A study on loss functions and decision thresholds for the segmentation of multiple sclerosis lesions on spinal cord MRI

Burhan Rashid Hussein, Cédric Meurée, Malo Gaubert, Arthur Masson, Anne Kerbrat, Benoît Combès, Francesca Galassi

A STUDY ON LOSS FUNCTIONS AND DECISION THRESHOLDS FOR THE SEGMENTATION OF MULTIPLE SCLEROSIS LESIONS ON SPINAL CORD MRI

Burhan Rashid Hussein[1], *Cédric Meurée*[1], *Malo Gaubert*[1,3], *Arthur Masson*[1], *Anne Kerbrat*[1,2], *Benoît Combès*[1*], *Francesca Galassi*[1*]

[1] Univ Rennes, Inria, CNRS, Inserm IRISA UMR 6074, Empenn ERL U 1228, Rennes, France

[2] Rennes University Hospital (CHU), Department of Neurology, Rennes, France

[3] Rennes University Hospital (CHU), Department of Neuroradiology, Rennes, France

[*] Authors contributed equally

ABSTRACT

Multiple sclerosis (MS) patients often present hyper-intense T2-w lesions in the spinal cord. The severe imbalance between background and lesion classes poses a major challenge to Deep Learning segmentation approaches, requiring for ad hoc strategies. Careful selection of the loss function and adjustment of the conventional 0.5-thresholding may help mitigating this issue. Our results show the performance advantages of loss functions based on the Tversky Index and the benefits of threshold tuning over more standard settings and the state-of-the-art model for MS lesion segmentation on spinal cord MRI.

***Index Terms*—** Multiple sclerosis, segmentation, spinal cord, loss function, decision threshold, medical imaging

1. INTRODUCTION

Multiple Sclerosis (MS) is a chronic inflammatory-demyelinating disease of the central nervous system [1]. Magnetic Resonance Imaging (MRI) is fundamental to characterizing and quantifying MS lesions. The number and volume of lesions are used for MS diagnosis, to track its progression, and to evaluate treatments [2]. Accurate identification of MS lesions in MR images is extremely difficult due to variability in lesion location, size, and shape, in addition to anatomical variability between subjects. Since manual segmentation requires expert knowledge, it is time-consuming and prone to intra- and inter-expert variability, methods to automatically segment lesions are required.

Deep Learning approaches have shown remarkable performances in medical imaging segmentation tasks. While automatic MS lesion segmentation on brain MRI is a well-studied and addressed problem [3], automatic MS lesion segmentation on spinal cord MRI has rarely been considered [2]. In this work we focus on the latter. It is important to emphasize that MS lesion segmentation in spinal cord MRI remains a more complex task than in brain MRI due to lesion contrast, image quality, a limited amount of annotated data, and image variability arising from the diversity of MR scanners and acquisition protocols.

A fundamental challenge in MS lesions segmentation in spinal cord MRI is handling the high-class imbalance. Loss functions used in the training of deep learning models differ in their robustness to class imbalance. Careful selection of the loss function is thus crucial. To inform loss function choice, we perform a large-scale loss function comparison.

Loss functions based on the Cross-Entropy, Dice Similarity Coefficient, and Tversky Index have been proposed in the literature to mitigate the effect of imbalanced data. Ma et al. [4] conducted one of the most exhaustive evaluation of segmentation loss functions in medical imaging, comparing 20 losses over four organ segmentation tasks. None of the losses could consistently achieve the best performance on the four tasks, indicating the importance of conducting a loss function comparison to identify the optimal loss for the specific dataset. Jadon et al. [5] reported significantly improved performance metrics for the segmentation of brain tissues using Tversky and Focal Tversky losses, compared to Dice score variants. In the work from Gros et al. [2] on spinal cord MS lesions segmentation, a standard Dice Loss function was used, without reporting the investigation of alternative losses.

Recent studies suggest that the decision threshold on the output probability map (p-map) may play an important role when dealing with highly imbalanced data [6]. In binary segmentation, the decision threshold is conventionally set to 0.5. Optimizing this value may improve the detection of the minority class, i.e. lesion class in our case. Compared to other approaches, adjustment of the decision threshold has the advantage of not altering the input data nor requiring retraining of the model, as it can be applied as a postprocessing step.

