# Calibrated Surrogate Maximization of Dice

Marcus Nordström[1,2(✉)], Han Bao[3,4], Fredrik Löfman[2], Henrik Hult[1], Atsuto Maki[1], and Masashi Sugiyama[4,3]

[1] KTH Royal Institute of Technology, Stockholm, Sweden
marcno@kth.se
[2] RaySearch Laboratories, Stockholm, Sweden
[3] The University of Tokyo, Tokyo, Japan
[4] RIKEN, Tokyo, Japan

**Abstract.** In the medical imaging community, it is increasingly popular to train machine learning models for segmentation problems with objectives based on the soft-Dice surrogate. While experimental studies have showed good performance with respect to Dice, there have also been reports of some issues related to stability. In parallel with these developments, direct optimization of evaluation metrics has also been studied in the context of binary classification. Recently, in this setting, a quasi-concave, lower-bounded and calibrated surrogate for the $F_1$-score has been proposed. In this work, we show how to use this surrogate in the context of segmentation. We then show that it has some better theoretical properties than soft-Dice. Finally, we experimentally compare the new surrogate with soft-Dice on a 3D-segmentation problem and get results indicating that stability is improved. We conclude that the new surrogate, for theoretical and experimental reasons, can be considered a promising alternative to the soft-Dice surrogate.

**Keywords:** Dice · Calibration · Segmentation

## 1 Introduction

With the introduction of the U-net [19], it became common in the medical imaging community to train neural network models evaluated using Dice with the cross-entropy loss [5,7,13]. However, because of problems associated with handling small structures, it was later proposed that a loss based on a smoothed version of Dice, referred to as soft-Dice, would yield better predictions [18]. This was confirmed in several studies [3,6,8,21], but because of some reported problems associated with handling noisy data [4,8,16] and an increased risk of convergence issues [8], it is common for practitioners to sacrifice some performance

**Electronic supplementary material** The online version of this chapter (https://doi.org/10.1007/978-3-030-59719-1_27) contains supplementary material, which is available to authorized users.


A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12264, pp. 269–278, 2020.
https://doi.org/10.1007/978-3-030-59719-1_27

for some stability by using the sum of the soft-Dice loss and the cross-entropy loss as the objective during training [8,9,24].

In parallel with these developments, several alternative performance metrics to accuracy, i.e., the probability of correct predictions, have been investigated to tackle the class imbalance problem in binary classification [11,14,15,17]. Since these metrics do not in general reduce to a sum of per-sample scores, one cannot in general consider the classical procedure of maximizing a concave approximation of the score without loosing statistical consistency [22]. However, for the case of the fractional utility metrics, recent work showed that one can consider such a procedure without loosing statistical consistency, provided concavity is replaced with the weaker notion of quasi-concavity [1].

In this work we address the stability issues reported for soft-Dice by making use of the recent progress in binary classification. More specifically, we propose a new surrogate that is both quasi-concave and a calibrated lower bound to Dice. We then prove that soft-Dice is neither quasi-concave nor a lower bound to Dice. Finally, we compare the surrogates experimentally on a kidney segmentation problem and report some evidence for improvement on the stability issues reported for soft-Dice.

## 2 Surrogate Maximization

Given a pair of $\mathbb{P}$-measurable random variables $(X, Y)$ taking values in $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^D \times \{\pm 1\}$ and a set of real valued functions $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$, the problem of binary classification is to find an $f \in \mathcal{F}$ such that $\operatorname{sgn}(f(X))$ predicts $Y$ as *good as possible* with respect to some score $S$:

$$f_* = \operatorname*{arg\,max}_{f \in \mathcal{F}} S(f, X, Y). \tag{1}$$

However, due to the discrete nature of the score functions often used, direct optimization is typically not feasible. Consequently, to simplify the problem, it is common to approximate $S$ with a concave surrogate $\tilde{S}$. In the sequel, we will also refer to the less common situation when concavity is replaced with quasi-concavity.

