sensors

MDPI

Article

An Analysis of Loss Functions for Heavily Imbalanced Lesion Segmentation

Mariano Cabezas [1,*,†] and Yago Diez [2,†]

[1] Brain and Mind Centre, The University of Sydney, Camperdown, NSW 2050, Australia
[2] Faculty of Science, Yamagata University, Yamagata 990-8560, Japan; yago@sci.kj.yamagata-u.ac.jp
* Correspondence: mariano.cabezas@sydney.edu.au
† These authors contributed equally to this work.

Abstract: Heavily imbalanced datasets are common in lesion segmentation. Specifically, the lesions usually comprise less than 5% of the whole image volume when dealing with brain MRI. A common solution when training with a limited dataset is the use of specific loss functions that rebalance the effect of background and foreground voxels. These approaches are usually evaluated running a single cross-validation split without taking into account other possible random aspects that might affect the true improvement of the final metric (i.e., random weight initialisation or random shuffling). Furthermore, the evolution of the effect of the loss on the heavily imbalanced class is usually not analysed during the training phase. In this work, we present an analysis of different common loss metrics during training on public datasets dealing with brain lesion segmentation in heavy imbalanced datasets. In order to limit the effect of hyperparameter tuning and architecture, we chose a 3D Unet architecture due to its ability to provide good performance on different segmentation applications. We evaluated this framework on two public datasets and we observed that weighted losses have a similar performance on average, even though heavily weighting the gradient of the foreground class gives better performance in terms of true positive segmentation.

Keywords: medical imaging sensors; magnetic resonance imaging; brain; lesion segmentation; imbalanced dataset; loss functions

check for updates

Citation: Cabezas, M.; Diez, Y. An Analysis of Loss Functions for Heavily Imbalanced Lesion Segmentation. *Sensors* **2024**, *24*, 1981. https://doi.org/10.3390/s24061981

Academic Editor: Alessandro Bevilacqua

Received: 11 January 2024
Revised: 13 March 2024
Accepted: 15 March 2024
Published: 20 March 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Magnetic resonance imaging (MRI) is one of the most common techniques using sensors to obtain imaging information from the brain. Furthermore, lesion segmentation is the most common image analysis task applied to brain MRI for different neurological pathologies, such as multiple sclerosis [1,2], Alzheimer's disease [3] or vascular cognitive impairment [4]. Brain lesions, usually identified as white matter hyperintensities on T2-weighted images, represent a small percentage of the whole brain volume, can have different shapes and volumes and their spatial distribution can greatly vary between patients. Due to these issues, the accurate detection and segmentation of white matter hyperintensities is a challenging segmentation task.

Recently, deep learning approaches have rapidly become the state of the art for medical image analysis [5,6], including binary image segmentation [7]. However, a severe data imbalance can bias the training of the models and produce unexpected results (i.e., reduce the accuracy of the lesion segmentation) [8,9]. The most common solutions are the use of sampling techniques to increase the number of positive samples [10] and loss functions that give different weights to the foreground and background classes, either implicitly [11,12] or explicitly [13]. While different losses have been presented for medical imaging [14], they are rarely analysed for datasets with heavy imbalances. Specifically, their effect on the underrepresented class is not properly understood, how the loss evolves through time per epoch is rarely studied, and the effect of randomness is usually excluded from the results,

even though weight optimisation is inherently a stochastic process. To that end, multiple studies have been published on understanding prediction uncertainty through probabilistic frameworks and deep ensembles [15–18], suggesting the importance of reporting results with multiple runs and random seeds. Furthermore, losses are usually defined based on their value, but they are rarely designed taking into account their derivative and their effect on optimisation.

In this paper, we present an analysis of the most common losses used for imbalanced datasets, and we study their evolution during training focusing on the effect on the under-represented class. Furthermore, we propose a new weighted loss function based on the gradient of the cross-entropy loss (as opposed to focusing only on the loss function alone). The rest of the paper is structured as follows: in Section 2, we present the framework and public datasets we used, followed by the results in Section 3, a short discussion in Section 4 and conclusions and future works in Section 5.

2. Materials and Methods

2.1. Public Datasets

For this analysis, we used two public datasets with manual annotations of brain lesions and a heavy imbalance. The statistics of the number of foreground and background voxels are summarised in Table 1 and examples of images are given in Figure 1. While the ratio between foreground and background voxels is bigger inside the training patches, they still represent less than 5% of the whole patch on average.

Table 1. Statistics on the number of voxels for the foreground and background class for the analysed datasets. % (brain) refers to the lesion voxel percentage mean ± standard deviation for the whole brain, while % (batch) refers to the lesion voxel percentage for all the batches. Numbers in brackets represent the global percentage for the whole dataset and batches.