This work investigates state-of-the-art segmentation losses and the relevance of adjusting the decision threshold in the complex, highly imbalanced and underexplored task of MS

Accepted at ISBI 2023

lesion segmentation on spinal cord MRI.

2. MATERIAL AND METHODS

The methods evaluated in the context of this work are integrated into a pipeline developed in Python 3.8 and based on TensorFlow 2.9 as well as the Spinal Cord Toolbox (SCT) [2].

2.1. Loss functions

Model convergence at training and performance scores depend on the loss function. To handle class imbalance in image segmentation, the Dice Loss was proposed as an alternative to the widely used Binary Cross Entropy (BCE) [7], and in combination with the BCE under the name of Combo Loss [8]. The Log-Cosh Dice Loss is a recently proposed variant that tackles the non-convex nature of the Dice Loss to facilitate model convergence [5]. The Generalized Dice Loss is another variant introduced to handle highly imbalanced segmentation tasks, relying on weighting factors applied to the different classes based on their relative representation in the training data [9]. The Tversky Loss [10], an extension of the Dice Loss, incorporates weights to control the trade-off between false positive (FP) and false negative (FN) samples. The Focal Tversky Loss extends the Tversky Loss by introducing a γ parameter to make the loss more sensitive to difficult samples and well-adapted to highly imbalanced problems [11]. In the context of this work, the default parameter values proposed in the original associated publications have been selected (i.e., for the Tversky Loss, $\alpha = 0.7$, and for the Focal Tversky Loss, $\gamma = 0.75$).

2.2. MRI Data Acquisition

The dataset includes 161 T2-w cervical and thoracic MRI scans from 108 subjects (35.84 ± 10.42 years old, 86 women). All patients were diagnosed with a form of MS. Acquisitions were performed in 13 different sites (four MR scanner brands). Image resolution is (0.58 ± 0.12, 0.58 ± 0.12, 2.81 ± 0.20) mm^3. Trained neurologist manually segmented lesions. Training, validation and test sets were generated by assigning a subject to a single set while balancing for lesion loads, as shown in figure 1. One test set (thr-optimization set) was used for optimising the decision threshold, the other test set (unseen set) was used at the test time only. All scans without expert-detected lesions were assigned to the unseen set.

2.3. Preprocessing

The preprocessing steps are similar to the ones proposed in [2]. Input images were reoriented and resampled to 0.5 mm isotropic images through linear interpolations. Images were cropped by selecting 48x48 2D patches in each axial plane around the SC centerline voxels [12]. 48x48x48 3D patches were then selected along the inferior-to-superior axis. An

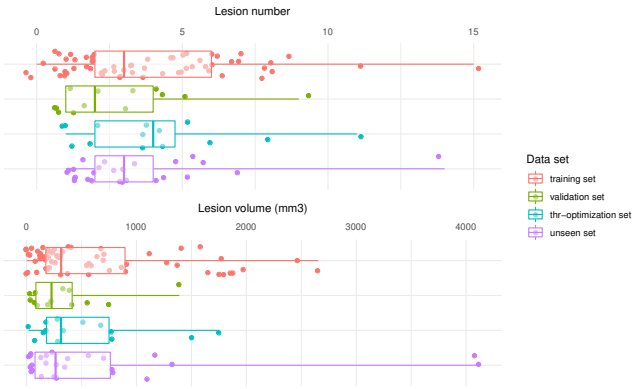

Fig. 1. Boxplots for lesion number and volume in the training (62 MRI scans), validation (13), thr-optimization (14) and unseen set (26, 46 with no MS lesions are not included here). Points are jittered to improve visualization.

overlap of 75% between consecutive patches was introduced as a data augmentation technique. Intensity normalization was applied on stacked patches obtained from a given volumetric image to homogenize the intensity distributions [13]. Finally, each patch was normalized to zero mean and unit standard deviation. Both patches with lesion voxels and without lesion voxels were used as inputs to the model.

