SCISPACE
formerly Typeset

Open access • Journal Article • DOI:10.1109/TMI.2020.3002417

Optimization for Medical Image Segmentation: Theory and Practice When Evaluating With Dice Score or Jaccard Index — Source link

Tom Eelbode, Jeroen Bertels, Maxim Berman, Dirk Vandermeulen ...+3 more authors

Institutions: Katholieke Universiteit Leuven

Published on: 15 Jun 2020 - IEEE Transactions on Medical Imaging (Institute of Electrical and Electronics Engineers (IEEE))

Topics: Jaccard index, Dice and Image segmentation

Related papers:

- U-Net: Convolutional Networks for Biomedical Image Segmentation
- Optimization for Medical Image Segmentation: Theory and Practice when evaluating with Dice Score or Jaccard Index
- Fully convolutional networks for semantic segmentation
- Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory and Practice
- The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)

Share this paper:

View more about this paper here: https://typeset.io/papers/optimization-for-medical-image-segmentation-theory-and-4beanbugqv

Optimization for Medical Image Segmentation: Theory and Practice when evaluating with Dice Score or Jaccard Index

Tom Eelbode, Jeroen Bertels, Maxim Berman, Dirk Vandermeulen, *Fellow, IEEE,* Frederik Maes, *Senior Member, IEEE,* Raf Bisschops, and Matthew B. Blaschko

arXiv:2010.13499v1 [eess.IV] 26 Oct 2020

***Abstract*—In many medical imaging and classical computer vision tasks, the Dice score and Jaccard index are used to evaluate the segmentation performance. Despite the existence and great empirical success of metric-sensitive losses, i.e. relaxations of these metrics such as soft Dice, soft Jaccard and Lovász-Softmax, many researchers still use per-pixel losses, such as (weighted) cross-entropy to train CNNs for segmentation. Therefore, the target metric is in many cases not directly optimized. We investigate from a theoretical perspective, the relation within the group of metric-sensitive loss functions and question the existence of an optimal weighting scheme for weighted cross-entropy to optimize the Dice score and Jaccard index at test time. We find that the Dice score and Jaccard index approximate each other relatively and absolutely, but we find no such approximation for a weighted Hamming similarity. For the Tversky loss, the approximation gets monotonically worse when deviating from the trivial weight setting where soft Tversky equals soft Dice. We verify these results empirically in an extensive validation on six medical segmentation tasks and can confirm that metric-sensitive losses are superior to cross-entropy based loss functions in case of evaluation with Dice Score or Jaccard Index. This further holds in a multi-class setting, and across different object sizes and foreground/background ratios. These results encourage a wider adoption of metric-sensitive loss functions for medical segmentation tasks where the performance measure of interest is the Dice score or Jaccard index.**

***Index Terms*—Dice, Jaccard, Risk minimization, Cross-entropy, Tversky**

Copyright (c) 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Submitted for review on November 16, 2019.

Published on June 15, 2020.

T. Eelbode and J. Bertels contributed equally to this work.

T. Eelbode, J. Bertels, M. Berman, D. Vandermeulen, F. Maes and M. Blaschko are with the Department of Electrical Engineering (ESAT), Center for Processing Speech and Images, KU Leuven, Belgium. E-mail: tom.eelbode@kuleuven.be

R. Bisschops is with the Department of Gastroenterology and Hepatology, UZ Leuven, Belgium.

This work is funded in part by Internal Funds KU Leuven (grant # C24/18/047). The computational resources were partly provided by the Flemish Supercomputer Center (VSC). J.B. is part of NEXIS, a project that has received funding from the European Union's Horizon 2020 Research and Innovations Programme (grant # 780026). R.B. is supported by FWO and Fujifilm. M.B. and M.B.B. acknowledge support from FWO (grant # G0A2716N), an Amazon Research Award, an NVIDIA GPU grant, and the Facebook AI Research Partnership. This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. The authors thank H. Willekens, C. Camps, C. Hassan, E. Coron, P. Bhandari, H. Neumann and O. Pech for their effort and collaboration.

I. Introduction

In medical image segmentation, the most commonly used performance metrics are the Dice score and Jaccard index [1], [2], [3], [4], [5]. These metrics are used rather than pixel-wise accuracy because they are a more adequate indicator for the perceptual quality of a segmentation. This is also recognized by Zijdenbos et al. who state that the Dice score better reflects size and localization agreement for object segmentation [6]. They also show that Dice score is a special case of the Kappa statistic in case the number of background voxels greatly outweighs that of the foreground voxels.

When we train a learning-based method for image segmentation, we are performing risk minimization. In this setting, it is important that we optimize a loss function during training which we will also use for evaluation during test time [7]. However, in the MICCAI 2018 proceedings, 68 out of 96 learning-based segmentation papers used a loss function which does not directly optimize Dice score or Jaccard index even though evaluation was performed with one of these two metrics (more details in Appendix A). Pixel-wise (weighted) cross-entropy loss is still frequently used in these papers [1], [8]. Nevertheless, differentiable approximations for Dice score and Jaccard index have been proposed for training discriminative models with a gradient-based optimization algorithm, such as stochastic gradient descent (SGD). For example, the soft Dice [9] and soft Jaccard [10], [11] are relaxations for their respective metrics and can be used to surrogate Dice score and Jaccard index during training. The Lovász-softmax is a more recent convex extension for the Jaccard index [12].

This great variety of loss functions being used in the literature raises the question whether there is any rationale for choosing a specific group of loss functions over the others and if this choice has an influence on the quality of the predictions. In this work, we theoretically investigate the link between Dice score and Jaccard index and find that they approximate each other under risk minimization. Despite the existence of this approximation bound, we can prove that no such approximation exists for cross-entropy and that no weighting will allow it to surrogate Dice or Jaccard. Analogously, we look for an optimal weighting of false positives and false negatives in the Tversky loss and find that this is the case when the latter is equivalent to the Dice loss. We then empirically validate these theories on six different medical image segmentation tasks. We find that the performance when training with (weighted)

cross-entropy is inferior to using any of the surrogates for Dice and Jaccard. However, there is no statistically significant difference between these surrogates.

This article is an extended version of [13]. We continued the analysis on an extra dataset from the WMH 2017 challenge, in which the images contain multiple small objects with a very low total foreground/background ratio. Additionally, we ran all experiment on this dataset under two different network architectures: additional to U-Net (results shown as WM17) we used DeepMedic [1] (results shown as WM17^{DM}). Furthermore, these runs were now executed in a patch-wise setting due to the larger image sizes. More details can be found under Sec. III and the results are listed in Tables I, II & III. We added an analysis of the Tversky loss (Eqs. (6) & (14), Prop. II.3, and experiments in Sec. IV-B), a comparison between the L^1 and L^2-based soft Dice variants (Sec. IV-A), visualizations of segmentation results and qualitative discussion in Sec. IV-B, multi-class segmentation experiments on the BRATS dataset, additional experiments analyzing the effect of object size on the POLYPS dataset in Sec. IV-D, and extended discussions.

II. Risk minimization with Dice and related similarities

When performing discriminative training of machine learning methods, such as stochastic gradient descent for a CNN [14], we are performing risk minimization. To learn a mapping f from an observed input x to a hidden variable y, empirical risk minimization optimizes the expectation of a loss function over a finite training set:

$$\arg\min_{f\in\mathcal{F}} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\ell(f(x_i),y_i)}_{=:\hat{\mathcal{R}}(f)}, \tag{1}$$

where ℓ is a loss function and $\mathcal{F}$ is a function class of interest, e.g. the set of functions that can be represented by a neural network with a given topology. We will denote the empirical distribution arising from a sample $\mathcal{S} := \{(x_i,y_i)\}_{1\leq i\leq n}$ of size n as P_n, and we may equivalently denote $\hat{\mathcal{R}}(f) = \mathbb{E}_{(x,y)\sim P_n}[\ell(x,y)]$.

In binary medical image segmentation, y can be thought of as a set of pixels labeled as foreground. It is therefore well defined to consider set theoretic notions such as $y\cap\tilde{y}$ for two different segmentations. This motivates the use of multiple set theoretic similarity measures between the ground truth segmentation y and the predicted segmentation $\tilde{y}$ including the Dice score D, the Jaccard index J, the Hamming similarity H, the weighted Hamming similarity H_γ, and the Tversky index $T_{\alpha,\beta}$:

$$D(y,\tilde{y}) := \frac{2|y\cap\tilde{y}|}{|y|+|\tilde{y}|}, \tag{2}$$

$$J(y,\tilde{y}) := \frac{|y\cap\tilde{y}|}{|y\cup\tilde{y}|}, \tag{3}$$

$$H(y,\tilde{y}) := 1-\frac{|y\setminus\tilde{y}|+|\tilde{y}\setminus y|}{d}, \tag{4}$$

$$H_\gamma(y,\tilde{y}) := 1-\gamma\frac{|y\setminus\tilde{y}|}{|y|}-(1-\gamma)\frac{|\tilde{y}\setminus y|}{d-|y|}, \tag{5}$$

$$T_{\alpha,\beta}(y,\tilde{y}) := \frac{|y\cap\tilde{y}|}{|y\cap\tilde{y}|+\alpha|\tilde{y}\setminus y|+\beta|y\setminus\tilde{y}|}, \tag{6}$$

where d denotes the number of pixels, $0\leq\gamma\leq 1$, $\alpha>0$, and $\beta>0$. We note that all these similarities are between 0 and 1, that H_γ generalizes H with equality when $\gamma=\frac{|y|}{d}$, and that the Tversky index is equivalent to the Dice score when $\alpha=\beta=\frac{1}{2}$ and equivalent to Jaccard when $\alpha=\beta=1$. A further important relationship is that between the Jaccard index and the Dice score. It is well known that

$$J(y,\tilde{y}) = \frac{D(y,\tilde{y})}{2-D(y,\tilde{y})} \text{ and } D(y,\tilde{y}) = \frac{2J(y,\tilde{y})}{1+J(y,\tilde{y})}. \tag{7}$$

Indeed, in the risk minimization framework for medical image segmentation, there are numerous examples where each of these measures are optimized [15], [9], [16].