Dataset	% (Brain)	% (Batch)	Mean Lesion Size	Lesion Number
WMH2017	1.54 ± 1.48 [1.57]	1.53 ± 2.40 [1.73]	111.1750	3679
LIT longitudinal	0.14 ± 0.18 [0.15]	0.48 ± 0.84 [0.51]	116.13	156

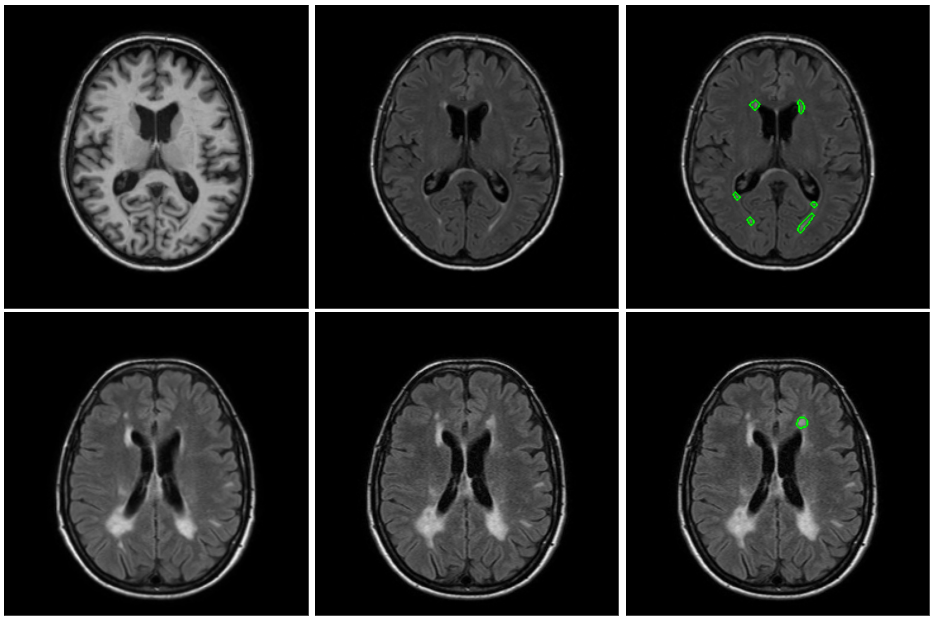

Figure 1. Example of slices of a randomly selected subject for each dataset. The first row depicts an example of a T1 slice, a FLAIR slice and the boundary of all the lesions for a WMH2017 subject, while the second row depicts an example of a baseline, follow-up image and the boundary of all the new lesions (not appearing on the baseline image) for a LIT longitudinal subject.

2.1.1. White Matter Hyperintensities (WMH) Challenge 2017 [4]

The WMH Segmentation Challenge was held during the MICCAI 2017 conference, and it provided a standardized assessment of automatic methods for the segmentation of WMH. The task for the challenge was defined as the segmentation of WMH of presumed vascular origin on brain MR images. A total of 60 labelled images were acquired from three different scanners with different sensors from three different vendors, in three different institutes. For each subject, a 3D T1-weighted and a 2D multi-slice FLAIR image were provided. Details on the acquisition and preprocessing steps for these images are summarised in the original challenge paper [4]. In brief, all images were manually labelled for white matter hyperintensities by an expert, and these annotations were afterwards reviewed by a second expert who created the final mask. Other pathological regions were also labelled. However, these regions were not evaluated during the challenge and, thus, we ignored this label for our analysis. Finally, T1-weighted images were co-registered to the FLAIR space during pre-processing, and we used these pre-processed images for our experiments.

The parameters of each acquisition are summarised as follows:

- UMC Utrecht, 3 T Philips Achieva: 3D T1-weighted sequence (192 slices, slice size: 256 × 256, voxel size: 1.00 × 1.00 × 1.00 mm^3, repetition time (TR)/echo time (TE): 7.9/4.5 ms), 2D FLAIR sequence (48 slices, slice size: 240 × 240, voxel size: 0.96 × 0.95 × 3.00 mm^3, TR/TE/inversion time (TI): 11,000/125/2800 ms).
- NUHS Singapore, 3 T Siemens TrioTim: 3D T1-weighted sequence (192 slices, slice size: 256 × 256, voxel size: 1.00 × 1.00 × 1.00 mm^3, TR/TE/TI: 2, 300/1.9/900 ms), 2D FLAIR sequence (48 slices, slice size: 256 × 256, voxel size: 1.00 × 1.00 × 3.00 mm^3, TR/TE/TI: 9000/82/2500 ms).
- VU Amsterdam, 3 T GE Signa HDxt: 3D T1-weighted sequence (176 slices, slice size: 256 × 256, voxel size: 0.94 × 0.94 × 1.00 mm^3, TR/TE: 7.8/3.0 ms), 3D FLAIR sequence (132 slices, slice size: 83 × 256, voxel size: 0.98 × 0.98 × 1.20 mm^3, TR/TE/TI: 8000/126/2340 ms).

2.1.2. Ljubljana Longitudinal Multiple Sclerosis Lesion Dataset [1]

This database contains baseline and follow-up MR images of 20 multiple sclerosis (MS) patients. The images were acquired on a 1.5 T Philips MRI machine at the University Medical Centre Ljubljana (UMCL), and the data were anonymized [1]. Each patient's MR acquisition contained a 2D T1-weighted (spin echo sequence, repetition time (TR) = 600 ms, echo time (TE) = 15 ms, flip angle (FA) = 90°, sampling of 0.9 × 0.9 × 3 mm with no inter-slice gap resulting in a 256 × 256 × 45 lattice), a 2D T2-weighted (spin echo sequence, TR = 4500 ms, TE = 100 ms, FA = 90°, sampling of 0.45 × 0.45 × 3 mm with no inter-slice gap, resulting in a 512 × 512 × 45 lattice), and a 2D FLAIR image (TR = 11,000, TE = 140, TI = 2800, FA = 90, sampling of 0.9 × 0.9 × 3 mm with no inter-slice gap, resulting in a 256 × 256 × 49 lattice). The median time between the baseline and follow-up studies was 311 days, ranging from 81 to 723 days with the interquartile range (IQR) of 223 days. For our analysis, we only focused on the FLAIR images that were skull stripped and bias corrected using N4, followed by coregistration to the follow-up space using rigid registration. To create a ground truth mask, an image neuroanalyst expert at our group labelled all positive activity (new and enlarging lesions) using the coregistered FLAIR images and their subtraction. Furthermore, 3 of the subjects were excluded due to a lack of positive activity after the labelling. Therefore, only 17 patients from the original dataset were used for the experiments.