2.4. Model Training

The models trained to segment spinal cord MS lesions were based on a 3D-Unet architecture [14]. 16, 32, and 64 filters were selected, associated with 3x3x3 convolutions, 2x2x2 max-pooling layers, and batch normalization applied after each convolution layer. Batch size was 8, Adam optimizer with an initial learning rate of 10^{-4} was used, except for the Combo Loss, where Adadelta optimizer and a learning rate of 1 were used. The learning rate was decreased by a factor of 2 every 15 epochs if no decrease in validation loss occurred. The training process was stopped after 300 epochs, the model providing the best validation loss. Data augmentation operations were applied, including randomly mirroring and rotating patches by 90°.

2.5. Prediction and Postprocessing

Performing predictions on overlapping patches has been shown to generate probability maps of increased reliability [15]. As in-house experiments confirmed this aspect, predictions were computed on 75% overlapping patches. Hence, for the overlapping regions, the mean probability at a given voxel was calculated to generate the output p-maps.
The output binary map was obtained by applying a decision threshold on the corresponding p-map. In our experiments, both the conventional 0.5 and the optimal decision threshold

values were assessed. The median voxel-wise F1 score of the first test set was used to select the optimal threshold for each loss function. For this purpose, threshold values were evaluated over the range from 0 to 1, with an interval of 0.05. The unseen test set was used to assess the influence of optimal thresholds as compared to the conventional value of 0.5.

2.6. Evaluation metrics

Evaluation metrics were based on those presented in Gros et al. [2] and brain MS segmentation challenges [3]. Sensitivity, precision, and voxel-wise F1-score (V-F1) were used to measure the voxel-wise agreement between the ground truth and the binarized predictions. Similar metrics were defined at the lesion level, including lesion-wise sensitivity, lesion-wise precision, and lesion-wise F1 score (L-F1). We considered a candidate lesion as correctly detected if the lesion connected voxels overlapped the ground truth connected voxels by at least 25% and did not go outside by more than 70%. Subject-level sensitivity and specificity were also calculated to assess the capability of a model to classify subjects with and without ground truth lesions correctly. The Wilcoxon signed-rank test was used to assess the significance of the metrics differences between two methods.

3. EXPERIMENTAL RESULTS

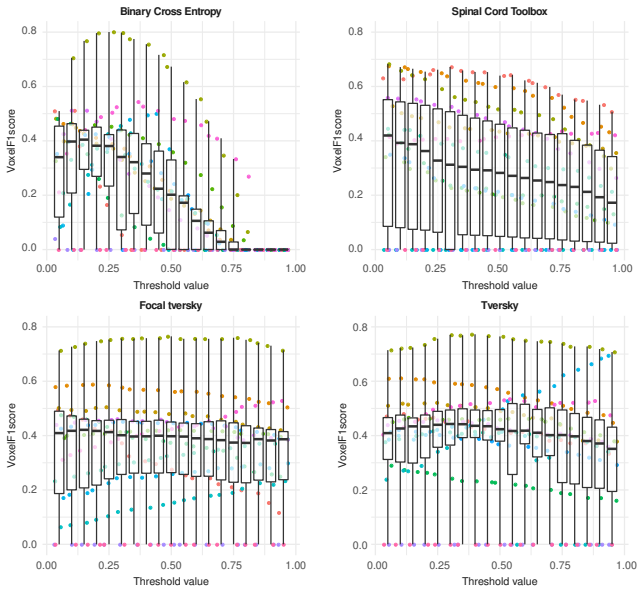

Fig. 2. Boxplots of voxel-wise F1 score for BCE, SCT, Focal Tversky and Tversky losses at each decision threshold value. One point corresponds to one subject.

The results presented in table 1 indicate that loss functions based on the standard DSC lead to high voxel-wise precision and low voxel-wise sensitivity scores, the predictions being biased towards the non-lesion class, i.e. many lesion voxels are missed yielding a high number of FN voxels. The use of the Generalised Dice Loss and loss functions based on the Tversky Index results in a better trade-off between voxel-wise precision and sensitivity. The highest V-F1 are achieved with the Focal Tversky Loss (0.398 on the threshold-optimization (thr-optimization) set (0.402 on the unseen set)) and Tversky Loss (0.424 (0.398)). The parameters of these loss functions were designed to weight sensitivity higher than precision, which resulted in a consistent improvement in the performance metrics on both datasets.

