize the problem, as it completely measures the overlap between both structures. [8] propose to use the $\min(BoundaryIoU, MaskIoU)$ to resolve this issue. Please note that the same limitations also affect other distance-based measures, such as the *NSD* or *HD* metrics. This pitfall also applies to segmentation tasks.

Pitfall: Perfect *Boundary IoU* for imperfect prediction

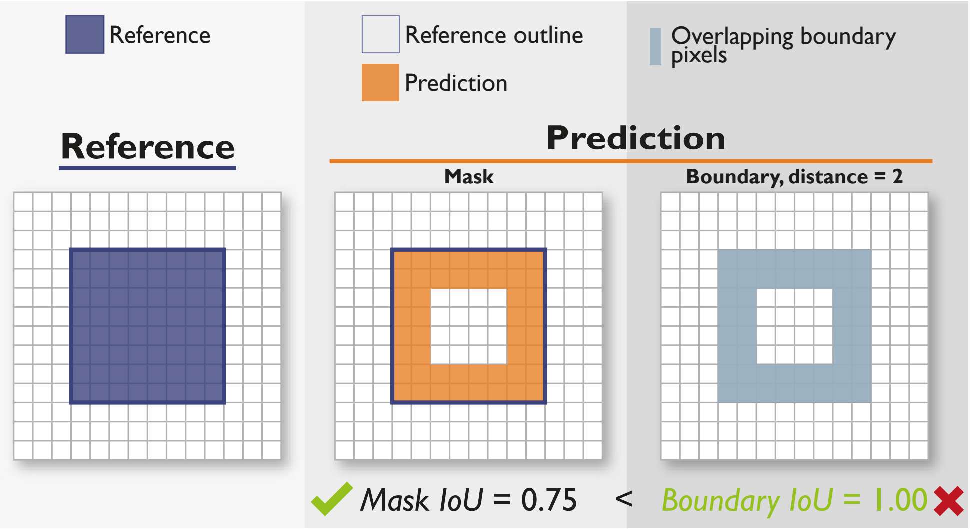


Fig. 42. Effect of a perfect *Boundary Intersection over Union (Boundary IoU)* score for an imperfect prediction. Overlapping pixels from the reference and prediction are shown in light blue. For a prediction with a hole in the middle, the *Boundary IoU* may result in a score of 1.00 if the distance to border contains all mask pixels (here: distance = 2). However, the *Mask IoU* spots the problem and results in a lower score.

Possibility of reference/prediction without the target structure and **`NaN`** ***handling*.** When validating an object detection problem per image rather than per data set, a reference or prediction image without the target structure(s) may become problematic as some metric values will turn into `NaN` due to division by zero errors (cf. Figure 34). Figure 43a shows potential scenarios for a validation per image categorized by the presence and absence of TP, FP and FN. Four occurrences of `NaN` are presented. To proceed with the validation, namely aggregating metric values for every image over the entire data set, a `NaN` strategy needs to be defined for every use case.