2.1.3. Data Preparation

For both datasets we extracted patches of size 32 × 32 × 32 using a sliding window (with an overlap of 16 voxels) on the brain bounding box to minimise the number of voxels that did not belong to the brain. These patches were then sampled to decrease the imbalance between foreground and background voxels. Only patches containing lesion voxels were used for training in groups of 16.

2.2. Network Architecture

In the last few years, the 3D Unet architecture [19] has become the state of the art in volumetric medical image segmentation for different tasks and images [7]. In order to train all the losses with a robust architecture, we decided to use a simple 3D Unet of 2,906,913 parameters using basic convolutional blocks (with 32, 64, 128 and 256 kernels of size $3 \times 3 \times 3$), followed by ReLU activation, group norm (with groups of 8 channels) and max pooling of size 2 (see Figure 2). These hyperparameters were chosen empirically based on the best performance of our previous works on other private datasets for lesion segmentation [20,21]. To guarantee a fair comparison, these hyperparameters were also shared for all the experiments (different seeds and loss functions). The weights were randomly initialised using the same seed for all the losses in a given experiment to reduce the effect of random initialisation. This seed was also used to ensure that the training batches per epoch were shuffled the same way for all the losses.

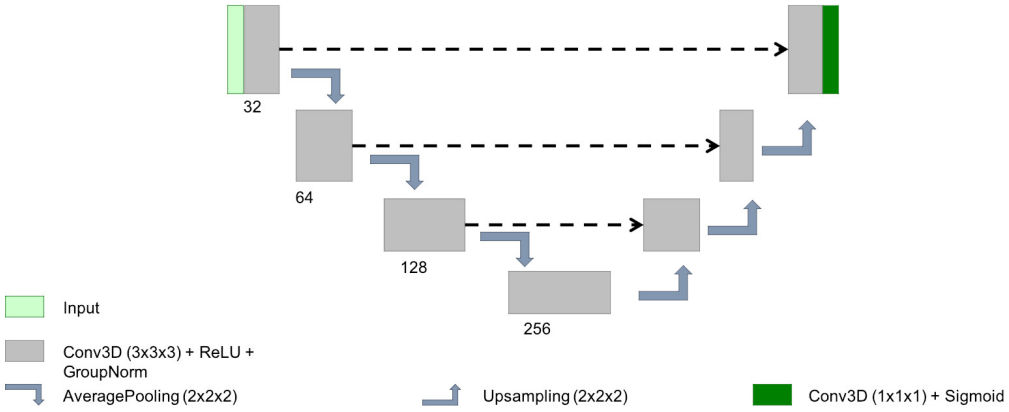

Figure 2. Scheme of the simple Unet architecture used for this analysis.

2.3. Loss Functions

2.3.1. Cross-Entropy

The cross-entropy loss (**xent**) is one of the most commonly used loss functions for deep neural networks. If we consider the distribution of the real labels (y_i) and the predicted ones ($\hat{y}_i$), the cross-entropy expresses the similarity of these two distributions in the following form:

$$\mathcal{L}_{CE}(\hat{y}_i, y_i) = \begin{cases} -\log \hat{y}_i, & \text{if foreground} \\ -\log(1-\hat{y}_i), & \text{if background} \end{cases} \quad (1)$$

In our case, this equation represents the loss for a single voxel, and the final value is computed as the mean of the loss for all the voxels. By definition, all voxels have an equal weight; therefore, during training, the optimiser will tend to favour the majority class.

2.3.2. Focal Loss

The focal loss function was presented by Tsung-Yi Lin et al. [13] for dense object detection in the presence of an imbalance between positive and negative samples. Through the use of a weighting variable (α) and a modulating factor (γ), they extended the cross-entropy loss to increase the importance of positive samples and reduce the effect of samples correctly classified with a high confidence. Following the notation for the cross-entropy loss, the focal loss for a prediction $\hat{y}_i$ can be defined as:

$$\mathcal{L}_{FL}(\hat{y}_i, y_i) = \begin{cases} -\alpha(1-\hat{y}_i)^\gamma \log \hat{y}_i, & \text{if foreground} \\ -(1-\alpha)\hat{y}_i{}^\gamma \log(1-\hat{y}_i), & \text{if background} \end{cases} \quad (2)$$

For our experiments, we used $\gamma = 2$ as suggested on the paper, and two different α values: $\alpha_1 = 0.25$ (**focal1**) and $\alpha_2 = 0.75$ (**focal2**). On the original paper, they suggest to

decrease α as γ is increased, and they noted that $\alpha = 0.25$ gave the best results with $\gamma = 2$. However, we also wanted to test the loss with a higher α value to give a higher weight to positive samples.