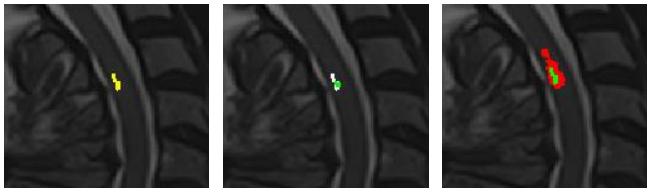

Fig. 3. From the left: ground truth (yellow) overlaid a sagittal T2-w image (from the unseen test set); segmentation with the Tversky Loss (Dice: 0.49), and with the Dice Loss (Dice: 0.29). The binary output was obtained by thresholding the p-map at the optimal value. FN and FP voxels are white and red, respectively; TP voxels are green.

The detection and count of lesions is vital in monitoring MS patients undergoing disease-modifying treatments. In this context, lesion-wise metrics have more value for clinicians than voxel-wise metrics. As both FP and FN lesions are undesirable, the L-F1 is favored, i.e. the harmonic mean of precision and sensitivity. Similarly to the voxel-wise analysis, the highest L-F1 is observed for losses based on the Tversky Index (TL: 0.523 on the thr-optimization set (0.336 on the unseen set)); Focal Tversky Index (FTL: 0.467 on the thr-optimization set (0.367 on the unseen set)). It has to be noted that the L-F1 score strictly depends on the chosen definition of correctly detected lesions.

In this study, the decision threshold was adjusted to maximize the V-F1. Figure 3 shows an example of the segmentation outputs obtained with the DSC Loss and the Tversky Loss on an image of the unseen test set using the optimized threshold values. The results presented in table 2 suggest that optimizing the decision threshold can affect the overall performance of a model with a significant boost in both the V-F1 and L-F1 metrics. Significant improvements were reported for BCE (V-F1 +20$\%$ (+32$\%$), L-F1 +37$\%$ (+27$\%$)), the state-of-the-art SCT model (V-F1 +14$\%$ (+16$\%$), L-F1 +20$\%$ (+39$\%$)), and slight improvements for DSC (V-F1 +20$\%$ (+1.5$\%$), L-F1 +37$\%$ (+1.5$\%$)). For all these loss functions, the performance metrics improved by lowering the decision threshold to a value lower than 0.5.

In figure 2, we report the voxel-wise F1 score boxplots for the BCE, SCT, Focal Tversky and Tversky losses. While the Tversky Loss showed minor improvements (V-F1 +2$\%$

Table 1. Metrics (median (mean ± std)) obtained on the thr-optimization and the unseen subsets (first and second row associated with each loss function respectively), using 0.5 as a decision threshold value.