**Definition 1.** *Let $\tilde{S}$ be some surrogate score.*

1. *$\tilde{S}$ is said to be concave if*

$$\tilde{S}(\alpha f_1 + (1-\alpha) f_2, X, Y) \geq \alpha \tilde{S}(f_1, X, Y) + (1-\alpha)\tilde{S}(f_2, X, Y), \tag{2}$$

*for any measurable functions, $f_1, f_2$, random variables $X, Y$ and $\alpha \in [0, 1]$.*

2. *$\tilde{S}$ is said to be quasi-concave if*

$$\tilde{S}(\alpha f_1 + (1-\alpha) f_2, X, Y) \geq \min\{\tilde{S}(f_1, X, Y), \tilde{S}(f_2, X, Y)\}, \tag{3}$$

*for any measurable functions $f_1, f_2$, random variables $X, Y$ and $\alpha \in [0, 1]$.*

To ensure that a surrogate is well behaved, it needs to relate to the score in some ways. For this purpose, it is common to consider the *calibration property*, which ensures that any solution to the surrogate maximization problem is also a solution to the original score maximization problem [2,20,23]. Another property that can be considered is the lower-bound property. Note however that in order for a lower-bound to be informative, it has to approximate the score closely.

**Definition 2.** *Let $S$ be a score function and $\tilde{S}$ be an associated surrogate.*

1. *$\tilde{S}$ is said to be a lower-bound to $S$ if for any pair of random variables $(X, Y)$ and any measurable function $f$, it holds that $\tilde{S}(f, X, Y) \leq S(f, X, Y)$.*

2. *$\tilde{S}$ is said to be calibrated with respect to $S$ if for any sequence of measurable functions $\{f_l\}_{l \geq 1}$ and any pair of random variables $(X, Y)$, it holds that $\tilde{S}(f_l, X, Y) \to \tilde{S}^* \Rightarrow S(f_l, X, Y) \to S^\dagger$ when $l \to \infty$, where $\tilde{S}^* \doteq \sup_f \tilde{S}(f, X, Y)$ and $S^\dagger \doteq \sup_f S(f)$ are the suprema taken over all measurable functions.*

The framework of surrogate maximization was initially developed for the case where the score function was taken to be accuracy:

$$S^{\mathrm{A}}(f, X, Y) = \mathbb{E}_{X,Y}[\mathbf{1}_{\geq 0}(f(X)Y)], \tag{4}$$

i.e., the probability of correct predictions. For this choice of score, several surrogates have been proposed and studied in the literature [2,22] . Among them is the logistic surrogate defined by

$$\tilde{S}^{\mathrm{A}}_{\log}(f, X, Y) = \mathbb{E}_{X,Y}[\log_2(2 \cdot \sigma(f(X)Y))], \tag{5}$$

where $\sigma(t) = 1/(1 + e^{-t})$ is the sigmoid function. It can be shown that the properties described above hold for this choice of score and surrogate, e.g., that $\tilde{S}^{\mathrm{A}}_{\log}$ is concave and a calibrated lower-bound to $S^{\mathrm{A}}$.

**Proposition 1.** *$\tilde{S}^{\mathrm{A}}_{\log}$ is concave and a calibrated lower-bound to $S^{\mathrm{A}}$.*

*Proof.* See [2].

When the data is very imbalanced, as when the probability $\mathbb{P}[Y = +1]$ is much higher than $\mathbb{P}[Y = -1]$ or vice versa, accuracy sometimes does not capture the essence of what practitioners want to study and other alternative scores are considered [11,14,15,17]. One such score is the $F_1$-score, which is commonly also referred to as Dice:

$$S^{\mathrm{D}}(f, X, Y) = \frac{\mathbb{E}_{X,Y}[2 \cdot \mathbf{1}_{\geq 0}(f(X)) \cdot \mathbf{1}_{\geq 0}(Y)]}{\mathbb{E}_{X,Y}[2 \cdot \mathbf{1}_{\geq 0}(f(X)) \cdot \mathbf{1}_{\geq 0}(Y) + \mathbf{1}_{<0}(f(X)) \cdot \mathbf{1}_{\geq 0}(Y) + \mathbf{1}_{\geq 0}(f(X)) \cdot \mathbf{1}_{<0}(Y)]}. \tag{6}$$