2.3.3. Generalised Dice Loss

The generalised Dice loss was presented by Sudre et al. [12] to balance the effect of positive and negative samples, based on a generalisation of the Dice similarity metric (DSC) [22]. This loss function can be defined as:

$$\mathcal{L}_{gDSC}(\hat{Y}, Y) = 1 - 2 \cdot \frac{(w_{fg} + w_{bg}) \cdot \sum_{i}^{N} \hat{y}_i \cdot y_i + w_{bg} \cdot (N - \sum_{i}^{N} (\hat{y}_i + y_i)}{(w_{fg} - w_{bg}) \cdot \sum_{i}^{N} (\hat{y}_i + y_i) + 2 \cdot N w_{bg}}, \tag{3}$$

where $\hat{Y}$ and Y represent the predicted and true segmentation for the N image voxels ($\hat{y}_i$ and y_i), and w_{bg} and w_{fg} are weights for the background and foreground class, respectively.

In the special case where $w_{bg} = 0$, the generalised Dice loss becomes the commonly used binary Dice loss originally presented by Milletari et al. [11]:

$$\mathcal{L}_{DSC}(\hat{Y}, Y) = 1 - 2 \cdot \frac{\sum_{i}^{N} \hat{y}_i \cdot y_i + \epsilon}{\sum_{i}^{N} (\hat{y}_i + y_i) + \epsilon}, \tag{4}$$

where ϵ is a small smoothing constant to avoid numerical issues when dividing by 0. In our experiments, we used the generalised loss (**gdsc**) as defined on the original paper ($w_{bg} = 1/N_{bg}^2$ and $w_{fg} = 1/N_{fg}^2$, where N_{bg} and N_{fg} refer to the number of background and foreground voxels), the binary Dice loss (**dsc**) ($w_{bg} = 0$ and $w_{fg} = 1$) and a combination of the binary Dice loss and cross-entropy with equal weights (**mixed**).

2.3.4. Weighted Gradient Loss

Similarly to the focal loss, we propose to introduce a weighting factor and a modular factor to the cross-entropy loss. However, instead of applying that factor on the loss function itself, we propose to apply it to the gradient of the cross-entropy when combined with the gradient of the sigmoid function as follows:

$$\frac{\partial \mathcal{L}_{wCE}(\hat{y}_i, y_i)}{\partial f_i} = \begin{cases} \alpha(1 - \hat{y}_i)^\gamma, & \text{if foreground} \\ (1 - \alpha)\hat{y}_i^\gamma, & \text{if background} \end{cases} \tag{5}$$

where f_i represents the output of the last layer before applying the final sigmoid activation. In our experiments, we set $\gamma = 2$ and $\alpha = N_{bg}/N$ (**new**).

2.4. Experimental Design

In order to counter the effect on random initialisation and shuffling, we developed a framework to train the same network with the same initialisation and batches by setting the same random seed for each loss for a given experiment. With that setup, we run 5 different 5-fold cross-validation experiments with the same training/testing splits. We then evaluate the models at the end of each epoch for a maximum of 25 epochs by analysing the Dice similarity coefficient ($DSC = 2 * \frac{TP}{2*TP+FN+FP}$); sensitivity, also known as true positive fraction or TP_f for short ($TP_f = \frac{TP}{TP+FN}$); and precision ($P = \frac{TP}{TP+FP}$). We compute all three metrics for the training patches, and the training and testing image segmentations reconstructed from patches. In brief, patches of $32 \times 32 \times 32$ are uniformly sampled for each image with a sliding window of $16 \times 16 \times 16$, and the final prediction for each voxel is averaged between all the overlapping results. For the binary metrics, we first binarise the predicted output of the network (with a threshold of 0.5). The goal of this setup is

to analyse whether the difference in different executions are actually due to the training variance caused by randomisation or the losses themselves.

2.5. Implementation Details

The framework was implemented using Python 3.6.9 and pytorch version 1.5.0 and the code is publicly available at https://github.com/marianocabezas/rethinking_dsc (accessed on 1 January 2024). The whole analysis was run on a NVIDIA DGX-1 server with 80 CPU cores, 504 GB of RAM and two Tesla V100-SXM2.

3. Results

To analyse the performance of the different trained models for each epoch and seed, we used the TP_f and precision to illustrate the balance between over- and undersegmentation and the *DSC* to highlight the overall segmentation. Ideally, we would like to maximise all metrics, but in practise, increasing the TP_f also increases the false positives, affecting the precision and *DSC* values. Table 2 summarises the best performance when the best epoch and random seed are known, while Figure 3 illustrates the curves for all folds and seeds for each dataset in terms of each of the three different metrics we analysed (plots for each fold are detailed in Figures 4 and 5). The curves are separated by the type of data being evaluated (input patches and whole image segmentations reconstructed from patch segmentations) and represent a summarised version of all the metrics for each different training seed setup. The mean metric value is represented with a solid coloured line, while error bands around that line are bounded by the minimum and maximum value per epoch for all the seeds. By using upper and lower bounds, we can represent stochastic variability that could lead to different conclusions if different seeds are used for each method. Furthermore, the error bands also illustrate how robust the models can be to the problems represented by each dataset.

Table 2. Summary of the measures for the average of the best results for each fold. Patch and train refer to the patch-wise and image-wise measures for the training set, while test stands for the image-wise measures for the testing set.