Loss	L-sensitivity	L-precision	L-F1	V-sensitivity	V-precision	V-F1	P-sensitivity	P-specificity
BCE	0 (0.13±0.27)	0 (0.21±0.32)	0 (0.15±0.28)	0.11 (0.14±0.16)	0.69 (0.54±0.43)	0.2 (0.21±0.21)	0.36	0.41
	0 (0.12±0.34)	0 (0.28±0.38)	0 (0.16±0.33)	0.01 (0.09±0.23)	0.38 (0.47±0.42)	0.01 (0.14±0.24)	0.31	0.91
Combo	0.43 (0.41±0.38)	0.42 (0.39±0.31)	0.5 (0.35±0.25)	0.31 (0.32±0.25)	0.48 (0.48±0.33)	0.41 (0.35±0.22)	0.71	0.91
	0 (0.23±0.38)	0 (0.3±0.3)	0 (0.25±0.3)	0.12 (0.18±0.34)	0.61 (0.46±0.28)	0.22 (0.23±0.25)	0.46	0.52
Dice	0.44 (0.46±0.34)	0.5 (0.43±0.35)	0.41 (0.4±0.3)	0.4 (0.42±0.26)	0.52 (0.49±0.29)	0.44 (0.4±0.2)	0.79	0.28
	0.33 (0.35±0.4)	0.28 (0.28±0.34)	0.3 (0.29±0.32)	0.28 (0.31±0.29)	0.48 (0.45±0.37)	0.35 (0.31±0.27)	0.62	0.3
Focal Tversky	0.5 (0.47±0.38)	0.38 (0.35±0.29)	0.47 (0.38±0.29)	0.47 (0.44±0.28)	0.4 (0.41±0.3)	0.4 (0.35±0.21)	0.71	0.37
	0.4 (0.46±0.36)	0.3 (0.3±0.44)	0.37 (0.34±0.36)	0.37 (0.36±0.28)	0.4 (0.38±0.45)	0.4 (0.33±0.28)	0.65	0.13
Generalized Dice	0.25 (0.33±0.33)	0.171 (0.26±0.31)	0.22 (0.25±0.25)	0.53 (0.53±0.29)	0.351 (0.33±0.23)	0.42 (0.35±0.21)	0.64	0.52
	0.46 (0.44±0.36)	0.33 (0.3±0.3)	0.38 (0.33±0.3)	0.52 (0.47±0.3)	0.38 (0.41±0.35)	0.39 (0.36±0.23)	0.65	0.28
Log-Cosh Dice	0.5 (0.54±0.37)	0.47 (0.47±0.34)	0.45 (0.45±0.28)	0.35 (0.34±0.19)	0.56 (0.52±0.31)	0.4 (0.37±0.18)	0.86	0.67
	0.23 (0.35±0.24)	0.16 (0.29±0.43)	0.19 (0.29±0.27)	0.24 (0.27±0.12)	0.45 (0.41±0.48)	0.35 (0.29±0.18)	0.54	0.41
SCT	0.23 (0.27±0.3)	0.25 (0.31±0.32)	0.24 (0.28±0.3)	0.17 (0.2±0.18)	0.8 (0.6±0.4)	0.28 (0.29±0.23)	0.64	0.13
	0 (0.28±0.39)	0 (0.36±0.35)	0 (0.3±0.32)	0.13 (0.22±0.28)	0.76 (0.53±0.35)	0.23 (0.27±0.25)	0.46	0.67
Tversky	0.5 (0.55±0.33)	0.5 (0.5±0.33)	0.52 (0.49±0.28)	0.41 (0.38±0.22)	0.47 (0.47±0.26)	0.42 (0.4±0.21)	0.86	0.3
	0.37 (0.4±0.41)	0.25 (0.32±0.33)	0.34 (0.32±0.31)	0.3 (0.32±0.3)	0.52 (0.42±0.31)	0.4 (0.32±0.25)	0.62	0.37

Table 2. Metrics (median (mean ± std)) obtained on the thr-optimization and the unseen subsets (first and second row associated with each loss function respectively), using an optimized decision threshold value based on the DSC. Significant metric differences (p-value ≤ 0.05) obtained by means of the 0.5 and optimal thresholds are highlighted in bold.