A. Differentiable loss surrogates

The similarities in Equations (2)-(6) are over a discrete space of binary segmentations, while the output of a predictive function, e.g. given by a neural network, is continuous. In order to perform gradient descent by backpropagation [17], [14] on the empirical risk (Equation (1)), we must replace a discrete loss

$$1-S:\{0,1\}^d\times\{0,1\}^d\to\mathbb{R}_+, \tag{8}$$

where S is any of the similarities in (2)-(6), with a surrogate that is differentiable in the predicted label

$$\ell:\{0,1\}^d\times\mathbb{R}^d\to\mathbb{R}_+. \tag{9}$$

For the Hamming similarity (Equation (4)), cross-entropy loss and other convex surrogates are statistically consistent [18]. To optimize the weighted Hamming similarity, one may employ weighted cross-entropy [19]. In this section, we will encode sets by binary vectors $y,\tilde{y}\in\{0,1\}^d$, allowing us to write the weighted cross entropy loss as:

$$\Delta_{H_\gamma}(y,\tilde{y}) = -\gamma\langle y,\log(\tilde{y})\rangle-(1-\gamma)\langle\mathbf{1}-y,\log(\mathbf{1}-\tilde{y})\rangle, \tag{10}$$

where Δ_S denotes a differentiable surrogate for a given similarity S, $\langle\cdot,\cdot\rangle$ denotes the canonical inner product, the logarithm is taken element-wise over the vector, and $\mathbf{1}$ denotes a d-dimensional vector of ones. To make the surrogate differentiable, one relaxes $\tilde{y}$ to take continuous values $\tilde{y}\in[0,1]^d$, where unbounded scores are constrained to lie in $[0,1]$ via a softmax operation [14, Equation (4.1)].

Similarly, differentiable surrogates have been proposed for the Dice score (e.g. soft Dice [20], [9]), Jaccard index (e.g.

soft Jaccard [21] or Lovász-softmax [12]), and Tversky index (e.g. soft Tversky [16]).

The soft Dice loss [20], [9] generalizes the Dice loss by noting that we have the identity $|y \cap \tilde{y}| = \langle y, \tilde{y} \rangle$. Furthermore, $|y| = |y \cap y| = \langle y, y \rangle$. We may now write the Dice loss as

$$\Delta_D(y, \tilde{y}) = 1 - \frac{2|y \cap \tilde{y}|}{|y| + |\tilde{y}|} \tag{11}$$

$$= 1 - \frac{2\langle y, \tilde{y} \rangle}{\langle y, y \rangle + \langle \tilde{y}, \tilde{y} \rangle}. \tag{12}$$

The soft Dice loss is simply the relaxation where $\tilde{y} \in [0, 1]^d$. We note that multiple generalizations are possible, e.g. by replacing $\langle \tilde{y}, \tilde{y} \rangle = \|\tilde{y}\|_2^2$ with $\|\tilde{y}\|_p^p$ for any L^p norm, and the statistical and optimization properties of the resulting surrogate are dependent on these choices. We will see below that $\|\tilde{y}\|_1$ is used in related surrogate definitions.

The soft Jaccard loss [21] is defined using similar ideas to the soft Dice loss:

$$\Delta_J(y, \tilde{y}) = 1 - \frac{\langle y, \tilde{y} \rangle}{\|y\|_1 + \|\tilde{y}\|_1 - \langle y, \tilde{y} \rangle}, \tag{13}$$

although as above, alternative relaxations of the discrete loss are possible, e.g. by replacing the L^1 norm with a squared L^2 norm.

The Lovász softmax is a strategy for constructing loss surrogates for submodular losses [12]. The Jaccard loss has been proven to be submodular [22, Proposition 11], while the Dice loss is not [23, Proposition 6]. As with other surrogates described in this section, the Lovász softmax first projects pixel scores into $[0, 1]$ using a softmax operation, and then computes the Lovász extension [24] of the submodular loss function, which is convex and differentiable almost everywhere, giving it favorable properties as a loss surrogate.

We finally describe the soft Tversky loss function:

$$\begin{aligned} \Delta_{T_{\alpha,\beta}}(y, \tilde{y}) &= 1 - \frac{\langle y, \tilde{y} \rangle}{\langle y, \tilde{y} \rangle + \alpha \langle \mathbf{1} - y, \tilde{y} \rangle + \beta \langle y, \mathbf{1} - \tilde{y} \rangle} \\ &= 1 - \frac{\langle y, \tilde{y} \rangle}{(1 - \alpha - \beta)\langle y, \tilde{y} \rangle + \alpha \|\tilde{y}\|_1 + \beta \|y\|_1}. \end{aligned} \tag{14}$$

Next, we discuss the absolute and relative approximations between Dice, Jaccard, and Tversky, and inspect the existence of an approximation through a weighted Hamming similarity.

B. Approximation bounds

Definition II.1 (Absolute approximation)**.** A similarity S is absolutely approximated by $\tilde{S}$ with error $\varepsilon \geq 0$ if the following holds for all y and $\tilde{y}$:

$$|S(y, \tilde{y}) - \tilde{S}(y, \tilde{y})| \leq \varepsilon. \tag{15}$$

Definition II.2 (Relative approximation)**.** A similarity S is relatively approximated by $\tilde{S}$ with error $\varepsilon \geq 0$ if the following holds for all y and $\tilde{y}$:

$$\frac{\tilde{S}(y, \tilde{y})}{1 + \varepsilon} \leq S(y, \tilde{y}) \leq \tilde{S}(y, \tilde{y})(1 + \varepsilon). \tag{16}$$

We note that both notions of approximation are symmetric in S and $\tilde{S}$.

Proposition II.1. J and D approximate each other with relative error of 1 and absolute error of $3 - 2\sqrt{2} = 0.17157\ldots$.

Proof. The relative error between J and D is given by (cf. Equation (7))

$$\min_{\varepsilon \geq 0} \varepsilon, \text{ s.t. } x \leq \frac{x}{2 - x}(1 + \varepsilon), \ \forall\ 0 \leq x \leq 1. \tag{17}$$

$$x \leq \frac{x}{2 - x}(1 + \varepsilon) \implies 1 - x \leq \varepsilon \implies \varepsilon = 1. \tag{18}$$

The absolute error between J and D is given by

$$\varepsilon = \sup_{0 \leq x \leq 1} \left| x - \frac{x}{2 - x} \right| = 3 - 2\sqrt{2}, \tag{19}$$

which can be verified straightforwardly by first order conditions:

$$\frac{\partial}{\partial x}\left(x - \frac{x}{2 - x} \right) = 0 \implies (2 - x)^2 - 2 = 0 \tag{20}$$

$$\implies x = 2 - \sqrt{2}. \tag{21}$$

□

Proposition II.2. D and H_γ (where γ is chosen to minimize the approximation factor between D and H_γ) do not relatively approximate each other, and absolutely approximate each other with an error of 1. We note that the absolute error bound is trivial as D and H_γ are both similarities in the range $[0, 1]$.

Proof. For relative error, consider the case that $|y \setminus \tilde{y}| = 0$, $|\tilde{y} \setminus y| = \alpha d$, and $|y \cap \tilde{y}| = \alpha^2 d$ for some $0 \leq \alpha < \frac{\sqrt{5}-1}{2}$ and d the number of pixels:

$$\begin{aligned} \inf_{\gamma} \sup_{y, \tilde{y}} \ & 1 - \gamma \frac{|y \setminus \tilde{y}|}{|y|} - (1 - \gamma) \frac{|\tilde{y} \setminus y|}{d - |y|} \\ & - \frac{2|y \cap \tilde{y}|}{|y \triangle \tilde{y}| + 2|y \cap \tilde{y}|}(1 + \varepsilon) \leq 0 \end{aligned} \tag{22}$$

$$\implies \sup_{0 \leq \alpha < \frac{\sqrt{5}-1}{2}} 1 - \frac{\alpha}{1 - \alpha^2} - \frac{2\alpha^2}{\alpha + 2\alpha^2}(1 + \varepsilon) \leq 0. \tag{23}$$

If we let $\alpha \to 0$, it must be the case that $\varepsilon \to \infty$.

To show that the absolute approximation error is 1, we similarly take

$$\lim_{\alpha \to 0} 1 - \frac{\alpha}{1 - \alpha^2} - \frac{2\alpha}{1 + 2\alpha} = 1. \qquad \square \tag{24}$$

□

Corollary II.1. D and H do not relatively approximate each other, and absolutely approximate each other with an error of 1.

From these bounds, we see that a (weighted) binary loss can be an arbitrarily bad approximation for Dice when segmenting small objects, while the Jaccard loss gives multiplicative and additive approximation guarantees. Furthermore, Eq. (7) implies that

$$1 - D(y, \tilde{y}) \leq 1 - J(y, \tilde{y}) \implies \tag{25}$$

$$\mathbb{E}_{(x,y) \sim P_n}[1 - D(y, f(x))] \leq \mathbb{E}_{(x,y) \sim P_n}[1 - J(y, f(x))] \tag{26}$$

and optimization with risk computed with the Jaccard loss minimizes an upper bound on risk computed with the Dice loss. Similarly setting $\varphi(x) = 2x/(1+x)$, by application of Jensen's inequality we arrive at

$$\mathbb{E}_{(x,y)\sim P_n}[1 - J(y, f(x))] = \tag{27}$$

$$\mathbb{E}_{(x,y)\sim P_n}[\varphi(1 - D(y, f(x)))]$$

$$\leq \varphi(\mathbb{E}_{(x,y)\sim P_n}[1 - D(y, f(x))]) \tag{28}$$

and optimizing the Dice loss minimizes an upper bound on the Jaccard loss as φ is a monotonic function over $[0, 1]$.

Proposition II.3. The Tversky index approximates the Dice score with an absolute error of $\max\left(\left|\frac{\sqrt{2\alpha}-1}{\sqrt{2\alpha}+1}\right|, \left|\frac{\sqrt{2\beta}-1}{\sqrt{2\beta}+1}\right|\right)$ and a relative error of $\max(2\alpha, 2\beta, 0.5\alpha^{-1}, 0.5\beta^{-1}) - 1$.