Loss	*DSC*			Sensitivity (TP_f)			Precision (P)		
	Patch	**Train**	**Test**	**Patch**	**Train**	**Test**	**Patch**	**Train**	**Test**
LIT longitudinal dataset									
xent	0.64	0.51	0.46	0.62	0.55	0.49	0.54	0.48	0.46
gdsc	0.93	0.54	0.53	0.96	0.81	0.77	**0.82**	0.20	0.29
dsc	**0.96**	0.55	0.52	0.99	0.89	0.84	0.54	0.29	0.24
mixed	0.93	0.55	0.54	0.98	0.86	0.82	0.74	0.18	0.24
focal1	0.59	0.48	0.46	0.52	0.49	0.49	0.55	**0.53**	**0.49**
focal2	0.89	**0.60**	**0.56**	0.97	0.70	0.63	0.78	0.48	0.38
new	0.79	0.58	**0.56**	**1.00**	**0.94**	**0.90**	0.36	0.14	0.23
WMH challenge 2017									
xent	**0.96**	0.68	0.42	0.95	0.60	0.55	**0.97**	**0.80**	0.47
gdsc	0.86	0.71	0.69	0.85	0.64	0.63	0.88	0.79	0.76
dsc	0.88	0.56	0.54	0.87	0.55	0.65	0.89	0.60	0.60
mixed	0.89	0.56	0.44	0.89	0.58	0.63	0.90	0.62	0.47
focal1	0.94	0.71	0.68	0.91	0.64	0.59	**0.97**	**0.80**	**0.81**
focal2	**0.96**	**0.72**	**0.72**	0.98	0.72	0.69	0.94	0.73	0.75
new	0.79	0.70	0.71	**1.00**	**0.85**	**0.81**	0.65	0.61	0.63

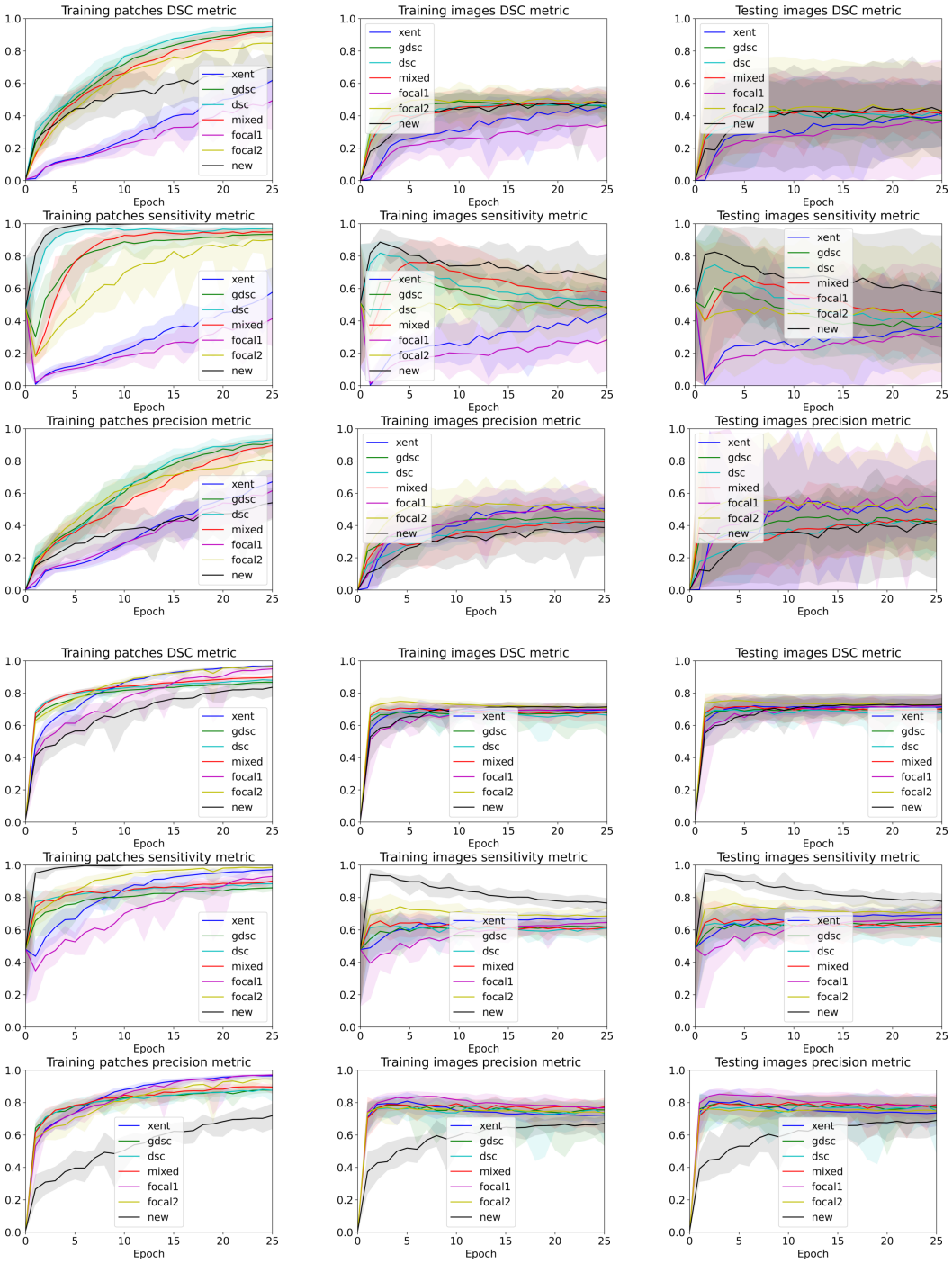

Figure 3. Band plots of the *DSC*, *sensitivity* and *precision* metric for the weights of each epoch for all the folds. The upper and lower bands represent the minimum and maximum values, while the middle line represents the mean for all random seeds and folds. The first three rows summarise the metrics for the LIT dataset, while the last three summarise the results for the WMH challenge dataset.