Loss	Optimal threshold	L-sensitivity	L-precision	L-F1	V-sensitivity	V-precision	V-F1	P-sensitivity	P-specificity
BCE	0.15	0.37 (0.41±0.37)	0.42 (0.39±0.34)	0.37 (0.35±0.29)	0.44 (0.47±0.29)	0.35 (0.39±0.27)	0.4 (0.36±0.2)	0.71	0.24
	0.15	**0.27 (0.33±0.36)**	0.25 (0.27±0.3)	0.27 (0.28±0.3)	**0.29 (0.34±0.32)**	0.4 (0.36±0.32)	**0.33 (0.29±0.24)**	0.58	0.24
Combo	0.05	0.25 (0.37±0.4)	0.23 (0.22±0.21)	0.21 (0.26±0.26)	0.48 (0.42±0.29)	0.38 (0.35±0.27)	0.42 (0.34±0.23)	0.64	0.17
	0.05	**0.27 (0.31±0.35)**	0.13 (0.2±0.26)	0.18 (0.22±0.26)	**0.28 (0.28±0.29)**	0.42 (0.39±0.33)	**0.3 (0.27±0.25)**	0.54	0.17
Dice	0.3	0.5 (0.49±0.36)	0.5 (0.38±0.28)	0.5 (0.39±0.26)	0.44 (0.45±0.27)	0.46 (0.44±0.28)	0.46 (0.4±0.2)	0.79	0.26
	0.3	0.33 (0.39±0.39)	0.29 (0.3±0.34)	0.31 (0.31±0.32)	**0.32 (0.34±0.31)**	**0.42 (0.41±0.33)**	0.36 (0.31±0.28)	0.62	0.26
Focal Tversky	0.15	0.5 (0.47±0.36)	0.31 (0.33±0.33)	0.41 (0.35±0.26)	0.55 (0.52±0.28)	0.34 (0.35±0.29)	0.42 (0.35±0.22)	0.71	0.22
	0.15	0.36 (0.41±0.38)	**0.17 (0.24±0.27)**	**0.26 (0.28±0.28)**	**0.43 (0.44±0.34)**	**0.23 (0.29±0.26)**	0.31 (0.31±0.25)	0.65	0.22
Generalized Dice	0.75	0.38 (0.38±0.33)	0.2 (0.31±0.35)	0.25 (0.3±0.28)	0.49 (0.49±0.29)	0.39 (0.38±0.27)	0.44 (0.37±0.21)	0.71	0.2
	0.75	0.45 (0.44±0.39)	0.28 (0.35±0.37)	0.34 (0.35±0.33)	**0.44 (0.41±0.33)**	**0.45 (0.46±0.32)**	0.41 (0.36±0.26)	0.66	0.2
Log-Cosh Dice	0.2	0.5 (0.58±0.35)	0.31 (0.42±0.34)	0.45 (0.42±0.25)	0.41 (0.41±0.23)	0.47 (0.45±0.3)	0.43 (0.38±0.19)	0.86	0.2
	0.2	0.21 (0.33±0.37)	0.12 (0.2±0.23)	0.17 (0.24±0.27)	**0.34 (0.35±0.34)**	**0.32 (0.33±0.32)**	0.37 (0.29±0.25)	0.54	0.2
SCT	0.05	0.37 (0.46±0.4)	0.44 (0.4±0.33)	0.44 (0.41±0.35)	0.33 (0.3±0.23)	0.62 (0.5±0.34)	0.42 (0.36±0.26)	0.71	0.33
	0.05	0.4 (0.42±0.4)	0.27 (0.34±0.35)	0.38 (0.34±0.32)	**0.3 (0.34±0.3)**	**0.63 (0.5±0.35)**	**0.39 (0.35±0.25)**	0.62	0.33
Tversky	0.3	0.55 (0.58±0.35)	0.44 (0.46±0.3)	0.52 (0.46±0.24)	0.46 (0.42±0.24)	0.42 (0.43±0.25)	0.44 (0.4±0.21)	0.86	0.26
	0.3	0.45 (0.44±0.39)	0.25 (0.33±0.32)	0.4 (0.34±0.31)	**0.36 (0.36±0.32)**	0.45 (0.42±0.33)	**0.41 (0.33±0.24)**	0.65	0.26

(+2%), L-F1 +0% (+6%)), the stability of its performance scores across the range of decision thresholds suggests that it is an apt solution for our task and dataset. A similar conclusion can be drawn for the Focal Tversky Loss, for which a minor deterioration was observed across datasets while maintaining a rather stable behavior over the threshold range. A significant decrease in model performance with conventional 0.5-thresholding can be observed for BCE and SCT model. Perturbations of as little as ±0.05 can induce a median reduction in V-F1 score of 10%.