Pitfall: NaN handling for empty reference/prediction

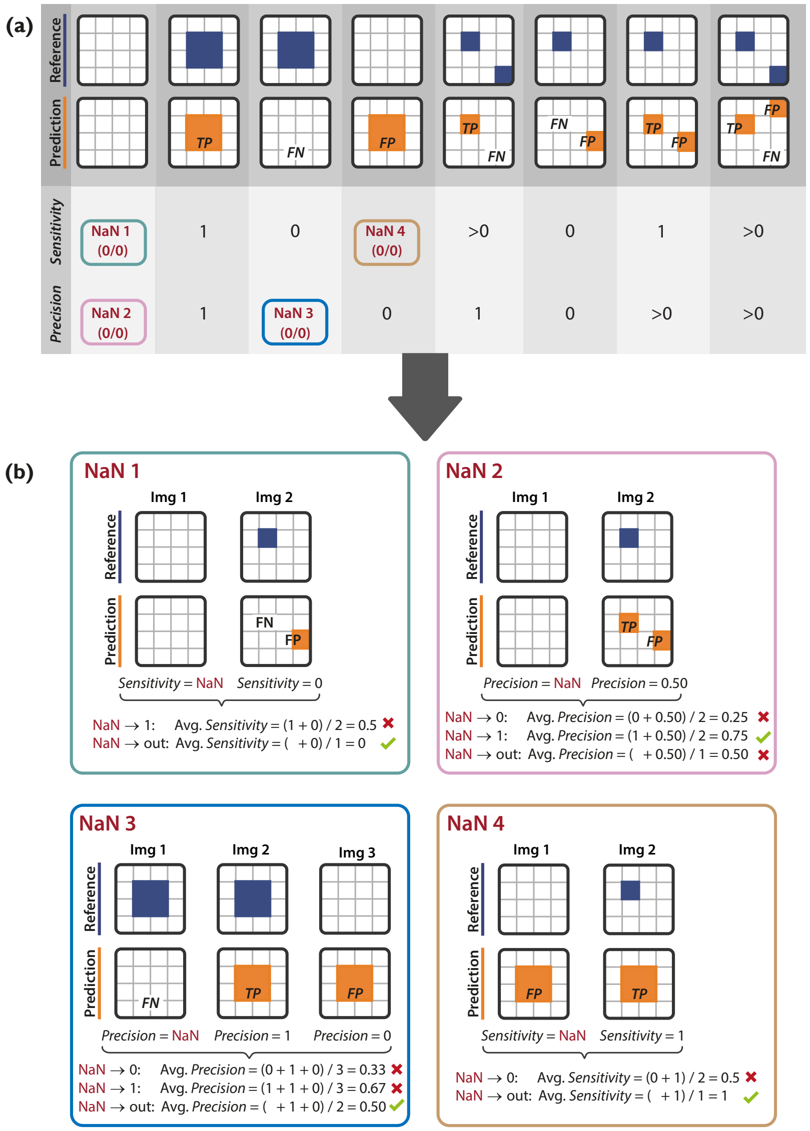

(a)
Reference
Prediction
TP
FN
FP
TP
FN
FN
FP
TP
FP
TP
FP
FN
Sensitivity
NaN 1 (0/0)
1
0
NaN 4 (0/0)
>0
0
1
>0
Precision
NaN 2 (0/0)
1
NaN 3 (0/0)
0
1
0
>0
>0
(b)
NaN 1
Img 1
Img 2
Reference
Prediction
FN
FP
Sensitivity = NaN
Sensitivity = 0
NaN → 1: Avg. Sensitivity = (1 + 0) / 2 = 0.5 ✗
NaN → out: Avg. Sensitivity = (+ 0) / 1 = 0 ✓
NaN 2
Img 1
Img 2
Reference
Prediction
TP
FP
Precision = NaN
Precision = 0.50
NaN → 0: Avg. Precision = (0 + 0.50) / 2 = 0.25 ✗
NaN → 1: Avg. Precision = (1 + 0.50) / 2 = 0.75 ✓
NaN → out: Avg. Precision = (+ 0.50) / 1 = 0.50 ✗
NaN 3
Img 1
Img 2
Img 3
Reference
Prediction
FN
TP
FP
Precision = NaN
Precision = 1
Precision = 0
NaN → 0: Avg. Precision = (0 + 1 + 0) / 3 = 0.33 ✗
NaN → 1: Avg. Precision = (1 + 1 + 0) / 3 = 0.67 ✗
NaN → out: Avg. Precision = (+ 1 + 0) / 2 = 0.50 ✓
NaN 4
Img 1
Img 2
Reference
Prediction
FP
TP
Sensitivity = NaN
Sensitivity = 1
NaN → 0: Avg. Sensitivity = (0 + 1) / 2 = 0.5 ✗
NaN → out: Avg. Sensitivity = (+ 1) / 1 = 1 ✓

Fig. 43. Effect of handling a NaN caused by reference or prediction without target structure(s) in object detection/instance segmentation problems validated per image. **(a)** Demonstration of how and when NaN can occur. Each column represents a potential scenario for per-image validation of objects, categorized by whether True Positive (TP), False Negative (FN), and False Positive (FP) are present (n > 0) or not present (n = 0) after matching/assignment. The sketches on the top showcase each scenario when setting "n > 0" to "n = 1". For each scenario, *Sensitivity* and *Precision* are calculated. **(b)** Effect of different NaN handling strategies based on different conventions for the aggregation across multiple images. Four examples are shown for the NaN scenarios from (a) (**NaN 1-4**). **NaN 1** and **4**: The intuitive penalization for FPs in "empty" images is already established by means of *Precision* scores (see **NaN 4**) and further penalization by means of *Sensitivity* is neither required nor appropriate. Instead, images without reference objects should be ignored when averaging *Sensitivity* scores over images. **NaN 2**: The intuitive penalization for FP in "empty" images is established when assigning a *Precision* of 1. **NaN 3**: The intuitive penalization for FP is established when removing images with FN and no FP from the aggregation of *Precision* scores.

Average Precision vs. Free-response ROC score. While the AP constitutes the standard metric for object detection and instance segmentation in the computer vision community, the *FROC* score is often favoured in the clinical context. In contrast to the AP, the FROC score takes into account the total number of images in the data set. As can be seen from Figure 44, both data sets D1 and D2 will yield the same AP score, although data set D1 contains two images and D2 contains four images. The FROC score, however, will reflect that the number of images is different for both data sets and that data set D2 contains two images that do not contain any FP. Thus, the FPPI will be lower in data set D2, yielding in a higher FROC score.

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

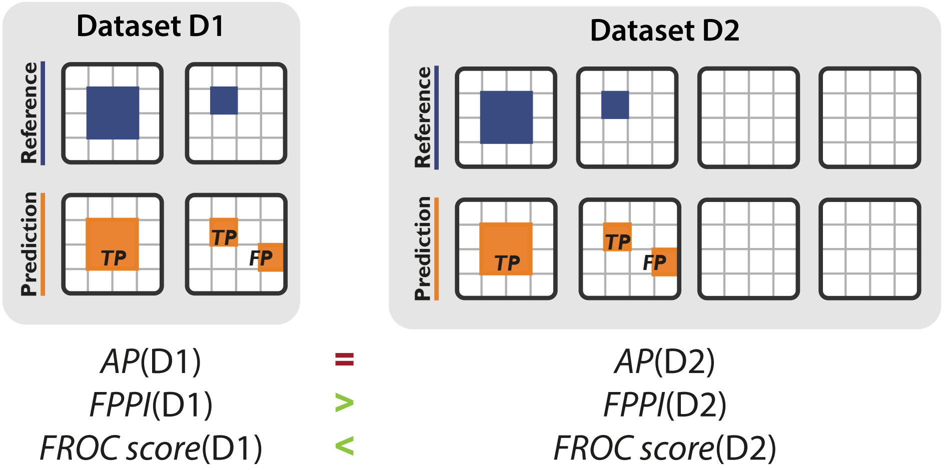


Fig. 44. Effect of the number of images per data set on the metric scores. The Average Precision (AP) metric does not take into account the total number of images, yielding the same score for data sets D1 and D2. The Free-Response Receiver Operating Characteristic (FROC) curve plots the average number 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.

Multi-threshold metric-related properties. In the next paragraphs, we highlight some limitations of the multi-threshold metrics, exemplarily for the *AP* metric, which can be transferred to other multi-threshold metrics, such as the *AUROC* [39]. By definition, multi-threshold metrics are ranking metrics, which rank the predicted class probabilities or confidence scores (cf. Figure 11). They are not designed to reflect the calibration of confidence or class scores, as shown in the following examples. Please note that we disregard the concrete choice of the localization criterion here for simplicity.

Predicted class probabilities. The *PR* curve and the resulting metric score *AP* highly depend on the ranking of predictions, based on their predicted class probabilities or confidence scores. Small changes in the scores can therefore significantly change the metric value, as shown in Figure 45. On the other hand, as long as the ranking remains unchanged among predictions, the predicted class probabilities themselves are not important for the result, although they should be (see Figure 46).