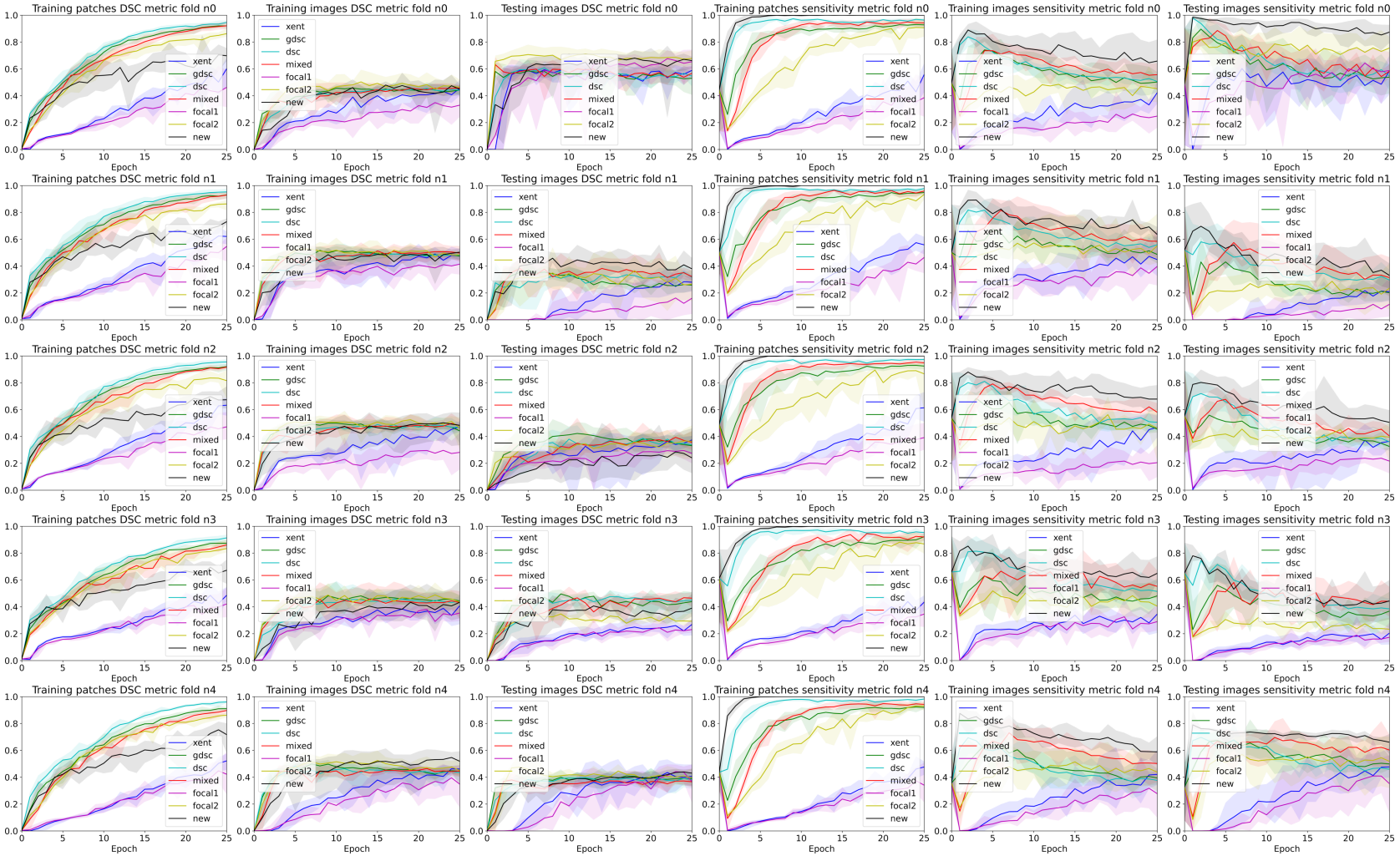

Figure 4. Band plots of the *DSC* and *sensitivity* metric for the weights of each epoch for each fold of the LIT dataset. The upper and lower bands represent the minimum and maximum values, while the middle line represents the mean for all the random seeds.