For this choice of score, it was recently shown in [1] that a surrogate given by

$$\tilde{S}^{\mathrm{D}}_{\mathrm{cal}}(f, X, Y) = \frac{\mathbb{E}_{X,Y}[2 \cdot \phi(f(X)) \cdot \mathbf{1}_{\geq 0}(Y)]}{\mathbb{E}_{X,Y}[2 \cdot \mathbf{1}_{\geq 0}(Y) + (1 - \phi(f(X))) \cdot \mathbf{1}_{\geq 0}(Y) + (1 - \phi(-f(X))) \cdot \mathbf{1}_{<0}(Y)]}, \tag{7}$$

where $\phi(t) = 1 + \log_2(\sigma(\max\{t, t/3\}))$, is a quasi-concave calibrated lower-bound.

**Proposition 2.** *$\tilde{S}^{\mathrm{D}}_{\mathrm{cal}}$ is quasi-concave and a calibrated lower-bound to $S^{\mathrm{D}}$.*

*Proof.* See [1].

We refer to this surrogate as *cal-Dice*.

## 3 Semantic Segmentation

Let $(I, S)$ be a pair of $\mathbb{P}$-measurable random variables taking values in $\mathcal{I} \times \mathcal{S} = \mathbb{R}^{M_1 \times \cdots \times M_D} \times \{\pm 1\}^{M_1 \times \cdots \times M_D}$ and $F : \mathbb{R}^{M_1 \times \cdots \times M_D} \to \mathbb{R}^{M_1 \times \cdots \times M_D \times K}$ be a feature extraction function, extracting $K$ features to each $D$-pixel (generalized pixel in $D$-dimensions). Furthermore, let a pair of (conditional) random variables $(X_{|I,S}, Y_{|I,S})$ be uniform over $\{(F(I)_j, S_j)\}_{j \in \mathcal{J}}$, where $\mathcal{J} = \{1, \ldots, M_1\} \times \cdots \times \{1, \ldots, M_D\}$. Now, given a set of real valued functions $\mathcal{F} \subset \mathbb{R}^K$, the problem of segmentation can be seen as to find an $f \in \mathcal{F}$ that maximizes some average score:

$$f_* = \arg\max_{f \in \mathcal{F}} \mathbb{E}_{I,S}[S(f, X_{|I,S}, Y_{|I,S})]. \tag{8}$$

Here, $I$ represents an input image and $S$ represents the associated ground truth. Furthermore, if we consider a U-net in 3D that uses zero-padding [5], then $F$ can be thought of as zero padded patches surrounding each voxel and $f$ can be thought of as the convolutional-kernel to the whole U-net.

Because of the discrete nature of the score functions often considered, $S$ is typically approximated by some surrogate $\tilde{S}$. Furthermore, since we do not have access to the full distribution $\mathbb{P}(I, S)$, we typically collect a set of samples $\{(I^i, S^i)\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{P}(I, S)$, and use the empirical distribution for approximation. This together yields

$$\mathbb{E}_{I,S}[S(f, X_{|I,S}, Y_{|I,S})] \approx \frac{1}{N} \sum_{i=1}^{N} \tilde{S}(f, X_{|I=I^i, S=S^i}, Y_{|I=I^i, S=S^i}). \tag{9}$$

Classically in the segmentation community, it has been standard to train models by minimizing the cross-entropy loss [5,7,13,19]:

$$L_{\mathrm{CE}}(f, F(I^i), S^i) = -\frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} [\mathbf{1}_{\geq 0}(S^i_j) \cdot \log_2(\sigma(f(F(I^i)_j))) + \mathbf{1}_{<0}(S^i_j) \cdot \log_2(1 - \sigma(f(F(I^i)_j)))]. \tag{10}$$

By simple computation, it can be shown that:

$$L_{\mathrm{CE}}(f,(I^i),S^i) = 1 - \tilde{S}^{\mathrm{A}}_{\log}(f, X_{|I=I^i,S=S^i}, Y_{|I=I^i,S=S^i}), \tag{11}$$

and so we have that minimizing $L_{\mathrm{CE}}$ is equivalent to maximizing $\tilde{S}^{\mathrm{A}}_{\log}$. Thus, if the evaluation score considered is accuracy, then the theory from the previous section motivates the choice.