Pitfall: Large effects of small changes in predicted class probabilities

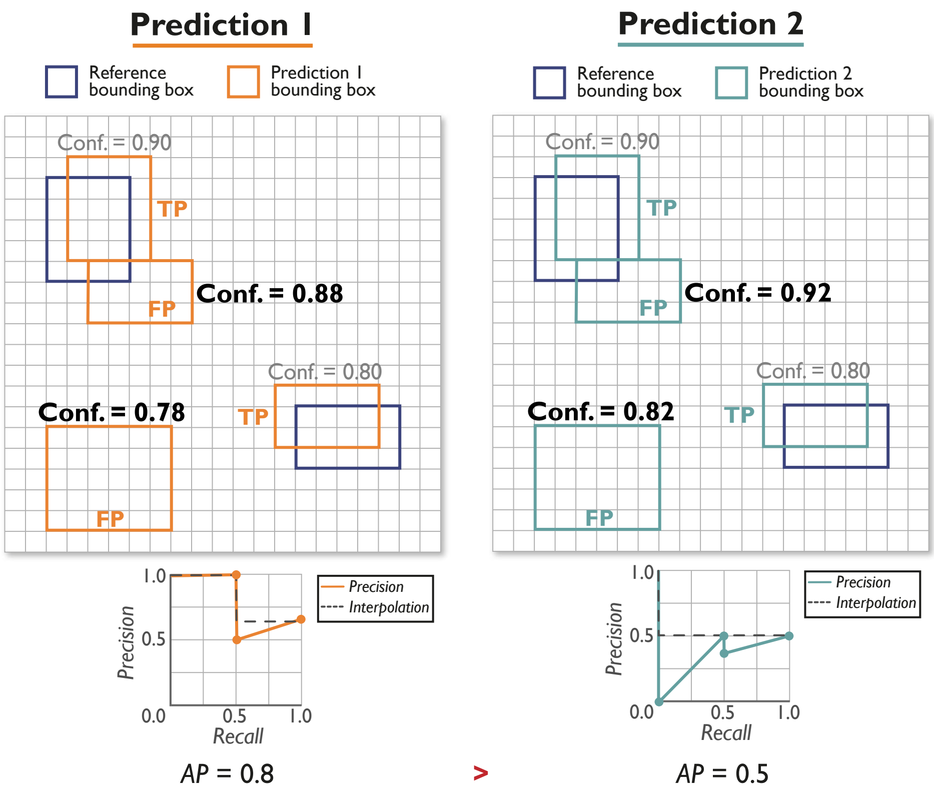


Fig. 45. Effect of small changes in predicted class probabilities. Reference bounding boxes are shown in dark blue. *Predictions 1* and *2* detect the exact same bounding boxes with minor variations in their predicted class probabilities (represented by the confidence scores (conf.)). This leads to a different ranking and therefore varying *Precision-Recall (PR)* curves, curve interpolations (dashed grey lines) and resulting *Average Precision (AP)* scores.

Pitfall: Predicted class probabilities neglected within ranking

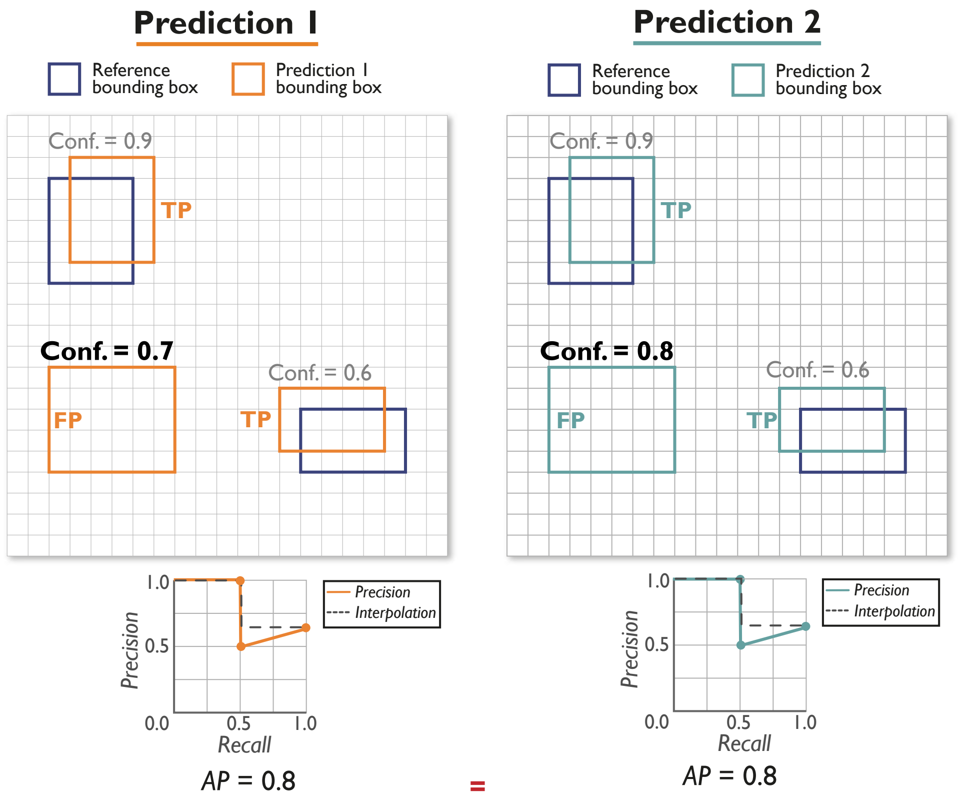


Fig. 46. Effect of neglecting (the absolute values of) predicted class probabilities within the ranking. Reference bounding boxes are shown in dark blue. *Predictions 1* and *2* detect the exact same bounding boxes with variations in their predicted class probabilities (represented by the confidence scores (conf.)) that do not affect the ranking. Therefore, the *Precision-Recall (PR)* curves, curve interpolations (dashed grey lines) and resulting *Average Precision (AP)* scores are the same; the predicted class probabilities are hence unimportant within the ranking.

FP with low predicted class probabilities. False positive predictions with lower predicted class probabilities than the last correctly predicted reference, corresponding to the end of the *PR* curve, do not affect the *AP* scores. Figure 47 shows two examples that are very similar, only differing in the number of wrongly predicted objects. *Prediction 2*, with two FP, performs worse than *Prediction 1* with only one FP. Nevertheless, the *AP* scores are the same for both models, given the low confidence of the second FP of *Prediction 2*.