Proof. The absolute error is given by

$$\sup_{y,\tilde{y}} |D(y, \tilde{y}) - T_{\alpha,\beta}(y, \tilde{y})| = \tag{29}$$

$$\sup_{y,\tilde{y}} \left| \frac{2|y \cap \tilde{y}|}{|y| + |\tilde{y}|} - \frac{|y \cap \tilde{y}|}{|y \cap \tilde{y}| + \alpha|\tilde{y} \setminus y| + \beta|y \setminus \tilde{y}|} \right| = \tag{30}$$

$$\sup_{a,b,c\geq 0,\ a+b+c=1} \left| \frac{b}{b + \alpha a + \beta c} - \frac{b}{b + 0.5a + 0.5c} \right|. \tag{31}$$

If $\alpha = \beta$, we can simplify this to an equation of one variable

$$\sup_{0\leq b\leq 1} \left| \frac{b}{b + \alpha(1-b)} - \frac{b}{b + 0.5(1-b)} \right|. \tag{32}$$

From first order conditions, this is maximized when $b = \frac{\alpha - \sqrt{2\alpha}}{\alpha - 2}$, and the resulting error is therefore $\left|\frac{\sqrt{2\alpha}-1}{\sqrt{2\alpha}+1}\right|$.

For the case that $\alpha \neq \beta$, we note that the directional derivative in the direction $a - c$ has constant sign (i.e. it is monotonic), indicating that at the maximum over the simplex, $a = 0$ or $c = 0$. This indicates that the problem simplifies to taking the maximum over the special cases that $a = 0$ or $c = 0$, which in both cases simplifies to Equation (32), and the absolute approximation error is $\max\left(\left|\frac{\sqrt{2\alpha}-1}{\sqrt{2\alpha}+1}\right|, \left|\frac{\sqrt{2\beta}-1}{\sqrt{2\beta}+1}\right|\right)$.

The relative error can similarly be computed as

$$1 + \varepsilon = \max\left(\sup_{y,\tilde{y}} \frac{T_{\alpha,\beta}(y, \tilde{y})}{D(y, \tilde{y})}, \sup_{y,\tilde{y}} \frac{D(y, \tilde{y})}{T_{\alpha,\beta}(y, \tilde{y}}\right) \tag{33}$$

$$= \max\left(\sup_{a,b,c} \frac{b + 0.5a + 0.5c}{b + \alpha a + \beta c}, \sup_{a,b,c} \frac{b + \alpha a + \beta c}{b + 0.5a + 0.5c}\right), \tag{34}$$

where in each sup, $0 < a + b + c \leq 1$ and $a, b, c \geq 0$. The first sup is maximized by determining $\max(0.5\alpha^{-1}, 0.5\beta^{-1})$ and setting $a = 1$ or $c = 1$, respectively. The second sup is similarly maximized by determining $\max(2\alpha, 2\beta)$ and setting either $a = 1$ or $c = 1$, respectively. The approximation error can then be determined by maximizing over all 4 cases. □

The absolute and relative errors are shown in Fig. 1. The Tversky index does give non-trivial approximation errors for the Dice score for both absolute and relative error. However, in both cases, the error bounds are increasing as α and β deviate from 0.5. This may provide the ability to trade off the performance on the Dice measure with that of other quality measures, but optimal performance on the Dice measure will be achieved at $\alpha = \beta = 0.5$, discounting eventual small sample effects or biases introduced by differentiable surrogates. Furthermore, if a trade-off between Dice and other quality measures is desired, one may simply take a weighted combination of loss functions for each of the desired metrics, directly encoding the relative importance of each.

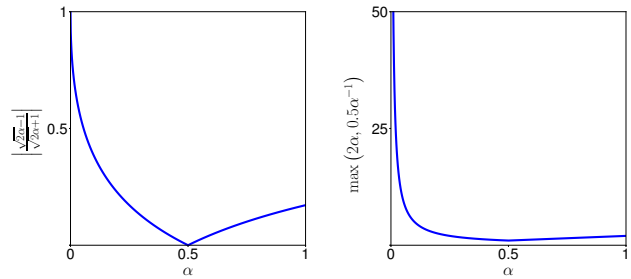


Fig. 1: Absolute (left) and relative (right) errors for the Dice score and Tversky indices as a function of the Tversky parameter that maximally deviates from 0.5 (see Proposition II.3).

In this section, we have shown that Dice, Jaccard, and Tversky similarities give approximation bounds to each other, while cross-entropy and weighted cross-entropy do not approximate the Dice score. This indicates that risk minimization with Dice, Jaccard, or Tversky will give some performance guarantees when measuring with the Dice score at test time. On the other hand, training with (weighted) cross-entropy can yield arbitrarily poor results when evaluating with the Dice score. Approximation bounds from the Tversky index indicate increasing degradation in performance on the Dice score as parameters deviate from $\alpha = \beta = 0.5$ suggesting best performance when training with the Dice loss, in correspondence with the risk minimization principle [7]. In the next section, we describe how we validate these theoretical results empirically in a range of medical image segmentation settings.

III. Empirical setup

To investigate whether the aforementioned properties hold in practice, we investigate the performance of segmentation networks trained with different loss functions: cross-entropy (CE), weighted cross-entropy (wCE), soft Dice (sDice), soft Jaccard (sJaccard), Lovász-sigmoid (Lovász; this is the binary equivalent to the Lovász-softmax) and soft Tversky loss (sTversky). The empirical validation is performed on the following six binary medical segmentation tasks:

- **BR18** - whole tumour segmentation on pre-operative MRI (BRATS 2018 [25], [26], [27], [3]) - $N = 285$
- **IS17** - final infarction after ischemic stroke treatment segmentation from acute MRI perfusion (ISLES 2017 [4]) - $N = 43$
- **IS18** - ischemic core segmentation on acute CT perfusion (ISLES 2018 [5]) - $N = 94$
- **MO17** - lower-left third molar segmentation on panoramic dental radiographs [28] - $N = 400$
- **PO18** - colorectal polyp segmentation on colonoscopy images [29] - $N = 1166$

- **WM17** - segmentation of white matter hyperintensities on MRI (WMH 2017 [30]) - $N = 60$

The datasets BR18, IS17, IS18 and WM17 are publicly available and 3D. The datasets MO17 and PO18 are in-house and 2D. An additional empirical validation for a multi-class segmentation task is given in Sec. IV-D for BRATS 2018 with segmentation of different glioma sub-regions. The segmentation volumes considered for evaluation were the whole tumor (WT), tumor core (TC) and enhancing tumor (ET). The inclusion of this allows us to empirically validate if the conclusions that will be drawn for binary image segmentation can still hold for individual structures, when the de facto optimization objective becomes the average loss across multiple structures.

A. Data preprocessing and network architectures

We use a 3D U-Net-like [2] architecture for the four 3D datasets BR18, IS17, IS18 and WM17. For MO17, we use the same architecture but with the 3D convolutions substituted by 2D convolutions. More specifically, we start from the No New-Net [31] architecture, which was a top-ranked implementation during BRATS 2018. We have one level less and compensate with 3x3(x3) pooling and upsampling to keep sufficient field-of-view. The number of filters evolve similarly, starting with 20 in the first layer. For a richer comparison, a fully convolutional network with a VGG16 backbone and atrous convolutions, pretrained on ImageNet, i.e. DeepLab [32], [33], is used for PO18. On WM17, additional runs with a DeepMedic-like [1] network architecture were done, further referred to as WM17^{DM}, depending on the context. In case of the multi-class segmentation task, we adapt the model with three sigmoid layers in the final layer. Thereby allowing a single voxel to be classified into one or more of the evaluation volumes.

We make use of all the available image modalities in each dataset to construct the input tensors (we exclude perfusion data for IS17 and IS18). In order to fit the memory constraints for whole-image processing, these input tensors are resized and cropped. The images of all public datasets are resampled to an isotropic voxel-size of 2 mm and, except for WM17, cropped to a fixed size of 136x136x82. For WM17, we did patch-wise processing (with the centroids of the patches sampled uniformly inside the head region) instead with a patch size of 136x136x82 or 15x15x15 for U-Net or DeepMedic, respectively. For MO17, we extract a 217x217 ROI around the geometrical center of the third molar from the panoramic radiograph at two times lower resolution. The images of PO18 are resampled to a fixed size of 384x288. All modalities are normalized according to its dataset's mean and standard deviation, except for PO18 where ImageNet statistics are used. We further use extensive data augmentation for all datasets, including Gaussian noise, lateral flipping and rigid transformations.

B. Training parameters

The CNNs are pretrained with CE to help final convergence. In preliminary experiments we noticed this can speed-up the experiments and help metric-sensitive loss functions to be more robust. For U-Net and DeepLab, we make use of the Adam optimizer [34] with the initial learning rate set at 10^{-3} (random initialization) or 10^{-4} (ImageNet initialization). For DeepMedic, we do SGD with an initial learning rate 10^{-1} The learning rate decreases by a factor of five when the validation loss on the complete images stops improving and the training stops when this loss increases. The input batch size is 2 for BR18, IS17 and WM17, 4 for MO17 and IS18, 16 for PO18 and 64 for WM17^{DM}.