For the more common situation where Dice is used for evaluation, it has become increasingly popular during the last couple of years to consider a smoothed version of Dice referred to as soft-Dice [3,6,8–10,18,21,24]:

$$\tilde{S}^{\mathrm{D}}_{\mathrm{soft}}(f,X,Y) = \frac{\mathbb{E}_{X,Y}[2\cdot\sigma(f(X))\cdot\mathbf{1}_{\geq 0}(Y)]}{\mathbb{E}_{X,Y}[2\cdot\sigma(f(X))\cdot\mathbf{1}_{\geq 0}(Y) + (1-\sigma(f(X)))\cdot\mathbf{1}_{\geq 0}(Y) + \sigma(f(X))\cdot\mathbf{1}_{<0}(Y)]}. \tag{12}$$

This choice of surrogate has been been shown to yield good results experimentally [3,6,8,10,21]. However, because of some reported problems associated with handling noisy data [4,8,16] and an increased risk of convergence issues [8], it is common for practitioners to sacrifice some performance for some stability by using the sum of the soft-Dice loss and the cross-entropy loss as the objective during training [8,9,24].

While there does not to our knowledge exist any work proving that soft-Dice is calibrated to Dice, we conjecture that this is the case because of how closely they are related. As for the other properties discussed in the previous section, we show in Theorem 1 that the surrogate in general is neither quasi-concave nor a lower bound to Dice.

**Theorem 1.** *$\tilde{S}^{\mathrm{D}}_{\mathrm{soft}}$ is neither quasi-concave, nor is it a lower-bound to $S^{\mathrm{D}}$.*

*Proof.* See the supplementary document.

These theoretical considerations together with the experimental reports from the previous works motivates the following hypotheses:

1. soft-Dice could yield better experimental results than cross-entropy when evaluated with Dice because it might be calibrated to Dice,
2. soft-Dice could be less stable than cross-entropy because of properties related to concavity.

In light of this, using a linear combination of soft-Dice and cross-entropy when evaluated with Dice can informally be seen as *trading some consistency for some concavity*. However, it is easy to verify that the resulting composite surrogate in general is neither quasi-concave nor a calibrated lower bound to Dice.

To avoid sacrificing performance or stability, we propose to replace the soft-Dice surrogate with the recently studied cal-Dice surrogate. Two arguments can be made to support this. Firstly, cal-Dice has been proven to be calibrated to Dice whereas soft-Dice, to the best of our knowledge, has only been conjectured to be calibrated to Dice. Secondly, cal-Dice is a quasi-concave lower-bound to

Dice whereas soft-Dice is neither quasi-concave nor a lower bound to Dice. If the hypotheses are valid, cal-Dice will achieve similar performance to soft-Dice without sacrificing stability. To see if this is the case, we proceed by comparing cal-Dice to soft-Dice in a realistic segmentation experiment.

## 4 Experiments

For our experiments, we take the 100 first cases from the Kits2019 competition [7]. Volumes of $256 \times 128 \times 64$ voxels with the kidney centered are then cut out on a resolution of $0.15 \times 0.15 \times 0.35$ (cm), where the last dimension is the *slice direction*. We also pre-process the data in the same way as the winners of the competition by clipping the CT-values to the interval $(-79, 304)$, subtracting by 101 and finally dividing by 76.9 [9]. In Fig. 1 we show an illustration of a slice from one of the patients together with the associated ground truth of the kidney.

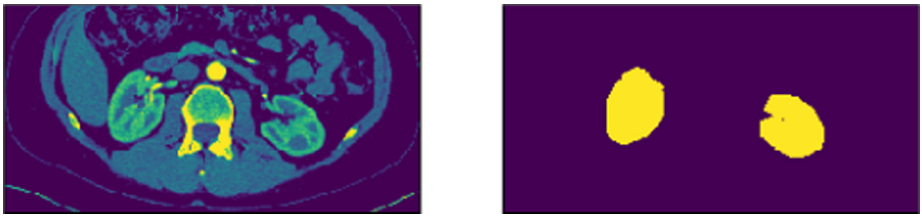

**Fig. 1.** Illustration of one of the 64 slices in one sample patient. To the left is the CT and to the right is the associated label map of the kidney.