Pitfall: FP with low predicted class probabilities

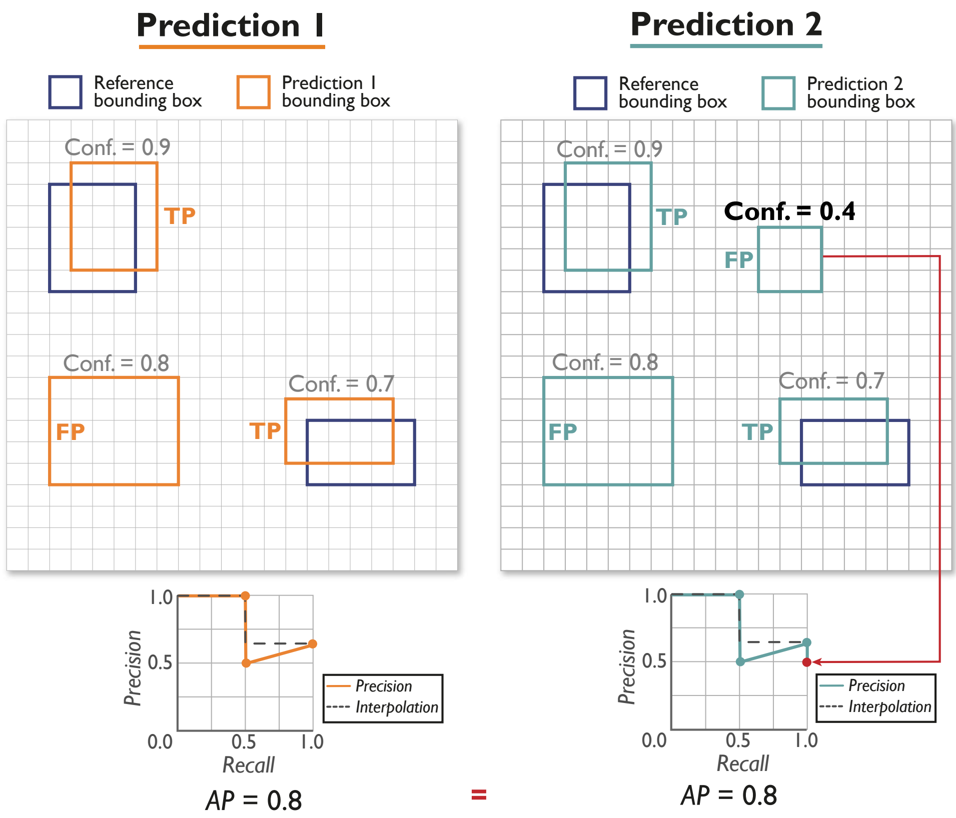


Fig. 47. Effect of False Positive (FP) predictions with low predicted class probabilities (represented by the confidence scores (conf.)). Reference bounding boxes are shown in dark blue. *Prediction 1* and *2* predict the exact same bounding boxes, but *Prediction 2* shows one additional FP detected box with low predicted class probabilities. This is not reflected in the *AP* score, as the FP is located at the tail of the *Precision-Recall (PR)* curve and does not change the curve interpolation (dashed grey lines).

8 PITFALLS RELATED TO ANALYSES AND POST-PROCESSING

A data set typically contains several hundreds or thousands of images. When analyzing, aggregating and combining metric values, a number of factors need to be taken into account. Pitfalls in this step are primarily related to the following aspects:

- Uninformative visualization (Figure 48)
- Metric aggregation for invalid algorithm output (e.g. `NaN`) (Figures 49 and 50)
- Hierarchical data aggregation (Figure 51)
- Aggregation in the presence of multiple classes (Figure 52)
- Combination of related metrics (Figure 53)

Uninformative visualization. Relying on only reporting aggregated metric scores may result in missing essential information on algorithm performance. Therefore, raw metric values (e.g. per image) should always be shown, for example in the shape of boxplots, as depicted in the top left of Figure 48. However, boxplots will only provide information on some key descriptive statistics, like median or 1st and 3rd quartiles. Another choice can be violin plots, which further visualize the raw data distribution. The top right of Figure 48 illustrates the multimodal distribution of the underlying data, invisible in the boxplot. Furthermore, using a violin plot and/or plotting the raw metric values for each data point on top (Figure 48, top right and bottom left) will reveal the complete data distribution. In the example below, many values lie below the 3rd quartile, although the box looks tight. Nevertheless, even these two visualizations may hide important information. Assume a data set with metric values of four different videos. Color- or shape-coding the metric values by the video type (Figure 48 bottom right) reveals a huge cluster of extremely low *DSC* values only affecting Video 4 (pink), which would have been hidden by the other two types of visualization.

Pitfall: Uninformative visualization

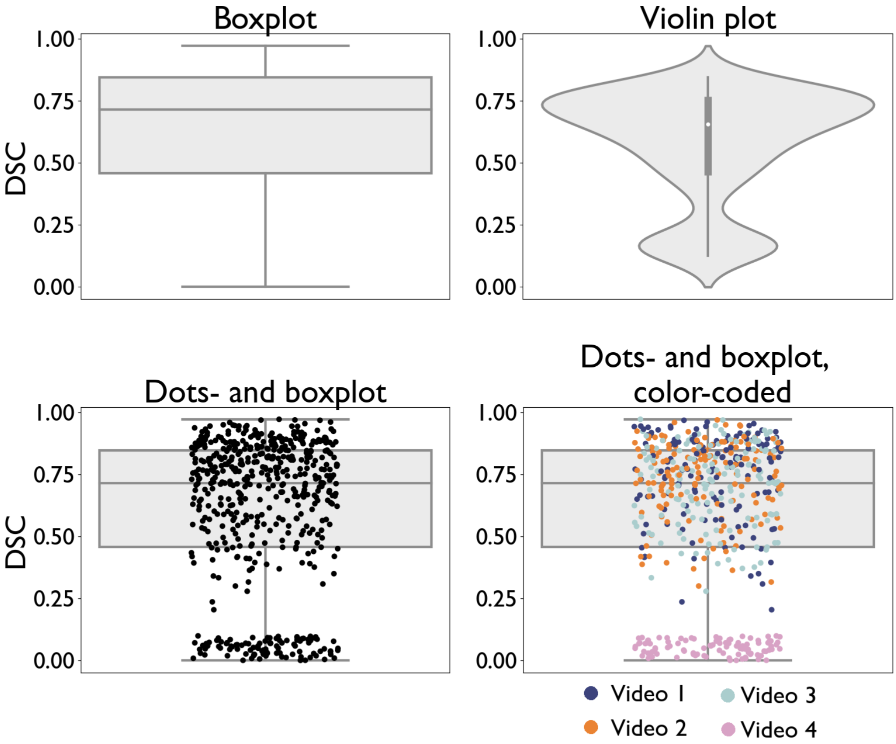


Fig. 48. Effect of different types of visualization. A single boxplot (top left) does not give sufficient information about the raw metric value distribution (here: *Dice Similarity Coefficient (DSC)*. Using a violin plot (top right) or adding the raw metric values as jittered dots on top (bottom left) adds important information. In the case of non-independent validation data, color/shape-coding helps reveal data clusters (bottom right).