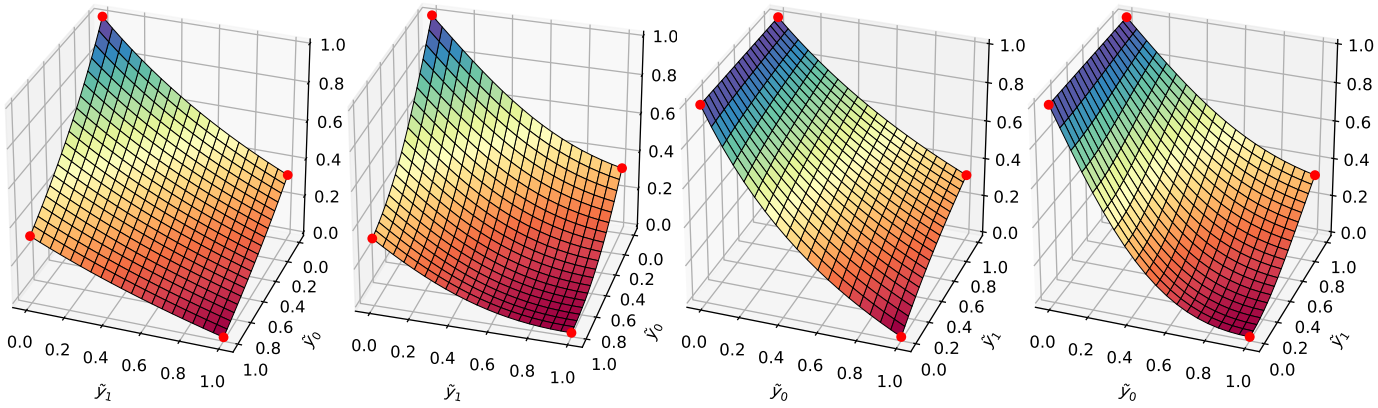
After CE-convergence, the training continues with CE, wCE, sDice, sJaccard, Lovász or Tversky with a reset optimizer state. Our theoretical analysis suggests there is no optimal Dice or Jaccard approximation for wCE that can be derived before training starts (see Sect. II). We therefore set the weights via the common heuristic of balancing out the number of foreground and background pixels/voxels [9], using a weight applied to the foreground class of $1/(2p)$ and a weight applied to the background class of $1/(2-2p)$. Here, p represents the foreground prior. For sTversky, two weighting factors need to be chosen. Our experiments are run for the case $\alpha + \beta = 1.0$ analogously to [16] and on a range of weight values with α ranging from 0.1 to 0.9 with a step size of 0.1. For $\alpha = 0.5$, sTversky is equivalent to sDice. We use similar optimization parameters as for CE pretraining, with the initial learning rate lowered to 10^{-4} for all datasets, except for MO17, which we found appropriate for convergence of the losses under study.

C. Model choice and statistical testing

In all experiments, we perform five-fold cross-validation and report the results aggregated over all the left-out subjects. For each fold we choose the model with the best validation loss (on the full images in case of patch-wise training) to test and report metrics. Final comparisons were done in a pair-wise setting between losses and tests for statistical significance were performed using non-parametric bootstrapping as in the BRATS and ISLES challenges [3], [4]. Implementation details are in [35], with ranks substituted with scores (in line with the hypothesis). To assess inferiority or superiority between pairs of optimization methods, significance level $p < 0.05$ is used. In all tables, cells highlighted in gray point to the top-ranked losses (i.e. not significantly inferior compared to the loss with the highest performance). Values in italic denote inferiority of one loss compared to all others. Boxplots for all measurements are reported separately in Supplementary Materials, Appendix B.

IV. Results and discussion

This section is divided in four parts: First, we compare two common versions of the soft Dice loss present in the literature. Second and most importantly, the performance of all the losses is evaluated on six binary medical segmentation tasks. Thirdly, we also perform this evaluation in a multi-class setting and in the final part, the influence of the class-imbalance on the performance of these respective losses is



(a) $y = [1, 1]$; L^1 soft Dice **(b)** $y = [1, 1]$; L^2 soft Dice **(c)** $y = [1, 0]$; L^1 soft Dice **(d)** $y = [1, 0]$; L^2 soft Dice

Fig. 2: Comparison between L^1 and L^2–based variants of the soft Dice loss for a two-pixel prediction $\tilde{y} = [\tilde{y}_0, \tilde{y}_1]$, and two different ground truth labels y. The red dots indicate the values of the discrete Dice loss. We see that the L^2 variants are flatter around the global minimum corresponding to the ground truth configuration.

investigated. We distinguish two groups of loss functions. The first group are the *CE-based losses*, which contains the (weighted) Hamming loss surrogates, CE and wCE. The second group are the *metric-sensitive losses*. It contains the sDice, sJaccard and Lovász losses, which are surrogates either for the Dice score or Jaccard index. The second group also contains the sTversky losses, which are surrogates for their respective Tversky indices.

A. Soft Dice variants: L^1 vs. L^2

As noted in Sec. II-A, we note that multiple soft generalizations for the Dice loss are possible by replacing $\langle \tilde{y}, \tilde{y} \rangle = \|\tilde{y}\|_2^2$ in the denominator of Eq. (12) by $\|\tilde{y}\|_p^p$ for any L^p norm, and this choice can influence the optimization properties. Both the L^2 [20] and the L^1-based [9] versions are present in the literature. In a preliminary experiment, we implemented the L^1 and L^2 version of the soft Dice loss and compare their performance when evaluating on Dice score and Jaccard index in Table I. There is a small favorable trend of the former definition and it significantly outperforms the latter on the IS18, PO18 and WM17^{DM} dataset. Figure 2 shows a visualization of the loss surfaces for the two soft dice variants, in the case of a simple two-pixel prediction. The fact that the L^2-based variants are flatter around the minimum could indicate that the L^2-based surrogates do not favour integer solutions, i.e. solutions that are close to an indicator vector – which is where the soft surrogate of the Dice metric meets the target discrete Dice metric (red dots on Fig. 2). This would explain why we found that the L^1-based surrogate may outperform the L^2-based surrogate in some cases.

In all further experiments, we restrict our analysis to the L^1-based relaxation, and refer to it simply as the soft Dice loss.

B. Comparison of losses for binary segmentation

The average Dice scores and Jaccard indexes for each dataset and evaluated loss function are shown in table II. The results for sTversky are shown in a different Table III for a better overview. Multiple observations can be made from these results:

TABLE I: Dice scores and Jaccard indexes obtained for each dataset, using the L^1 and L^2 generalization for soft Dice loss. Values highlighted in grey are significantly superior compared to the other loss.

	Loss *Data* →	BR18	IS17	IS18	MO17	PO18	WM17	WM17^{DM}
Dice	L^1	0.870	0.362	0.538	0.944	0.656	0.712	0.717
	L^2	0.869	0.358	*0.525*	0.943	*0.639*	0.708	*0.703*
Jacc.	L^1	0.780	0.244	0.407	0.897	0.598	0.571	0.577
	L^2	0.780	0.245	*0.395*	0.896	*0.567*	0.566	*0.563*

TABLE II: Dice scores and Jaccard indexes obtained for each dataset, using a CE-based loss (CE, wCE) or a metric-sensitive loss (sDice, sJaccard, Lovász). Cells in gray highlight the top-ranked losses. Values in italic point to a significant lower value compared to all other losses.

		group →	CE-based		Metric-sensitive		
	Dataset	*loss* →	CE	wCE	sDice	sJaccard	Lovász
Dice	BR18		0.848	*0.808*	0.870	0.868	0.871
	IS17		0.281	0.310	0.362	0.363	0.357
	IS18		0.463	0.487	0.538	0.528	0.508
	MO18		0.942	*0.903*	0.944	0.942	0.944
	PO18		0.635	*0.602*	0.656	0.651	0.649
	WM17		0.672	*0.492*	0.712	0.712	0.700
	WM17^{DM}		0.671	*0.383*	0.717	0.714	0.711
Jaccard	BR18		0.751	*0.695*	0.780	0.780	0.784
	IS17		0.191	0.212	0.244	0.246	0.245
	IS18		0.345	0.357	0.407	0.399	0.382
	MO18		0.894	*0.835*	0.897	0.896	0.896
	PO18		0.541	*0.488*	0.559	0.554	0.553
	WM17		0.530	*0.355*	0.571	0.573	0.558
	WM17^{DM}		0.529	*0.273*	0.577	0.576	0.572

1) Equivalence of J and D: Theory indicates that Dice score and Jaccard index are equivalent (Prop. II.1) and that they should provide the same ranking of the different loss functions. This also shows empirically in our results. If a loss ℓ_A performs better than loss ℓ_B in terms of Dice, it will perform better in terms of Jaccard index, and vice-versa.

2) Performance of the surrogates: The results show clearly that use of the metric-sensitive losses provides superior Dice scores and Jaccard indexes than (w)CE. There is only one dataset, MO17, for which the performance of CE is not inferior to that of the metric-sensitive losses. This is probably because the class imbalance is less pronounced in this dataset. These results are what we expected since we already saw that CE and the metric-sensitive losses are theoretically divergent. A second observation that we make is that there is no significant difference between sDice, sJacard and Lovász as a loss function. This implies a free choice within the group of metric-sensitive losses.

For the Tversky loss, we observe that in no situation, $\alpha = 0.5$-weighting is significantly inferior to any of the unequal weightings. With α going further away from the 0.5 baseline, the performance w.r.t. Dice score drops significantly once the weighting is too extreme. Therefore we can conclude that in our experiments $\alpha = \beta = 0.5$ (equivalent to Dice loss) is the optimal weighting for all cases when evaluating on Dice score.

3) Weighting of cross-entropy: We can see from our results that wCE is typically underperforming w.r.t. CE. Choosing a different weight might increase the performance of wCE, but it is apparent that the choice of weighting is non-trivial and very application-dependent. Finding an optimal weight is therefore not easy and will probably require additional parameter tuning as compared to the metric-sensitive losses. Moreover, we show in section IV-C that performance of wCE is better for a smaller range of object sizes. Therefore, it is highly unlikely that one specific weighting could provide an appropriate surrogate for all the object sizes and datasets which is in agreement with our theory as well.

4) Qualitative inspection: In Fig. 3 we present example slices for each dataset. From left to right the images represent the input (with ground truth delineations in blue), the ground truth segmentation and the predictions using each of the original loss functions (with ground truth delineations in blue and predicted delineations in red). The slices shown had a median Dice, both across cases and within one case (having a minimum of one foreground pixel), for the CE model and were checked for being representative. We note a wider value range for the outputs of the cross-entropy based losses. In particular, at a threshold of 0.5, the CE model seems to systematically under-segment the structures. We further note that the outputs of the metric-sensitive losses are highly similar and have a narrow value range close to 1.0. Only for both ISLES datasets, the Lovász predictions have a value range close to 0.5. However, after thresholding at 0.5 the predicted delineations are similar to sDice and sJaccard.