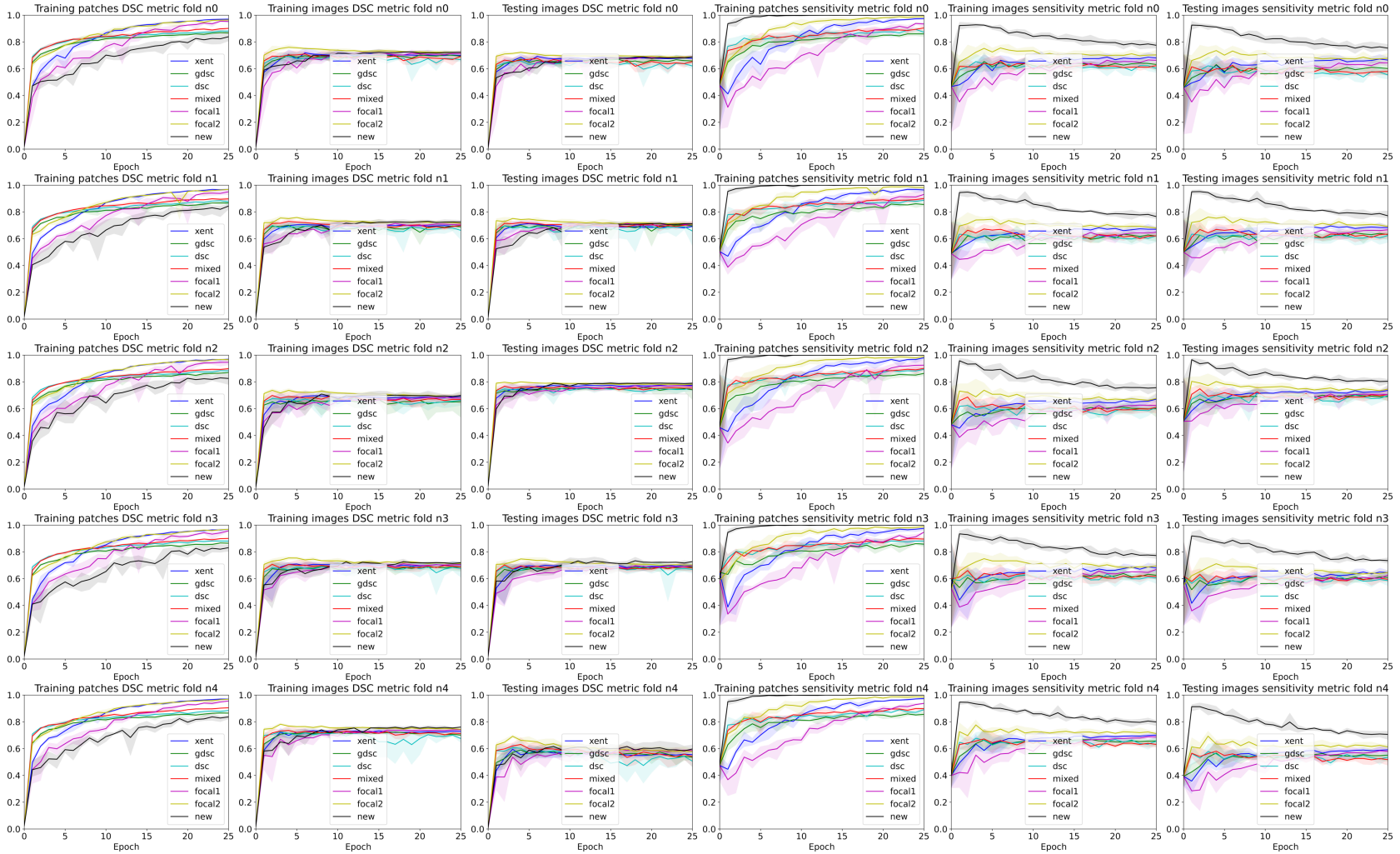

Figure 5. Band plots of the *DSC* and *sensitivity* metric for the weights of each epoch for each fold of the WMH dataset. The upper and lower bands represent the minimum and maximum values, while the middle line represents the mean for all the random seeds.

Observing the table and curves, there is a clear discrepancy between the performance on the training patches and the overall segmentation for the whole image. This difference is emphasised on the Dice-based losses. Furthermore, it is also clear that weighting the foreground class, either explicitly (focal and weighted loss) or implicitly (Dice based), improves the performance in terms of the final *DSC* metric. However, the larger the weight, the better the results. Furthermore, a high weight (higher than 0.9 for the weighted loss) has a direct impact on the TP_f measure throughout the whole training process.

A closer analysis of the curves highlights the trade-off between the *precision* and TP_f segmentation measures (also illustrated in the qualitative examples from Figures 4 and 5). As the training progresses, the TP_f metric decreases slowly, while the *precision* improves (leading to an overall slightly better *DSC* value). Furthermore, if we observe the bands for the losses, there is a clear overlap between all of them for the *DSC* metric. This suggests that there is a large variability in the model performance depending on the chosen random seed and fold. In general, the observed training pattern is similar for the weighted losses (Dice based, focal and weighted) with an initial higher peak in terms of TP_f and a slower decrease for the losses with a higher weight for the foreground class. While the behaviour is similar for both datasets, there is a bigger difference between losses for the longitudinal dataset, which also presents a higher imbalance between classes. In fact, the variability in the LIT dataset is also larger (wider error bands), leading to a larger overlap and an increased difficulty to separate between methods. Curves for the WMH dataset, on the other hand, have thinner error bands, and there is a clear separation between our proposed loss function and the rest of methods for the TP_f and *precision* metrics on the first few epochs, even though the *DSC* values remain similar for the training and testing images. Nonetheless, the curves converge to similar results as the number of epochs trained increases. Finally, the curves on Figures 4 and 5 show the same trends observed on the general dataset curves, albeit with reduced error bands (fewer samples are analysed per fold). Analysing folds carefully (each row on the figures), the LIT shows larger differences between folds. Specifically, the first row shows overall higher TP_f and *DSC* values for all methods when compared to the other folds.