The architecture used is a 3D U-net [5,19] with the following properties. Each layer but the last uses instance normalization, relu-activations and has a convolutional kernel of size $7 \times 5 \times 3$. The last layer does not use instance normalization or any activation function and furthermore has a convolutional kernel of $1 \times 1 \times 1$. The first layer uses 32 filters, then for each downsampling, the number of filters is doubled. Five downsamplings and upsamplings are performed in total using 2-strided convolutions and 2-strided transposed-convolutions, and after each downsampling and upsampling there is one regular convolutional layer.

Since we in practice often only have access to a few cases for training, we conduct a 10-fold study using only 10 patients for training and 90 for test. The model is trained using the Adam optimizer [12] with a learning rate of $10^{-3}$ and a batch size of 1. Furthermore, we shuffle the samples for each epoch and train for 1000 epochs in total. The result from the experiments is depicted in Fig. 2 and Fig. 3.

Based on two observations, we argue that the outcome of the experiment support the claim that cal-Dice achieves similar performance to soft-Dice without

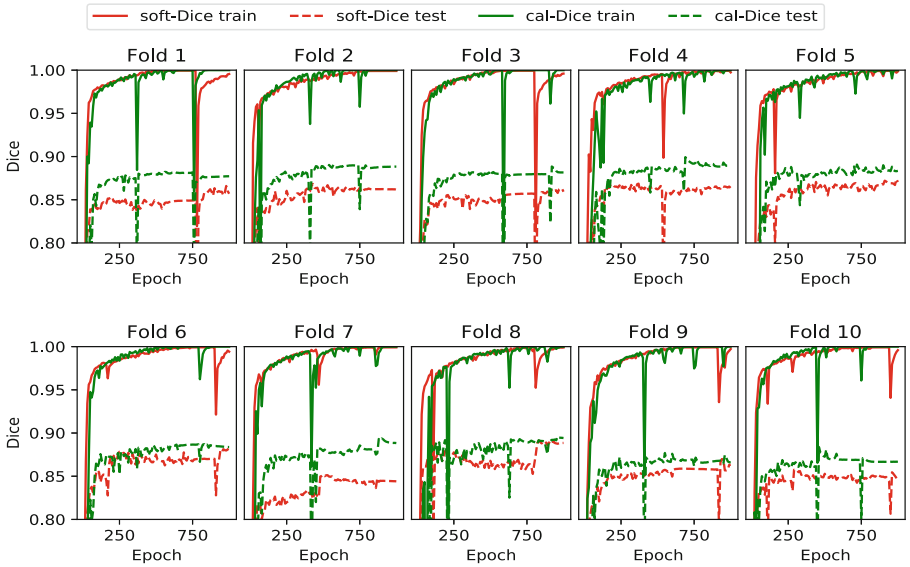


**Fig. 2.** Illustration of the ten fold experiments when training on 10 patients and testing on 90 patients. The whole red line illustrates the Dice score of the training data during training when using the soft-Dice surrogate and the dashed red line illustrates the Dice score of the testing data during training when using the soft-Dice surrogate. Similarly, the whole green illustrates the Dice score of the training data during training when using the cal-Dice surrogate and the dashed red line illustrates the Dice score of the testing data during training when using the cal-Dice surrogate. (Color figure online)