5) Network architecture and patch-wise training: It is clear that results are consistent over different state-of-the-art network architectures: U-Net (used in BR18, IS17, IS18, MO17 and WM17), DeepLab (used in PO18) and DeepMedic (used in WM17^{DM}). All architectures were shown to be optimized according to the implemented loss functions. This was somehow expected since these architectures are used almost interchangeably in recent biomedical segmentation challenges [4], [3].

Interestingly, the results are also consistent for patch-wise training, whereas it is unclear how the effective, patch-wise training loss relates to the actual test metric that is evaluated on full images.

6) Generalizability to absolute and distance-based metrics: The results above show that the use of metric-sensitive losses provides superior Dice scores and Jaccard indices compared to (w)CE. However, often multiple metrics are of interest for evaluation such as volume errors or distance-based metrics [3], [30]. In this light, we report three additional metrics for each of the public datasets: accuracy (ACC), Hausdorff distance (HDD) and absolute volume difference (AVD). The performance on these metrics is compared for the experiments with CE and sDice in table IV (equivalence within each group of loss functions was already shown). For HDD and AVD, sDice is never inferior to CE and in certain cases significantly better. For ACC, CE is often superior, which is to be expected since it directly optimizes the evaluation metric.

C. Influence of class imbalance

Most datasets for image segmentation consist of objects of different sizes. The metric-sensitive losses are typically expected to help most with segmentation of small-sized objects because they are invariant to scale [12]. The approximation bound for the Hamming loss in Eq. (5) is dependent on the number of foreground pixels in the ground truth and this makes us think that optimization with cross-entropy will lead to a lower Dice score for small objects.

We study the performance of the different loss functions w.r.t. the object size in Fig. 4. This figure shows the average Dice scores in function of the number of foreground pixels in the ground truth. We show that optimization with one of the metric-sensitive losses is beneficial for all object sizes and that CE can provide poor Dice scores for almost the entire range of scales. CE is underperforming the other losses for all datasets, but especially in BR18 and IS18. We further note that the performance of wCE as a loss function can greatly depend on the lesion size (which is not the case for CE). In fact, wCE does improve performance for some sample-size ranges, but never across the entire range and it can also perform much worse than the metric-sensitive losses. This shows again that simply giving a weight to cross-entropy will not be able to surrogate the target metric across all sample scales.

In order to further investigate the influence of the foreground to background ratio (fg/bg ratio) on the performance of the previously described loss functions at the dataset level, we virtually create new datasets from PO18. Following Sect. III-A, the apparent fg/bg ratios of the datasets are 0.00518, 0.0264, 0.0348, 0.0424, 0.0671 and 0.0705 for WM17, IS17, IS18, PO18, BR18 and MO17, respectively. With the lower end of the fg/bg spectrum covered by WM17 and the converging trend of the losses at the case level (i.e. object size) observed in Fig. 4, we question the hypothesis for higher fg/bg ratios. For each of the ratios we want to analyse (i.e. 0.05, 0.1, 0.2,

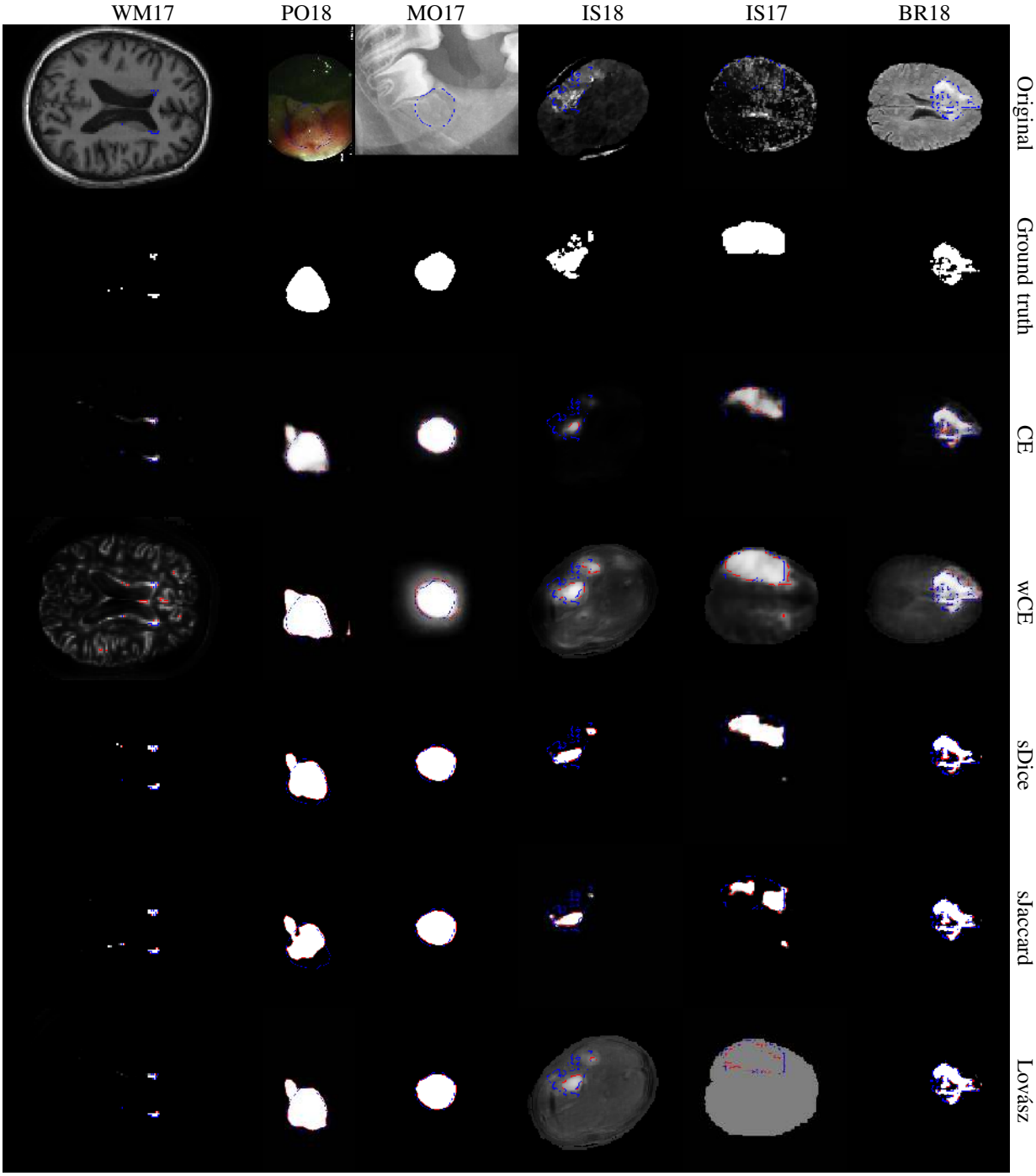
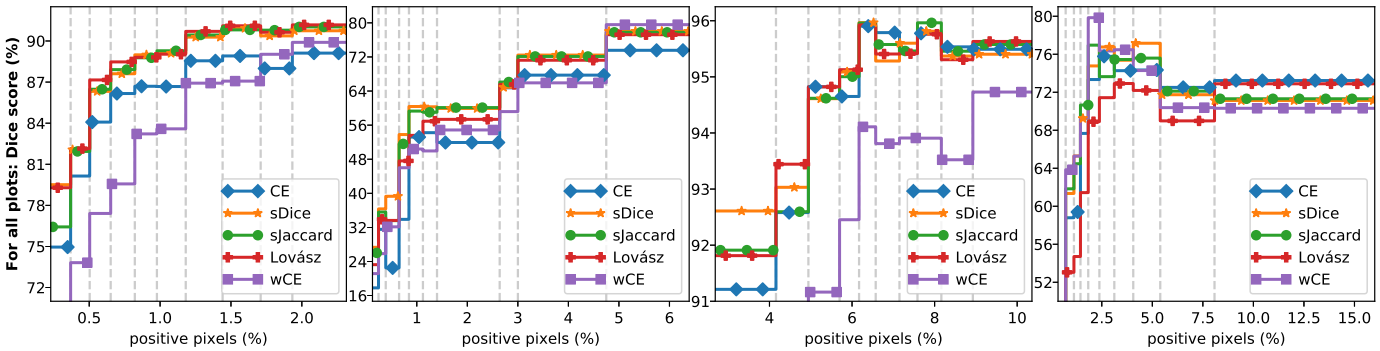


Fig. 3: A qualitative evaluation of the resulting predictions from networks trained with CE, wCE, sDice, sJaccard and Lovász, columns 3-7, respectively. The predicted delineations are obtained after thresholding at 0.5 and are overlayed in red. The original images inputted to the network (or part of) are shown on the left. The ground truths are shown in the second column and their delineations are overlayed in blue on the predictions. The slices shown had a median Dice, both across cases and within one case (having a minimum of one foreground pixel), for the CE model and were checked visually for being representative.

TABLE III: Dice scores and Jaccard indexes obtained for all datasets using the range of sTversky losses with varying alpha/beta. Cells in gray highlight the top-ranked losses. There is no weighting scheme producing significantly higher values than sTversky 0.5/0.5.