Metric aggregation for invalid algorithm output (e.g.* NaN*). In challenges or benchmarking experiments, metric values are often aggregated over all test cases to produce a challenge ranking [30]. Missing data plays a crucial role when aggregating metric values and occurs primarily due to two reasons: Invalid output of the algorithm or metric routine output resulting in NaN, non-submission of single cases (by accident or even for cheating [40]). Figures 49 and 50 illustrate why a strategy on how to handle missing values may be crucial.

In the case of metrics with fixed boundaries, such as the *DSC* or the *IoU*, missing values can easily be set to the worst possible value (here: 0). For spatial distance-based measures without lower/upper bounds, the strategy of how to treat missing values is not trivial. In the case of the *HD*, for example, one may choose the maximum distance of the image or normalize the metric values to $[0, 1]$ and use the worst possible value (here: 1). Another possibility is to employ a case-based ranking scheme [30] and assign the last rank for every missing submission. Furthermore, aggregating with the mean may not be a good choice as results are unlikely to be normally distributed. Crucially, however, every choice will produce a different aggregated value (Figure 50), thus potentially affecting the ranking. Another way of handling missing values would lie in rejecting the entire submission in a challenge.

However, metric values may also be undefined (NaN)) if either reference or prediction or both are empty. In the case of empty reference and prediction, an undefined metric value (e.g. *DSC*) may be a desirable outcome and should therefore not necessarily be penalized.

Pitfall: Missing values for metrics with fixed upper/lower bounds

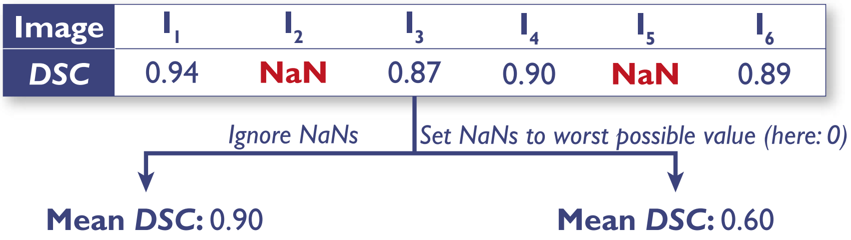


Image	I_1	I_2	I_3	I_4	I_5	I_6
DSC	0.94	NaN	0.87	0.90	NaN	0.89

Fig. 49. Effect of missing values when aggregating metric values. In this example, ignoring missing values leads to a substantially higher *Dice Similarity Coefficient (DSC)*) compared to setting missing values to the worst possible value (here: 0).

Pitfall: Missing values for metrics without fixed upper/lower bounds

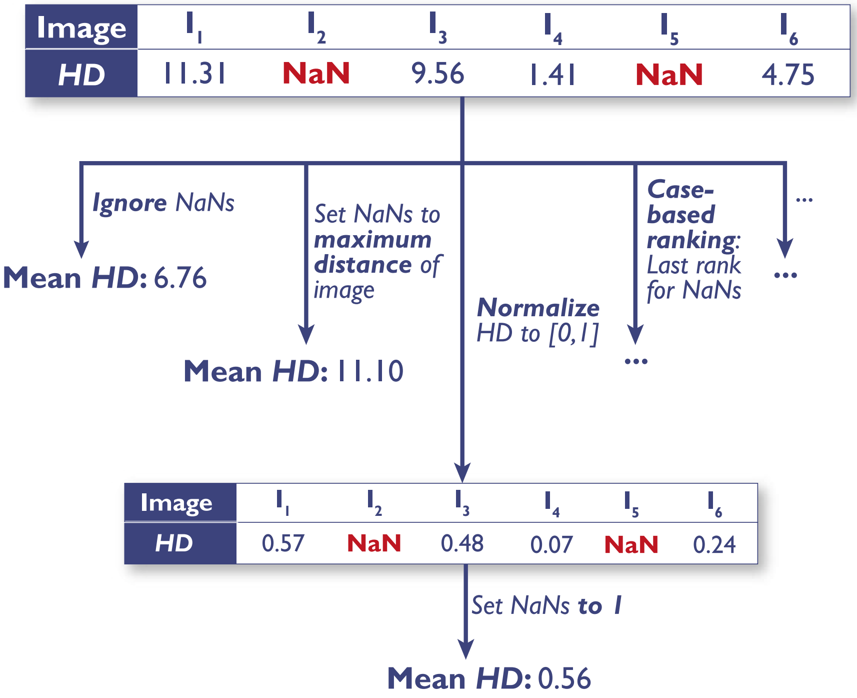


Image	I_1	I_2	I_3	I_4	I_5	I_6
HD	11.31	NaN	9.56	1.41	NaN	4.75

Image	I_1	I_2	I_3	I_4	I_5	I_6
HD	0.57	NaN	0.48	0.07	NaN	0.24

Fig. 50. Effect of missing values when aggregating metric values for metrics without fixed boundaries (here: *Hausdorff Distance (HD)*. In this example, ignoring or treating missing values in different ways leads to substantially different *HD* values.

Hierarchical data aggregation. Nowadays, most data sets are inherently hierarchically structured, meaning that the test cases are not independent. Data may, for example, come from several centers or hospitals, and for every center or even within one, different devices may be used for image acquisition, and images may be drawn from different subjects or patients. This should be kept in mind when visualizing and aggregating data points, especially if the individual tree nodes end in a large variation in the size of images. Figure 51 shows an example of five patients with an unequal number of images associated with them. Just averaging all metric values for every image would result in a high average *DSC* of 0.8. Averaging metric values per patient reveals that the *DSC* values are much higher for *Patient 1*, overruling the other patients given the high number of samples for this patient. Aggregating per patient first and averaging subsequently will resolve this issue.