Such observations indicate that when dealing with heterogeneous data and prone to high-class imbalance, certain schemes of the loss function and decision threshold are better suited to accommodate the metrics of interest, i.e. V-F1 and L-F1. Similarly to voxel-wise and lesion-wise metrics, Tversky Loss tends to have a better trade-off when looking at the P-sensitivity and P-specificity (see table 1 and table 2). While threshold optimization based on V-F1 scores tends to improve patient detection rate, it can decrease the performance of a model when presented with non-lesion subjects.

To put the work into context, our results indicate that the combination of the Tversky Loss with an adjusted decision threshold performs very well compared to the latest results in T2-w MS spinal cord lesion segmentation [2].

4. CONCLUSION

This study evaluates segmentation loss functions proposed in the literature to mitigate the issue of imbalanced data, with reference to the complex task of MS lesion segmentation in spinal cord MRI. In addition, the adjustment of the decision threshold on the output p-map is explored to further improve the performance metrics of interest. Results indicate that loss functions based on the Tversky Index and a minimal adjustment of the decision threshold can yield higher median scores and less dispersed output scores than more standard settings. Future work will aim to understand how the variability due to acquisition scanners and protocol, and patient population, can affect the test metrics and can be eventually addressed.

5. COMPLIANCE WITH ETHICAL STANDARDS

All patients provided written consent and were included in one of the four clinical studies registered on clinicaltrials.gov (NCT02889965, NCT02117375, NCT04220814, NCT04918225), and compliant with French data confidentiality regulations. The study was approved by the relevant ethics committee.

6. ACKNOWLEDGMENTS

This study was supported by the French National Research Agency (ANR) within the France 2030 program, 3rd PIA (reference ANR-21-RHUS-0014). Data collection was supported by a grant provided by the French State and handled by ANR within the France 2030 program (reference ANR-10-COHO-002 OFSEP).

7. REFERENCES

[1] S. Leguy, B. Combès, E. Bannier, and A. Kerbrat, "Prognostic value of spinal cord MRI in multiple sclerosis patients," *Revue Neurologique*, vol. 177, no. 5, pp. 571–581, May 2021.

[2] C. Gros et al., "Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks," *NeuroImage*, vol. 184, pp. 901–915, Jan. 2019.

[3] O. Commowick et al., "Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure," *Scientific Reports*, vol. 8, no. 1, Sept. 2018.

[4] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A.L. Martel, "Loss odyssey in medical image segmentation," *Medical Image Analysis*, vol. 71, pp. 102035, 2021.

[5] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. Oct. 2020, IEEE.

[6] N. Bice, N. Kirby, R. Li, D. Nguyen, T. Bahr, C. Kabat, P. Myers, N. Papanikolaou, and M. Fakhreddine, "A sensitivity analysis of probability maps in deep-learning-based anatomical segmentation," *J. Appl. Clin. Med. Phys.*, vol. 22, no. 8, pp. 105–119, Aug. 2021.

[7] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," 2016.

[8] S.A. Taghanaki, Y. Zheng, S.K. Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, and G. Hamarneh, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 24–33, 2019.

[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," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 240–248. Springer International Publishing, 2017.

[10] 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*, pp. 379–387. Springer International Publishing, 2017.

[11] N. Abraham and N.M. Khan, "A novel focal tversky loss function with improved attention u-net for lesion segmentation," *CoRR*, vol. abs/1810.07842, 2018.

[12] C. Gros, B. De Leener, S.M. Dupont, A.R. Martin, M.G. Fehlings, R. Bakshi, S. Tummala, V. Auclair, D.G. McLaren, V. Callot, J. Cohen-Adad, and M. Sdika, "Automatic spinal cord localization, robust to MRI contrasts using global curve optimization," *Medical Image Analysis*, vol. 44, pp. 215–227, Feb. 2018.

[13] L.G. Nyul, J.K. Udupa, and X. Zhang, "New variants of a method of MRI scale standardization," *IEEE Transactions on Medical Imaging*, vol. 19, no. 2, pp. 143–150, 2000.

[14] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pp. 424–432. Springer International Publishing, 2016.

[15] N. Pielawski and C. Wählby, "Introducing hann windows for reducing edge-effects in patch-based image segmentation," *PLOS ONE*, vol. 15, no. 3, pp. e0229839, Mar. 2020.