	Dataset $\alpha/\beta \rightarrow$	0.1/0.9	0.2/0.8	0.3/0.7	0.4/0.6	0.5/0.5	0.6/0.4	0.7/0.3	0.8/0.2	0.9/0.1	0.75/0.75	1.0/1.0
Dice	BR18	0.801	0.844	0.859	0.863	0.870	0.867	0.857	0.831	*0.787*	0.871	0.868
	IS17	0.352	0.345	0.374	0.374	0.362	0.346	0.356	0.345	0.293	0.371	0.363
	IS18	0.481	0.522	0.533	0.540	0.538	0.527	0.519	0.490	*0.445*	0.528	0.528
	MO18	0.907	0.928	0.936	0.941	0.944	0.942	0.938	0.932	0.902	0.944	0.942
	PO18	0.615	0.628	0.638	0.647	0.656	0.651	0.647	0.633	0.611	0.646	0.651
	WM17	0.581	0.649	0.689	0.700	0.712	0.706	0.693	0.660	0.565	0.712	0.712
	$WM17^{DM}$	0.599	0.668	0.696	0.709	0.717	0.706	0.675	0.628	*0.496*	0.706	0.714
Jaccard	BR18	0.680	0.740	0.765	0.772	0.780	0.775	0.764	0.726	*0.666*	0.783	0.780
	IS17	0.239	0.232	0.257	0.257	0.244	0.231	0.238	0.230	*0.188*	0.250	0.246
	IS18	0.349	0.389	0.401	0.407	0.407	0.398	0.390	0.361	*0.321*	0.399	0.399
	MO18	0.834	0.870	0.884	0.892	0.897	0.895	0.886	0.876	0.826	0.896	0.896
	PO18	0.504	0.523	0.536	0.548	0.559	0.556	0.554	0.539	0.511	0.550	0.554
	WM17	0.422	0.496	0.541	0.556	0.571	0.565	0.548	0.510	0.410	0.571	0.573
	$WM17^{DM}$	0.441	0.516	0.551	0.566	0.577	0.566	0.532	0.481	*0.360*	0.565	0.576



(a) BR18 **(b)** IS18 **(c)** MO17 **(d)** PO18

Fig. 4: Dice score as a function of object size (# foreground pixels in %) for BR18 (a), IS18 (b), MO17 (c) and PO18 (d). The results for IS17 and WM17 are omitted due to the small dataset size and lack of statistical relevance. For visualization purposes, the Dice scores are averaged within every 10th percentile (bordered by the dashed lines). The superior performance for metric-sensitive losses holds across the entire range of object sizes.

TABLE IV: Accuracy (ACC), Hausdorff distance (HDD) and absolute volume difference (AVD) for each of the public datasets and for networks trained with CE and sDice. Values highlighted in grey point to a significantly better value compared to the other loss. Italic formatting denotes significantly inferior values.

	Loss	*Dataset* $\rightarrow$ BR18	IS17	IS18	WM17	$WM17^{DM}$
ACC	CE	*0.997*	0.995	0.989	0.999	0.999
	sDice	0.998	*0.993*	*0.988*	0.999	0.999
HDD	CE	22.673	*73.102*	*42.994*	34.210	30.526
	sDice	22.911	47.043	33.366	34.806	28.695
AVD	CE	*15.579*	21.042	12.029	*5.628*	*5.865*
	sDice	13.375	35.895	11.954	3.895	4.064

0.3, 0.4, 0.5), a new dataset is created as follows: a binary rectangular mask is created for each image as shown in Fig. 5. This rectangle has the same aspect ratio as the original image and contains the entire polyp (or part of it if it doesn't fit). The size of this rectangle is the same for all images in each virtual dataset and is determined such that the average fg/bg ratio over all images is equal to the desired ratio. This binary mask is then used to mask the output of the final layer of the network during training and testing. This way, only the pixels inside the rectangle are used to update the weights of the network, but the field-of-view (FOV) doesn't change between the virtual datasets. Otherwise, if we would just crop the input images before feeding them to the network, segmentation could become harder when increasing the fg/bg ratio because the FOV decreases.

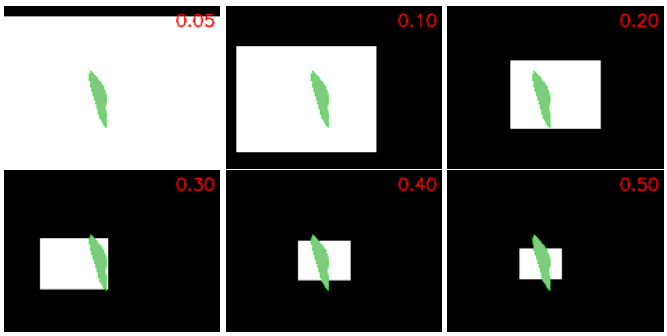


Fig. 5: For each foreground to background ratio in red, a rectangular mask is created with the same aspect ratio as the original image and which contains the entire polyp (or part of it if not possible) in green.

We then analysed the loss functions' performance for each of the six virtual datasets analogously to the original PO18 dataset in Section III. The same network, training procedure and training images from PO18 are used with one difference: the output of the final layer is masked with the corresponding mask for a given fg/bg ratio. The results are shown in Table V. We only post-trained with CE and sDice since we already

showed that there is no significant difference within each group of loss functions. We see no clear trend when going from 0.05 to 0.5, indicating that the superiority of the metric-sensitive loss functions is not dependent on the fg/bg ratio. Even for a ratio of 0.5, post-training with Dice loss shows a significantly higher performance w.r.t. Dice score than CE.

TABLE V: Dice scores and Jaccard indexes obtained for each virtual PO18 dataset and for networks trained with CE or sDice. Values highlighted in grey point to a significantly higher value compared to the other loss. Italic formatting denotes significantly inferior values.

	Loss	*Dataset* →	0.05	0.1	0.2	0.3	0.4	0.5
Dice	CE		*0.627*	*0.645*	*0.673*	*0.691*	*0.730*	*0.757*
	sDice		0.639	0.662	0.683	0.701	0.740	0.765
Jacc.	CE		*0.534*	*0.552*	*0.579*	*0.597*	*0.639*	*0.672*
	sDice		0.541	0.564	0.584	0.604	0.646	0.676

D. Comparison of losses for multi-class segmentation

The results for the original loss functions are presented in Table VI. Again, we observe a consistent, superior performance of the metric-sensitive losses compared to the cross-entropy losses w.r.t. the Dice score and Jaccard index. Given foreground-background ratios of 0.065, 0.028 and 0.012 for WT, TC and ET, respectively, we observe a similar trend as in Fig. 4. The metric-sensitive losses obtain higher Dice scores and Jaccard indexes compared to the cross-entropy losses over multiple foreground-background ratios.

TABLE VI: Dice scores and Jaccard indexes obtained for the multi-class BRATS dataset using the original loss functions. Results are presented for the three volumes that were considered for evaluation during the challenge: whole tumor (WT), tumor core (TC) and enhancing tumor (ET). We also report the average across the three types (all). Values in gray are the significantly best performing losses. There are no values in italic since no single loss performed significantly worse than all the others.

		group →	CE-based		Metric-sensitive		
	Type	*loss* →	CE	wCE	sDice	sJaccard	Lovász
Dice	All		0.639	0.665	0.740	0.747	0.716
	WT		0.798	0.799	0.849	0.859	0.858
	TC		0.620	0.636	0.725	0.737	0.692
	ET		0.498	0.560	0.646	0.645	0.598
Jaccard	All		0.522	0.554	0.632	0.643	0.606
	WT		0.683	0.688	0.753	0.767	0.766
	TC		0.487	0.518	0.609	0.627	0.569
	ET		0.395	0.456	0.534	0.534	0.484

The results for the range of sTversky losses are presented in Table VII.

Similar to the results regarding the range of sTversky losses in Sect. IV, there is no weighting scheme that is significantly superior to the Dice equivalent weighting of $\alpha = \beta = 0.5$. There is a clear trend of sub-optimal weighting when α and/or β deviate from 0.5, which relates back to Fig. 1 where relative and absolute errors are bound to increase.

V. Conclusion

More and more metric-sensitive loss functions find their way into the optimization of CNNs for image segmentation, both in the context of medical imaging and classical computer vision. Nonetheless, we saw a great mass of research in the MICCAI 2018 proceedings still using per-pixel losses while their evaluation was based primarily on the Dice score or Jaccard index. In this work we questioned the latter approach from both theoretical and empirical point of views.

On the one hand, theory suggests that the Dice score and Jaccard index approximate each other relatively and absolutely. On the other hand, we found that no such approximations exist for a weighted Hamming similarity. For the Tversky index, the approximation gets monotonically worse when weighting differently from the soft Dice setting. We were able to confirm these findings in an extensive empirical validation on five binary medical segmentation tasks. Further experiments reaffirmed their superior performance across different object sizes, foreground/background ratios and in a multi-class setting.

We conclude that segmentation tasks may benefit from a wider adoption of metric-sensitive loss-functions when evaluation is performed using only the Dice score or the Jaccard index.

Appendix A
MICCAI 2018 proceedings

In the introduction, it is stated that only 30% of all MICCAI 2018 proceedings actually use a metric-sensitive loss function. This, even though the Dice score is reported as their evaluation metric. We categorized all MICCAI 2018 proceedings [36] as indicated by the flowchart in Figure 6. From the total number of MICCAI 2018 proceedings (n = 372), there are 103 papers dealing with learning-based segmentation. Almost all of them (n=96) use the Dice score or Jaccard index to evaluate the performance of their algorithm, but only 28 proceedings report to use soft Dice or any of the other metric-sensitive loss functions described in this work. Many methods use cross-entropy or its weighted variant (n=38) and the rest have used a variety of other losses such as adverserial loss, focal loss, mean-squared-error, etc.

TABLE VII: Dice scores and Jaccard indexes obtained for the multi-class BRATS dataset using the range of sTversky losses with varying alpha/beta. Results are presented for the three volumes that were considered for evaluation during the challenge: whole tumor (WT), tumor core (TC) and enhancing tumor (ET). We also report the average across the three types (all). Cells in gray highlight the top-ranked losses and italic values are significantly inferior to all others in the row. There is no weighting scheme producing significantly higher values than sTversky 0.5/0.5.