Pitfall: Non-independence of validation data

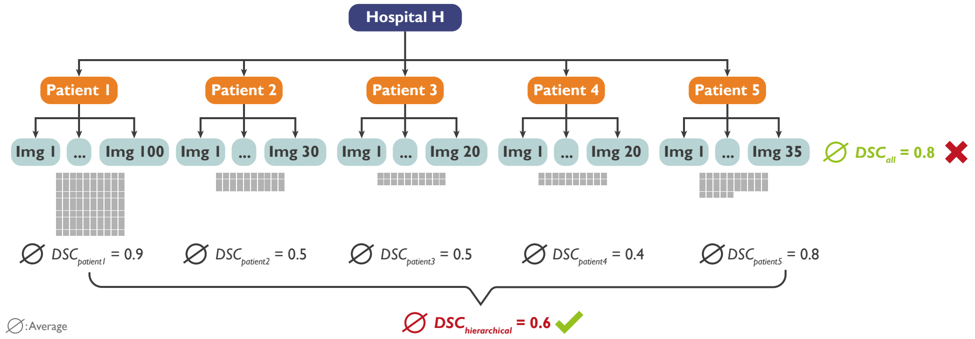


Fig. 51. Effect of non-independence of validation data, here caused by unequal numbers of data points per subject. The number of images taken from *Patient 1* is much higher compared to those acquired from *Patients 2-5*. Averaging over all *Dice Similarity Coefficient (DSC)* values results in a high averaged score. However, aggregating metric values per patient reveals much larger scores for *Patient 1* compared to the others, which would have been hidden by simple aggregation. ∅ refers to the average *DSC* values.

Aggregation per class. Similar approaches should be chosen in the presence of multiple classes in a data set. The performance may differ significantly for the individual classes, as shown in Figure 52. The background class in particular will result in a nearly perfect averaged *DSC* value, whereas the average scores for classes 2 and 3 are much lower. Aggregating over all values, not considering the class, would hide this information. An alternative approach to the problem lies in the application of metrics that explicitly handle class balance, such as using the *Generalized DSC* [45].

Pitfall: Ignoring multiple classes when aggregating

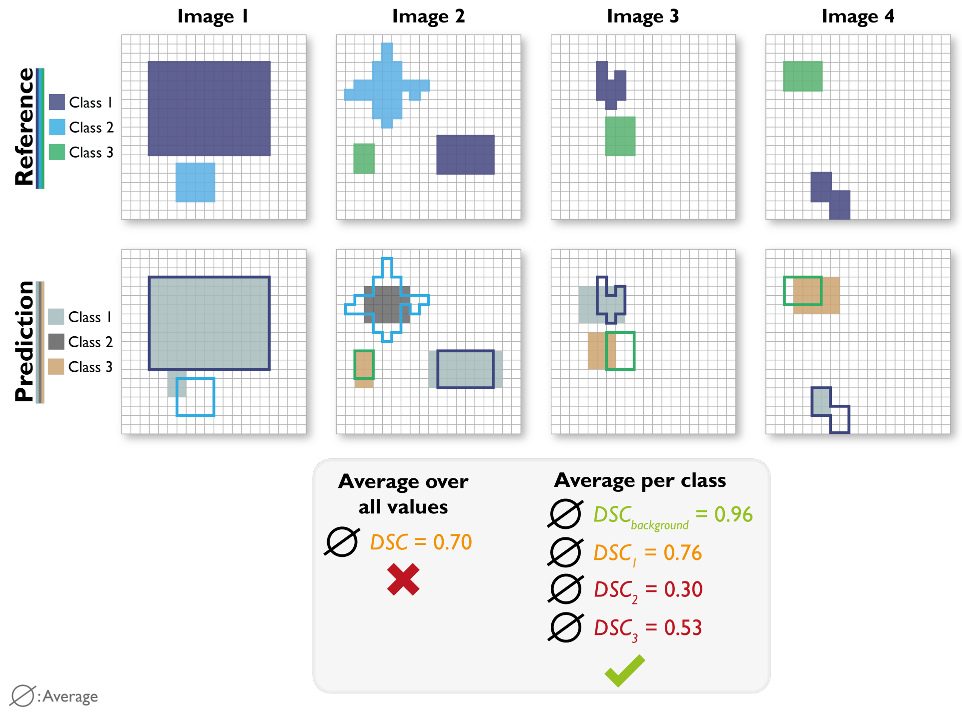
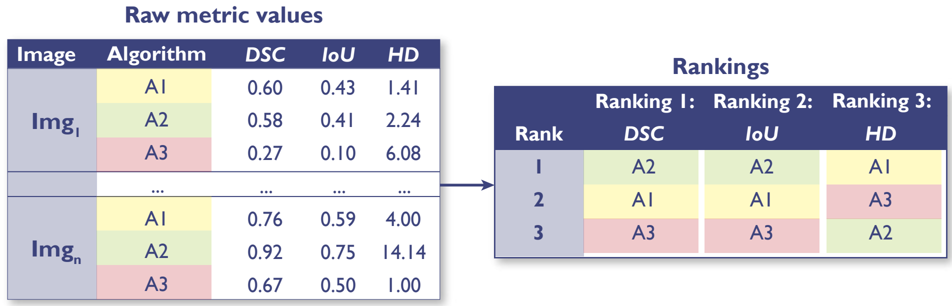


Fig. 52. Effect of ignoring the presence of multiple classes when aggregating metric values. The overall average of all *Dice Similarity Coefficient (DSC)* scores for the four images results in a *DSC* score of 0.7. Averaging per class reveals a very low performance for classes 2 and 3. ∅ refers to the average *DSC* values.

Metric combination. A single metric typically does not reflect all aspects that are essential for algorithm validation. Hence, multiple metrics with different properties are often combined. However, the selection of metrics should be well-considered as some metrics are mathematically related to each other [47, 48]. A prominent example is the *IoU* – the most popular segmentation metric in computer vision – which highly correlates with the *DSC* – the most popular segmentation metric in medical image analysis. In fact, the *IoU* and the *DSC* are mathematically related (see Sec. 3.2) [47].

Combining metrics that are related will not provide additional information for a ranking. Figure 53 illustrates how the ranking can change when adding a metric that measures different properties.