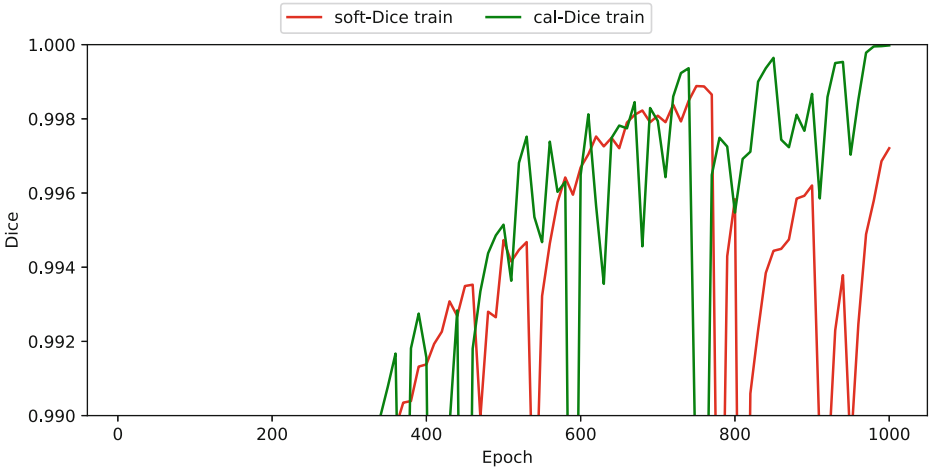


**Fig. 3.** Average performance on the training set over all of the folds when using soft-Dice and cal-Dice. The plot illustrates that cal-Dice pushes to perfect Dice whereas soft-Dice on average starts to get unstable when getting close to perfect Dice. (Color figure online)

sacrificing stability. Firstly, in Fig. 2, a systematic improvement in generalization is clearly visible. Since the training set is rather small, noise will affect the training even though the problem is not considered a particularly noisy segmentation problem. Hence, the improved generalization can be interpreted as an improvement in handling noise. Secondly, in Fig. 3, the average performance on the training set over all of the folds is depicted for soft-Dice and cal-Dice. We see that the neural network when trained using cal-Dice is able to perfectly represent the training data, but also, that this on average is not the case when the neural network is trained using soft-Dice. This can be interpreted as if cal-Dice has less convergence issues than soft-Dice.

We end with a speculation of why these effects are observed. Consider a segmentation problem with two pixels where the ground truth labels are both positive. In Fig. 4, the analytic gradient path for such a problem is depicted from a specific starting point for both soft-Dice and cal-Dice. Since the maximum Dice is given when $f(x_1) \geq 0$ and $f(x_2) \geq 0$, both trajectories lead to an optimal solution. However, soft-Dice focuses on improving one pixel at a time which might encourage the learning of many concrete features. On the other hand, cal-Dice, focuses on improving both pixels simultaneously, which might encourage the learning of few abstract features. This might, since making decisions based on few abstract features often is more robust to noise than making decisions based on many concrete features, be the reason to why we observe better generalization for cal-Dice than for soft-Dice. Furthermore, when learning pixels sequentially compared to learning pixels simultaneously, there might be an increased risk of getting into situations where a feature that is learnt to represent one pixel later is forgotten when focus is on another pixel. This could explain why we observe more convergence issues with soft-Dice than with cal-Dice.

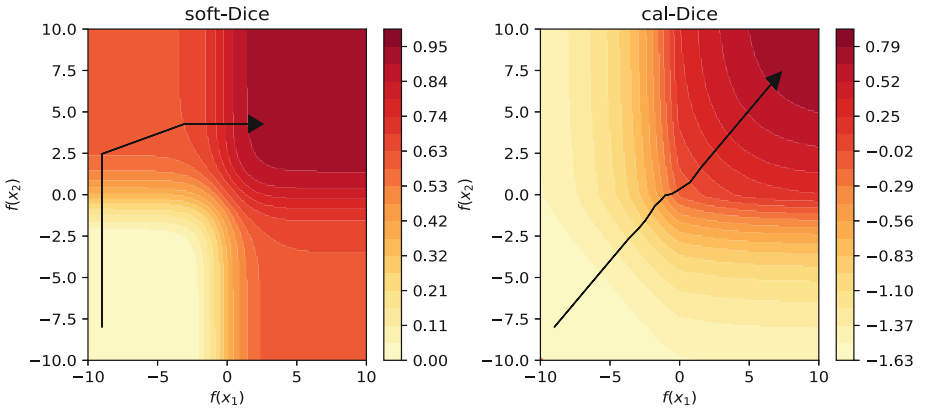


**Fig. 4.** Illustration of analytic gradient trajectories for soft-Dice and cal-Dice in a two pixel segmentation problem when the ground truth labels for both pixels are positive.