	Type $\alpha/\beta \rightarrow$	0.1/0.9	0.2/0.8	0.3/0.7	0.4/0.6	0.5/0.5	0.6/0.4	0.7/0.3	0.8/0.2	0.9/0.1	0.75/0.75	1.0/1.0
Dice	All	0.649	0.702	0.725	0.739	0.740	0.738	0.709	0.704	0.649	0.741	0.747
	WT	0.771	0.822	0.844	0.853	0.849	0.848	0.826	0.816	0.757	0.858	0.859
	TC	0.652	0.703	0.724	0.735	0.725	0.725	0.687	0.686	0.623	0.731	0.737
	ET	*0.525*	0.582	0.606	0.631	0.646	0.642	0.614	0.611	0.568	0.634	0.645
Jaccard	All	0.517	0.580	0.610	0.630	0.632	0.629	0.595	0.586	0.521	0.634	0.643
	WT	0.640	0.711	0.742	0.757	0.753	0.750	0.719	0.704	0.627	0.765	0.767
	TC	0.513	0.572	0.602	0.617	0.609	0.609	0.567	0.563	0.493	0.615	0.627
	ET	0.397	0.457	0.487	0.515	0.534	0.528	0.500	0.491	0.444	0.522	0.534

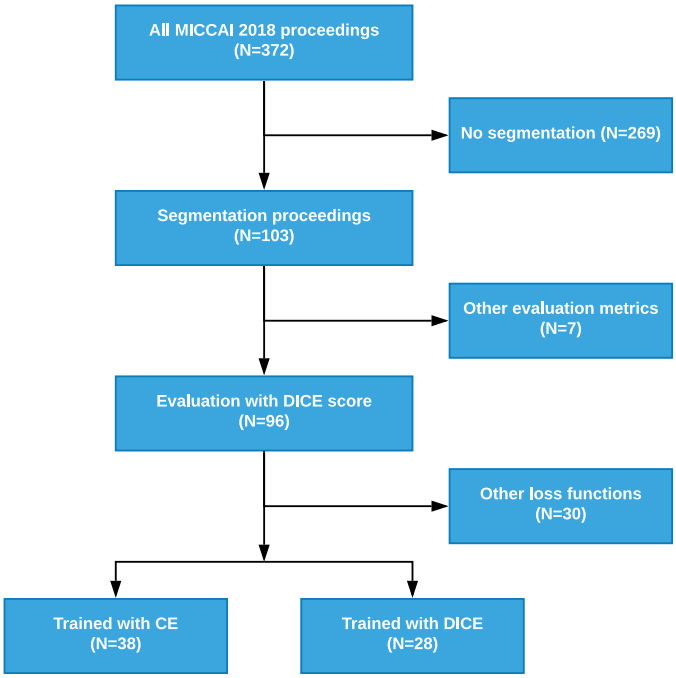


Fig. 6: Flowchart used for categorization of MICCAI 2018 proceedings. All segmentation papers are selected and from these, we retain all the papers that used DICE score for evaluation. We then analysed which type of loss function was used.

REFERENCES

[1] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841516301839

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *MICCAI*, pp. 234–241, 2015.

[3] "BRATS challenge," 2018. [Online]. Available: https://www.med.upenn.edu/sbia/brats2018.html

[4] "ISLES challenge," 2017. [Online]. Available: http://www.isles-challenge.org/ISLES2017/

[5] "ISLES challenge," 2018. [Online]. Available: http://www.isles-challenge.org/ISLES2017/

[6] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, "Morphometric analysis of white matter lesions in MR images: Method and validation," *IEEE Trans. Medical Imaging*, vol. 13, no. 4, pp. 716–724, 1994.

[7] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[8] L. Chen, P. Bentley, and D. Rueckert, "Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks," *NeuroImage: Clinical*, vol. 15, no. May, pp. 633–643, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.nicl.2017.06.016

[9] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations," *LNCS*, vol. 10553, pp. 240–248, 2017.

[10] D. Tarlow and R. P. Adams, "Revisiting uncertainty in graph cut solutions," in *CVPR*, 2012, pp. 2440–2447.

[11] S. Nowozin, "Optimal decisions from probabilistic models: the intersection-over-union case," in *CVPR*, 2014, pp. 548–555.

[12] M. Berman, A. Rannen Triki, and M. B. Blaschko, "The Lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *CVPR*, 2018, pp. 4413–4421.

[13] J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. Blaschko, "Optimizing the Dice score and Jaccard index for medical image segmentation: Theory and practice," in *Medical Image Computing and Computer-Assisted Intervention*, 2019.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[15] J. R. England and P. M. Cheng, "Artificial intelligence for medical image analysis: A guide for authors and reviewers," *AJR Am J Roentgenol*, vol. 212, no. 3, pp. 513–519, 2019.

[16] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," in *Machine Learning in Medical Imaging*, Q. Wang, Y. Shi, H.-I. Suk, and K. Suzuki, Eds. Springer, 2017, pp. 379–387.

[17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Neurocomputing: Foundations of research," J. A. Anderson and E. Rosenfeld, Eds. MIT Press, 1988, ch. Learning Representations by Back-propagating Errors, pp. 696–699.

[18] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006.

[19] C. X. Ling and V. S. Sheng, "Cost-sensitive learning," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Springer, 2010, pp. 231–235.

[20] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Fourth International Conference on 3D Vision*, 2016, pp. 565–571.

[21] M. A. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in *Advances in Visual Computing*, G. Bebis, R. Boyle, B. Parvin, D. Koracin, F. Porikli, S. Skaff, A. Entezari, J. Min, D. Iwai, A. Sadagic, C. Scheidegger, and T. Isenberg, Eds. Springer, 2016, pp. 234–244.

[22] J. Yu and M. B. Blaschko, "The Lovász hinge: A novel convex surrogate for submodular losses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 735–748, 2020.

[23] J. Yu and M. Blaschko, "A convex surrogate operator for general non-modular loss functions," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Gretton and C. C. Robert, Eds., vol. 51, 2016, pp. 1032–1041.

[24] L. Lovász, *Submodular functions and convexity*. Springer, 1983, pp. 235–257. [Online]. Available: https://doi.org/10.1007/978-3-642-68874-4_10

[25] B. H. Menze, A. Jakab, S. Bauer *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, 2015.

[26] S. Bakas, H. Akbari, A. Sotiras *et al.*, "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, no. March, pp. 1–13, 2017.

[27] S. Bakas, M. Reyes, A. Jakab *et al.*, "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge," 2018. [Online]. Available: http://arxiv.org/abs/1811.02629

[28] J. De Tobel, P. Radesh, D. Vandermeulen, and P. W. Thevissen, "An automated technique to stage lower third molar development on panoramic radiographs for age estimation: A pilot study," *Journal of Forensic Odonto-Stomatology*, vol. 35, no. 2, pp. 49–60, 2017.

[29] T. Eelbode, I. Demedts, R. Bisschops, P. Roelandt, C. Hassan, E. Coron, P. Bhandari, H. Neumann, O. Pech, A. Repici *et al.*, "Incorporation of temporal information in a deep neural network improves performance level for automated polyp detection and delineation." *Gastrointestinal Endoscopy*, vol. 89, no. 6, pp. AB618–AB619, 2019.

[30] H. J. Kuijf, A. Casamitjana, D. L. Collins, M. Dadar, A. Georgiou, M. Ghafoorian, D. Jin, A. Khademi, J. Knight, H. Li, X. Lladó, J. M. Biesbroek, M. Luna, Q. Mahmood, R. Mckinley, A. Mehrtash, S. Ourselin, B. Y. Park, H. Park, S. H. Park, S. Pezold, E. Puybareau, J. De Bresser, L. Rittner, C. H. Sudre, S. Valverde, V. Vilaplana, R. Wiest, Y. Xu, Z. Xu, G. Zeng, J. Zhang, G. Zheng, R. Heinen, C. Chen, W. Van Der Flier, F. Barkhof, M. A. Viergever, G. J. Biessels, S. Andermatt, M. Bento, M. Berseth, M. Belyaev, and M. J. Cardoso, "Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge," *IEEE Transactions on Medical Imaging*, vol. 38, no. 11, pp. 2556–2568, 2019.

[31] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No New-Net," *arXiv preprint arXiv:1809.10483*, 2018, arXiv:1809.10483. [Online]. Available: http://arxiv.org/abs/1809.10483

[32] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE-TPAMI*, vol. 40, no. 4, 2018. [Online]. Available: https://doi.org/10.1109/TPAMI.2017.2699184

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[36] A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings*. Springer, 2018, vol. 11073.

Supplementary material

Appendix A
F measures

We show in Section IV that the weighting $\alpha = \beta = 0.5$ for the Tversky loss (equivalent to Dice loss) is the optimal weighting when evaluation is performed on Dice score. Alternative weightings of Tversky loss can however be beneficial in case other metrics are of interest. In Table B.1, we show that a different weighting scheme can be very effective to achieve superior performance when evaluation is done with alternative F-measures ($F^{0.5}$, $F^{1.0}$, $F^{1.5}$ and $F^{2.0}$). A relative higher weight for α gives better performance for the lower order F-measures, whereas the best weighting for $F^{2.0}$ is found in the lower range of values for α. In this table, results for the weightings $\alpha = \beta = 0.5$ and $\alpha = \beta = 1.0$ are also provided which correspond to sDice and sJaccard respectively as well as the balanced version between both, i.e. weighting $\alpha = \beta = 1.0$. No clear difference can be observed between these three.

Appendix B
Boxplots

This section gives a complementary boxplot for all tables in the main manuscript providing more detailed spread information for all measurements. It can be observed that the IS17, IS18 and PO18 datasets have the highest variability which is to be expected due to their difficult nature and smaller sample size.

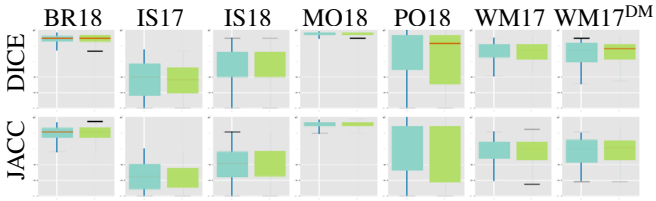


Fig. B.1: Boxplots for the Dice scores and Jaccard indexes obtained for each dataset, using the L^1 and L^2 generalization for soft Dice loss.

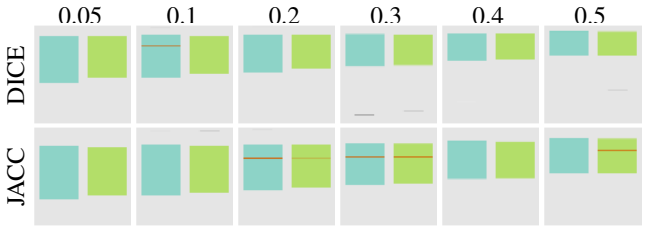


Fig. B.2: Boxplots for the Dice scores and Jaccard indexes obtained for each virtual PO18 dataset and for networks trained with CE or sDice.