Pitfall: Related metrics



Raw metric values

Image	Algorithm	DSC	IoU	HD
Img_1	A1	0.60	0.43	1.41
	A2	0.58	0.41	2.24
	A3	0.27	0.10	6.08
	...	...	...	...
Img_n	A1	0.76	0.59	4.00
	A2	0.92	0.75	14.14
	A3	0.67	0.50	1.00

Rankings

	Ranking 1:	Ranking 2:	Ranking 3:
Rank	DSC	IoU	HD
1	A2	A2	A1
2	A1	A1	A3
3	A3	A3	A2

Fig. 53. Effect of using mathematically closely related metrics. The *Dice Similarity Coefficient (DSC)* and *Intersection over Union (IoU)* typically lead to the same ranking, whereas metrics from different families (here: *Hausdorff Distance (HD)* may lead to substantially different rankings.

9 CONCLUSION

Choosing the right metric for a specific image processing task is a nontrivial undertaking. With this (dynamic) paper, we wish to raise awareness about some of the common flaws of the most frequently used reference-based validation metrics in the field of image processing and provide guidance of their use, encouraging researchers to reconsider common workflows.

10 ACKNOWLEDGEMENTS

This work was initiated by the Helmholtz Association of German Research Centers in the scope of the Helmholtz Imaging Platform (HIP). It was further supported in part by the Intramural Research Program of the National Institutes of Health Clinical Center as well as by the National Cancer Institute (NCI) and the National Institute of Neurological Disorders and Stroke (NINDS) of the National Institutes of Health (NIH), under award numbers NCI:U01CA242871 and NINDS:R01NS042645. The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH. RMS receives royalties for patents and/or software licenses from iCAD, Philips, ScanMed, Translation Holdings, and PingAn, and has a Cooperative Research and Development Agreement with PingAn. JSR reports funding from GSK and Sanofi and fees from Travere Therapeutics and Astex Pharmaceuticals.

We would like to thank Amine Yamlahi, Marco Hübner, Dominik Michael, You Suhang, Yannick Suter, Amith Kamath and for proofreading the document.

REFERENCES

[1] Shadi AlZu'bi, Mohammed Shehab, Mahmoud Al-Ayyoub, Yaser Jararweh, and Brij Gupta. 2020. Parallel implementation for 3d medical volume fuzzy segmentation. *Pattern Recognition Letters* 130 (2020), 312–318.

[2] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. 2021. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review* 54, 1 (2021), 137–178.

[3] D Bamira and MH Picard. 2018. Imaging: Echocardiology—Assessment of Cardiac Structure and Function. (2018).

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

[5] Neil G Burnet, Simon J Thomas, Kate E Burton, and Sarah J Jefferies. 2004. Defining the tumour and target volumes for radiotherapy. *Cancer Imaging* 4, 2 (2004), 153.

[6] 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.

[7] Dev P Chakraborty and Xuetong Zhai. 2019. Analysis of data acquired using ROC paradigm and its extensions.

[8] 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.

[9] Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics* 21, 1 (2020), 1–13.

[10] 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.

[11] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning.* 233–240.

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

[13] 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.

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

[15] Steven Hicks, Inga Strüke, Vajira Thambawita, Malek Hammou, Pål Halvorsen, Michael Riegler, and Sravanthi Parasa. 2021. On evaluation metrics for medical applications of artificial intelligence. *medRxiv* (2021).

[16] 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.

[17] Mohammad Hossin and Md Nasir Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process* 5, 2 (2015), 1.

[18] 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.

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

[20] Paul F Jaeger, Simon AA Kohl, Sebastian Bickelhaupt, Fabian Isensee, Tristan Anselm Kuder, Heinz-Peter Schlemmer, and Klaus H Maier-Hein. 2020. Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. In *Machine Learning for Health Workshop.* PMLR, 171–183.

[21] Paul Ferdinand Jäger. 2020. *Challenges and Opportunities of End-to-End Learning in Medical Image Classification.* Ph.D. Dissertation. Karlsruher Institut für Technologie (KIT).

[22] Leo Joskowicz, D Cohen, N Caplan, and J Sosna. 2019. Inter-observer variability of manual contour delineation of structures in CT. *European radiology* 29, 3 (2019), 1391–1399.

[23] Jens N Kaftan, Atilla P Kiraly, Annemarie Bakai, Marco Das, Carol L Novak, and Til Aach. 2008. Fuzzy pulmonary vessel segmentation in contrast enhanced CT data. In *Medical Imaging 2008: Image Processing*, Vol. 6914. International Society for Optics and Photonics, 69141Q.

[24] 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.

[25] 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).

[26] 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.

[27] Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly* 2, 1-2 (1955), 83–97.

[28] 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.

[29] Hua Ma, Andriy I Bandos, Howard E Rockette, and David Gur. 2013. On use of partial area under the ROC curve for evaluation of diagnostic performance. *Statistics in medicine* 32, 20 (2013), 3449–3458.

[30] 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.

[31] 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.

[32] 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.

[33] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. 2014. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging* 34, 10 (2014), 1993–2024.

[34] Pawel Mlynarski, Hervé Delingette, Hamza Alghamdi, Pierre-Yves Bondiau, and Nicholas Ayache. 2020. Anatomically consistent CNN-based segmentation of organs-at-risk in cranial radiotherapy. *Journal of Medical Imaging* 7, 1 (2020), 014502.

[35] 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.

[36] Tanya Nair. 2018. *Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation.* Ph.D. Dissertation. McGill University.

[37] Nudrat Nida, Aun Irtaza, Ali Javed, Muhammad Haroon Yousaf, and Muhammad Tariq Mahmood. 2019. Melanoma lesion detection and segmentation using deep region based convolutional neural network and fuzzy C-means clustering. *International journal of medical informatics* 124 (2019), 37–48.

[38] 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.

[39] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. 2018. Localization recall precision (LRP): A new performance metric for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 504–519.

[40] 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.

[41] Azriel Rosenfeld and John L Pfaltz. 1966. Sequential operations in digital picture processing. *Journal of the ACM (JACM)* 13, 4 (1966), 471–494.

[42] Tobias Roß, Annika Reinke, Peter M Full, Martin Wagner, Hannes Kenngott, Martin Apitz, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, et al. 2021. Comparative validation of multi-instance instrument segmentation in endoscopy: results of the ROBUST-MIS 2019 challenge. *Medical image analysis* 70 (2021), 101920.