## 5 Conclusion

In this work, we have gone through the theoretical background for surrogate maximization and discussed a new surrogate for Dice that we refer to as cal-Dice. We have theoretically compared cal-Dice with soft-Dice and showed that cal-Dice has some better properties. Finally, we have shown some experimental results that support the claim that cal-Dice improves on the stability issues previously reported for soft-Dice. We conclude that cal-Dice, for theoretical and for experimental reasons, can be considered a promising alternative to soft-Dice.

**Acknowledgement.** Marcus Nordström, Fredrik Löfman, Henrik Hult and Atsuto Maki were supported by RaySearch Laboratories. Masashi Sugiyama was supported by the International Research Center for Neurointelligence (WPI-IRCN) at The University of Tokyo Institutes for Advanced Study.

## References

1. Bao, H., Sugiyama, M.: Calibrated surrogate maximization of linear-fractional utility in binary classification. In: International Conference on Artificial Intelligence and Statistics, pp. 2337–2347 (2020)
2. Bartlett, P.L., Jordan, M.I., McAuliffe, J.D.: Convexity, classification, and risk bounds. J. Am. Stat. Assoc. **101**(473), 138–156 (2006)
3. Bertels, J., et al.: Optimizing the dice score and Jaccard index for medical image segmentation: theory and practice. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 92–100. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_11
4. Bertels, J., Robben, D., Vandermeulen, D., Suetens, P.: Optimization with soft dice can lead to a volumetric bias. In: Crimi, A., Bakas, S. (eds.) BrainLes 2019. LNCS, vol. 11992, pp. 89–97. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-46640-4_9
5. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
6. Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C.: The importance of skip connections in biomedical image segmentation. In: Carneiro, G., et al. (eds.) LABELS/DLMIA - 2016. LNCS, vol. 10008, pp. 179–187. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46976-8_19
7. Heller, N., et al.: The Kits19 Challenge Data: 300 Kidney Tumor Cases With Clinical Context, CT Semantic Segmentations, and Surgical Outcomes (2019)
8. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: No new-net. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 234–244. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_21
9. Isensee, F., Maier-Hein, K.H.: An attempt at beating the 3D U-net. arXiv preprint arXiv:1908.02182 (2019)

10. Jadon, S., et al.: A comparative study of 2D image segmentation algorithms for traumatic brain lesions using CT data from the ProTECTIII multicenter clinical trial. In: Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications. vol. 11318, p. 11318 0Q. International Society for Optics and Photonics (2020)
11. Kar, P., Narasimhan, H., Jain, P.: Surrogate functions for maximizing precision at the top. In: International Conference on Machine Learning, pp. 189–198 (2015)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (2015)
13. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
14. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. IEEE Trans. Syst. Man Cybern. Part B (Cybern. ) **39**(2), 539–550 (2008)
15. Liu, X.Y., Zhou, Z.H.: The influence of class imbalance on cost-sensitive learning: an empirical study. In: Sixth International Conference on Data Mining (ICDM 2006), pp. 970–974. IEEE (2006)
16. Mehrtash, A., Wells III, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. arXiv preprint arXiv:1911.13273 (2019)
17. Menon, A., Narasimhan, H., Agarwal, S., Chawla, S.: On the statistical consistency of algorithms for binary classification under class imbalance. In: International Conference on Machine Learning, pp. 603–611 (2013)
18. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: Fourth International Conference on 3D vision (3DV), pp. 565–571. IEEE (2016)
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Steinwart, I.: How to compare different loss functions and their risks. Constr. Approximat. **26**(2), 225–287 (2007)
21. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS - 2017. LNCS, vol. 10553, pp. 240–248. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_28
22. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (2013). https://doi.org/10.1007/978-1-4757-3264-1
23. Zhang, T.: Statistical behavior and consistency of classification methods based on convex risk minimization. Ann. Stat. **32**, 56–85 (2004)
24. Zhang, Y., et al.: Cascaded volumetric convolutional network for kidney tumor segmentation from CT volumes. arXiv preprint arXiv:1910.02235 (2019)