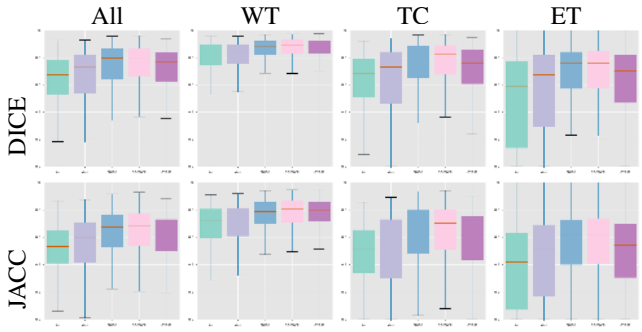


Fig. B.3: Boxplots for the Dice scores and Jaccard indexes obtained for the multi-class BRATS dataset using the original loss functions. Results are presented for the three volumes that were considered for evaluation during the challenge: whole tumor (WT), tumor core (TC) and enhancing tumor (ET). We also report the average across the three types (all).

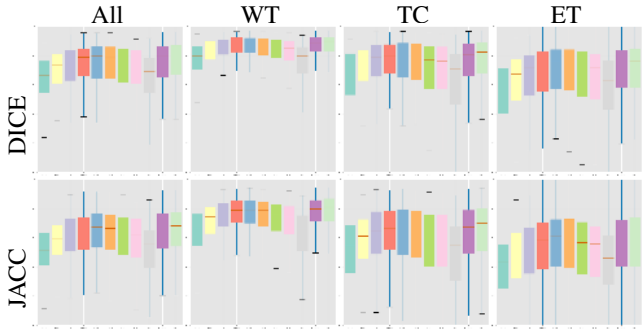


Fig. B.4: Boxplots for the Dice scores and Jaccard indexes obtained for the multi-class BRATS dataset using the range of sTversky losses with varying alpha/beta. Results are presented for the three volumes that were considered for evaluation during the challenge: whole tumor (WT), tumor core (TC) and enhancing tumor (ET). We also report the average across the three types (all).

Fig. B.5: Boxplots for the accuracy, Hausdorff distance and absolute volume difference obtained for each dataset, using a cross-entropy loss (CE, wCE) or a metric-sensitive loss (sDice, sAVD, Lovász).

TABLE B.1: Alternative F measures obtained for each dataset using the range of sTversky losses with varying alpha/beta. Cells in grey highlight the top-ranked losses and values that are significantly inferior to all others are shown in italic.

	Measureα/β →	0.1/0.9	0.2/0.8	0.3/0.7	0.4/0.6	0.5/0.5	0.6/0.4	0.7/0.3	0.8/0.2	0.9/0.1	0.75/0.75	1.0/1.0
BR18	F0.5	*0.739*	0.806	0.841	0.853	0.878	0.890	0.896	0.894	0.879	0.878	0.877
	F1.0	0.801	0.844	0.859	0.863	0.870	0.867	0.857	0.831	*0.787*	0.871	0.868
	F1.5	0.850	0.874	0.875	0.873	0.867	0.855	0.837	0.798	*0.741*	0.869	0.865
	F2.0	0.882	0.893	0.885	0.880	0.867	0.850	0.827	0.782	*0.719*	0.869	0.865
IS17	F0.5	0.316	0.318	0.358	0.372	0.368	0.367	0.387	0.382	0.363	0.367	0.362
	F1.0	0.352	0.345	0.374	0.374	0.362	0.346	0.356	0.345	0.293	0.371	0.363
	F1.5	0.398	0.384	0.402	0.393	0.378	0.351	0.360	0.342	*0.274*	0.394	0.384
	F2.0	0.439	0.420	0.430	0.415	0.396	0.362	0.372	0.348	*0.270*	0.419	0.408
IS18	F0.5	*0.411*	0.468	0.495	0.516	0.528	0.528	0.539	0.543	0.530	0.521	0.520
	F1.0	0.481	0.522	0.533	0.540	0.538	0.527	0.519	0.490	*0.445*	0.528	0.528
	F1.5	0.551	0.573	0.571	0.568	0.555	0.537	0.517	0.470	*0.413*	0.544	0.544
	F2.0	0.607	0.612	0.600	0.590	0.569	0.547	0.520	0.464	*0.399*	0.558	0.558
MO17	F0.5	*0.866*	0.900	0.918	0.930	0.943	0.948	0.951	0.954	0.948	0.943	0.942
	F1.0	0.907	0.928	0.936	0.941	0.944	0.942	0.938	0.932	0.902	0.944	0.942
	F1.5	0.936	0.947	0.949	0.948	0.945	0.940	0.930	0.919	*0.876*	0.945	0.944
	F2.0	0.954	0.959	0.957	0.953	0.946	0.939	0.926	0.913	*0.862*	0.946	0.945
PO18	F0.5	*0.588*	0.609	0.625	0.644	0.664	0.664	0.671	0.669	0.664	0.656	0.657
	F1.0	0.614	0.628	0.637	0.647	0.656	0.650	0.647	0.633	0.611	0.646	0.651
	F1.5	0.643	0.651	0.654	0.658	0.660	0.651	0.641	0.620	*0.591*	0.650	0.656
	F2.0	0.665	0.669	0.668	0.668	0.665	0.655	0.640	0.616	*0.582*	0.655	0.662
WM17	F0.5	*0.498*	0.592	0.659	0.690	0.724	0.742	0.761	0.759	0.720	0.723	0.730
	F1.0	0.581	0.649	0.689	0.700	0.712	0.706	0.693	0.660	0.565	0.712	0.712
	F1.5	0.657	0.700	0.718	0.713	0.711	0.692	0.660	0.615	*0.501*	0.712	0.708
	F2.0	0.714	0.736	0.739	0.724	0.713	0.686	0.645	0.593	*0.472*	0.714	0.708
$WM17^{DM}$	F0.5	*0.521*	0.617	0.671	0.709	0.736	0.758	0.759	0.752	0.647	0.731	0.736
	F1.0	0.599	0.668	0.696	0.709	0.717	0.706	0.675	0.628	*0.496*	0.706	0.714
	F1.5	0.670	0.713	0.722	0.717	0.712	0.682	0.636	0.574	*0.436*	0.697	0.707
	F2.0	0.721	0.744	0.739	0.723	0.712	0.672	0.618	0.548	*0.409*	0.695	0.706

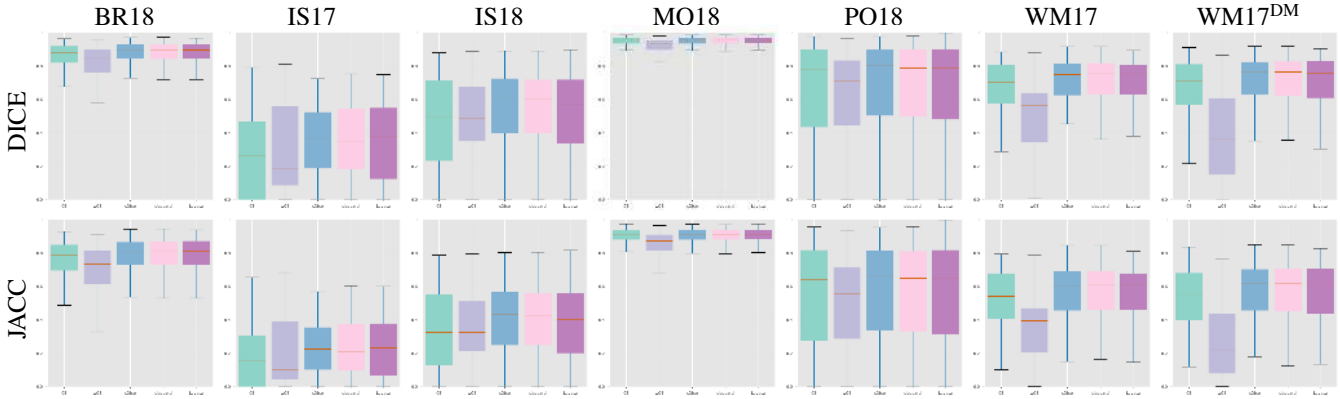


Fig. B.6: Boxplots for the Dice scores and Jaccard indexes obtained for each dataset, using a cross-entropy loss (CE, wCE) or a metric-sensitive loss (sDice, sJaccard, Lovász).

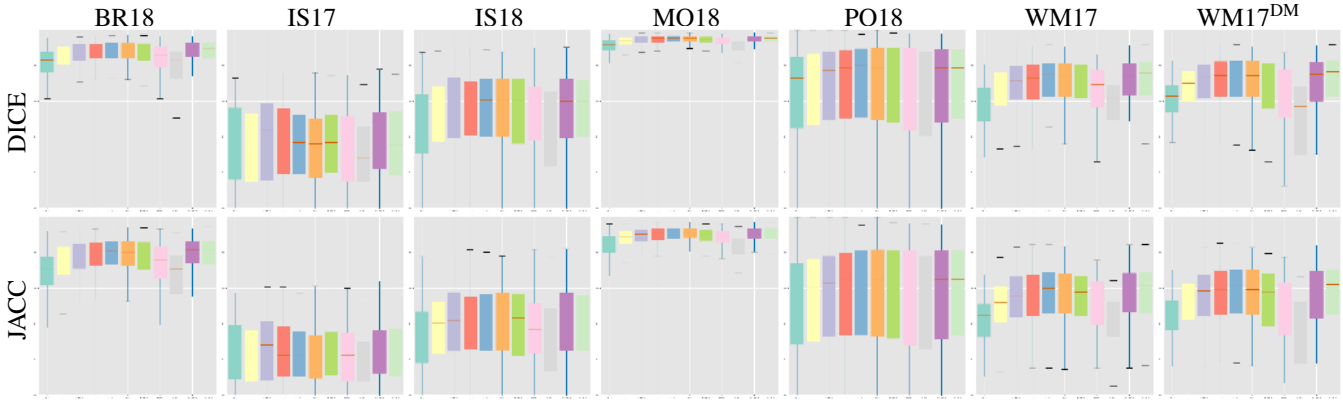


Fig. B.7: Boxplots for the Dice scores and Jaccard indexes obtained for all datasets using the range of sTversky losses with varying alpha/beta.

Fig. B.8: Boxplots of the alternative F measures obtained for each dataset using the range of sTversky losses with varying alpha/beta.