[43] Anindo Saha, Joeran Bosma, Jasper Linmans, Matin Hosseinzadeh, and Henkjan Huisman. 2021. Anatomical and Diagnostic Bayesian Segmentation in Prostate MRI – Should Different Clinical Objectives Mandate Different Loss Functions? *arXiv preprint arXiv:2110.12889* (2021).

[44] 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.

[45] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 240–248.

[46] Yanmin Sun, Andrew K. C. Wong, and Mohamed S. Kamel. 2009. Classification of Imbalanced Data: a Review. *Int. J. Pattern Recognit. Artif. Intell.* 23 (2009), 687–719.

[47] 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.

[48] Abdel Aziz Taha, Allan Hanbury, and Oscar A Jimenez del Toro. 2014. A formal method for selecting evaluation metrics for image segmentation. In *2014 IEEE international conference on image processing (ICIP)*. IEEE, 932–936.

[49] Uchila N Umesh, Robert A Peterson, and Matthew H Sauber. 1989. Interjudge agreement and the maximum value of kappa. *Educational and Psychological Measurement* 49, 4 (1989), 835–850.

[50] 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.

[51] Bram Van Ginneken, Tobias Heimann, and Martin Styner. 2007. 3D segmentation in the clinic: A grand challenge. In *MICCAI Workshop on 3D Segmentation in the Clinic: A Grand Challenge*, Vol. 1. 7–15.

[52] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* 17, 3 (2020), 261–272.

[53] 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.

APPENDIX

A FULL AUTHOR AFFILIATIONS

Annika Reinke, a.reinke@dkfz.de, German Cancer Research Center (DKFZ), Div. Intelligent Medical Systems and HIP Helmholtz Imaging Platform, 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 HIP Helmholtz Imaging Platform, Heidelberg, Germany
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
Matthias Eisenmann, German Cancer Research Center (DKFZ), Div. Intelligent Medical Systems and HIP Helmholtz Imaging Platform, Heidelberg, Germany
Tim Rädsch, German Cancer Research Center (DKFZ), Div. Intelligent Medical Systems and HIP Helmholtz Imaging Platform, Heidelberg, Germany and understandAI GmbH, Karlsruhe, Germany
Michael Baumgartner, German Cancer Research Center (DKFZ), Div. Medical Image Computing, 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, 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
Ben Glocker, Imperial College London, Biomedical Image Analysis Group, Department of Computing, London, UK
Patrick Godau, German Cancer Research Center (DKFZ), Div. Intelligent Medical Systems, Heidelberg, Germany and Heidelberg University, Faculty of Mathematics and Computer Science, Heidelberg, Germany
Fred Hamprecht, Heidelberg University, Heidelberg Collaboratory for Image Processing (HCI), Interdisciplinary Center for Scientific Computing (IWR), Heidelberg, Germany
Daniel A. Hashimoto, Case Western Reserve University School of Medicine, University Hospitals Cleveland Medical Center, Cleveland, Ohio, USA
Doreen Heckmann-Nötzel, German Cancer Research Center (DKFZ), Div. Intelligent Medical Systems, Heidelberg, Germany
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, University Medical Center Utrecht, Department of Radiology, Utrecht, The Netherlands
Fabian Isensee, German Cancer Research Center (DKFZ), HIP Applied Computer Vision Lab, Division of Medical Image Computing, Heidelberg, Germany
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
Alexandros Karargyris, IHU Strasbourg, Strasbourg, France
Alan Karthikesalingam, Google Health Deepmind, London, UK
Bernhard Kainz, Imperial College London, Department of Computing, Faculty of Engineering, London, UK
Emre Kavur, German Cancer Research Center (DKFZ), Div. Intelligent Medical Systems, HIP Applied Computer Vision Lab, Division of Medical Image Computing, Heidelberg, Germany
Hannes Kenngott, Heidelberg University Hospital, Department of General, Visceral and Transplantation Surgery, Heidelberg, Germany
Jens Kleesiek, University Medicine Essen, Translational Image-guided Oncology (TIO), Institute for AI in Medicine (IKIM), Essen, Germany
Thijs Kooi, Lunit Inc, Seoul, South Korea
Michal Kozubek, Masaryk University, Centre for Biomedical Image Analysis, Brno, Czech Republic
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, 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, US
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
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
Gorkem Polat, Middle East Technical University, Graduate School of Informatics, Ankara, Turkey
Nasir Rajpoot, University of Warwick, Tissue Image Analytics Laboratory, Department of Computer Science, Coventry, West Midlands, UK
Mauricio Reyes, University of Bern, ARTORG Center for Biomedical Engineering Research, Bern, Switzerland
Nicola Rieke, NVIDIA GmbH, Munich, Germany
Michael A. Riegler, Simula Metropolitan Center for Digital Engineering, Oslo, Norway and UiT The Arctic University of Norway, Oslo, Norway
Hassan Rivaz, Concordia University, Department of Electrical and Computer Engineering, Montreal, Canada
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
Julien Schroeter, McGill University, Centre for Intelligent Machines, Montreal, Canada
Anindo Saha, Radboud University Medical Center, Diagnostic Image Analysis Group, Nijmegen, The Netherlands
Shravya Shetty, Google, Google Health, Palo Alto, US
Bram Stieltjes, University Hospital of Basel, Department of Radiology, Basel, Switzerland
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
Bram van Ginneken, Fraunhofer MEVIS, Bremen, Germany and Radboud University Medical Center, Radboud Institute for Health Sciences, Nijmegen, The Netherlands
Gaël Varoquaux, INRIA Saclay-Île de France, Parietal Project-team, Palaiseau, France
Manuel Wiesenfarth, German Cancer Research Center (DKFZ), Div. Biostatistics, Heidelberg, Germany
Ziv R. Yaniv, Bioinformatics and Computational Bioscience Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA
Annette Kopp-Schneider, German Cancer Research Center (DKFZ), Div. Biostatistics, Heidelberg, Germany
Paul Jäger, German Cancer Research Center (DKFZ), Interactive Machine Learning Group, Heidelberg, Germany
Lena Maier-Hein, German Cancer Research Center (DKFZ), Div. Intelligent Medical Systems and HIP Helmholtz Imaging Platform, Heidelberg, Germany and Heidelberg University, Faculty of Mathematics and Computer Science and Medical Faculty, Heidelberg, Germany.