To complement the analysis of the curves, qualitative results for epoch 25 (the last epoch) are provided in Figures 6 and 7 for the LIT and WMH dataset, respectively. While there seems to be a general agreement in terms of true positives (green colour), most losses and examples show an increase in false negatives (blue colour) when compared to the new proposal (as also evidenced by the numerical results). In particular, small lesions tend to be missed by losses that do not heavily weight positive predictions,with cross-entropy having the lowest detection. This is contrasted by a decrease in the number of false positives (red colour), also evidenced by the *precision* metric on the numerical results. From the Dice-based losses, mixed has the overall best results, with the dsc loss having a large number of false positives (especially on the LIT dataset). Comparing the two focal losses, the results are fairly similar with small differences in detection depending on the case, suggesting that the value of α might not be as important as expected.

Finally, we summarise the *DSC*, TP_f and *precision* metrics on the training and testing images using a violin plot in Figure 8 to analyse the distribution of these metrics for each loss according to the random seeds. Once again, we can observe that the best performance (highest bound) is obtained with the heavily weighted gradient loss (new), even though the plots present a similar shape for *DSC* in both datasets for most of the losses and a large variability in general. However, we can also observe some important differences on the TP_f distribution for both datasets. On the longitudinal dataset, the focal losses and the cross-entropy present a larger concentration of lower TP_f values when compared to the other losses, while only the lightly weighted focal loss (focal1) and cross-entropy present a lower performance on the WMH dataset. These two losses also present the worst overall performance on the longitudinal dataset in terms of *DSC* (as illustrated by a longer and spread plot, concentrated on the lower bound).

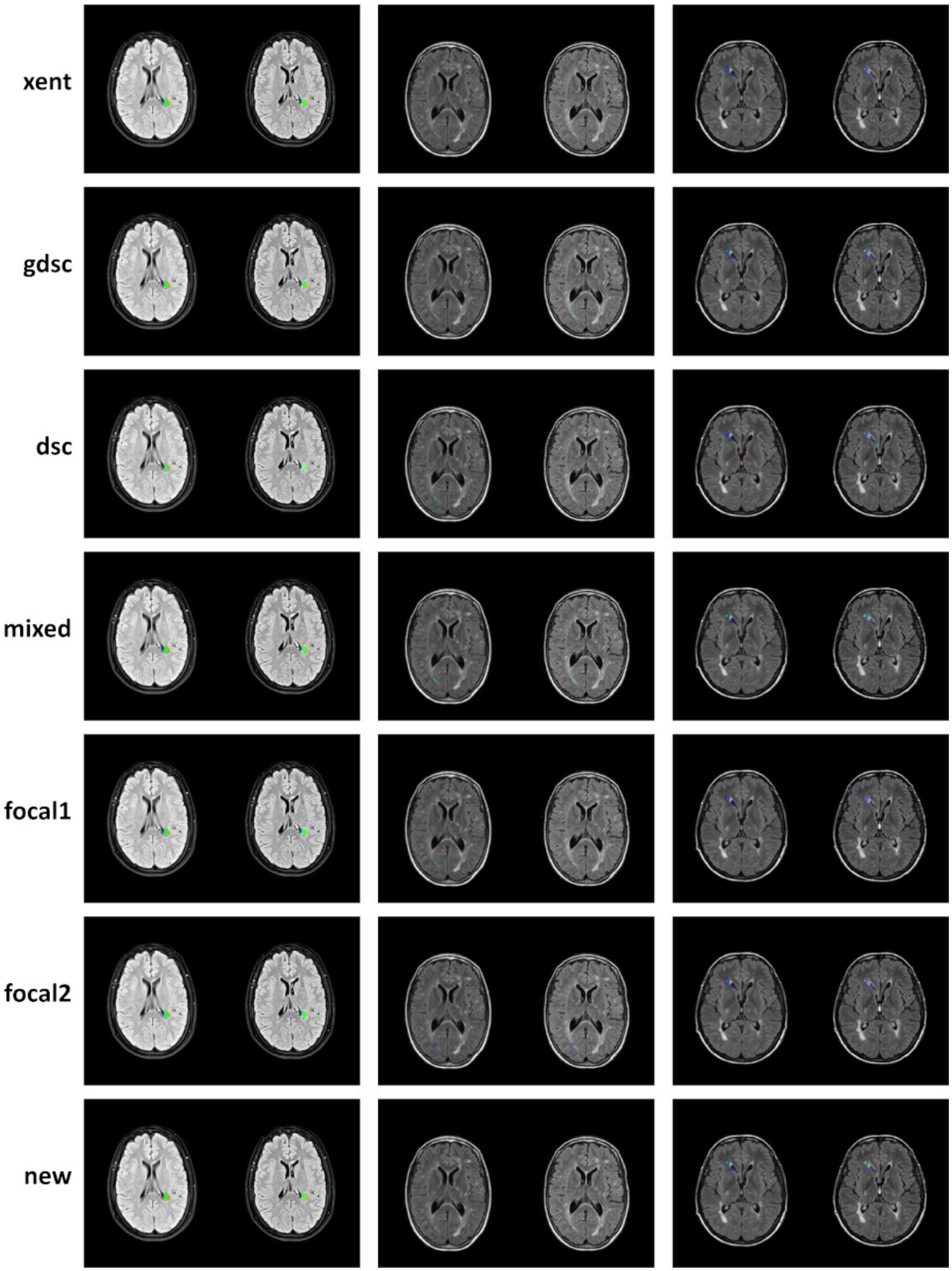

Figure 6. Qualitative examples of three subject from the LIT dataset where new lesions are segmented (those that appear only on the follow-up image on the right). Classified voxel are colour coded to represent true positives (green), false positives (red) and false negatives (blue).

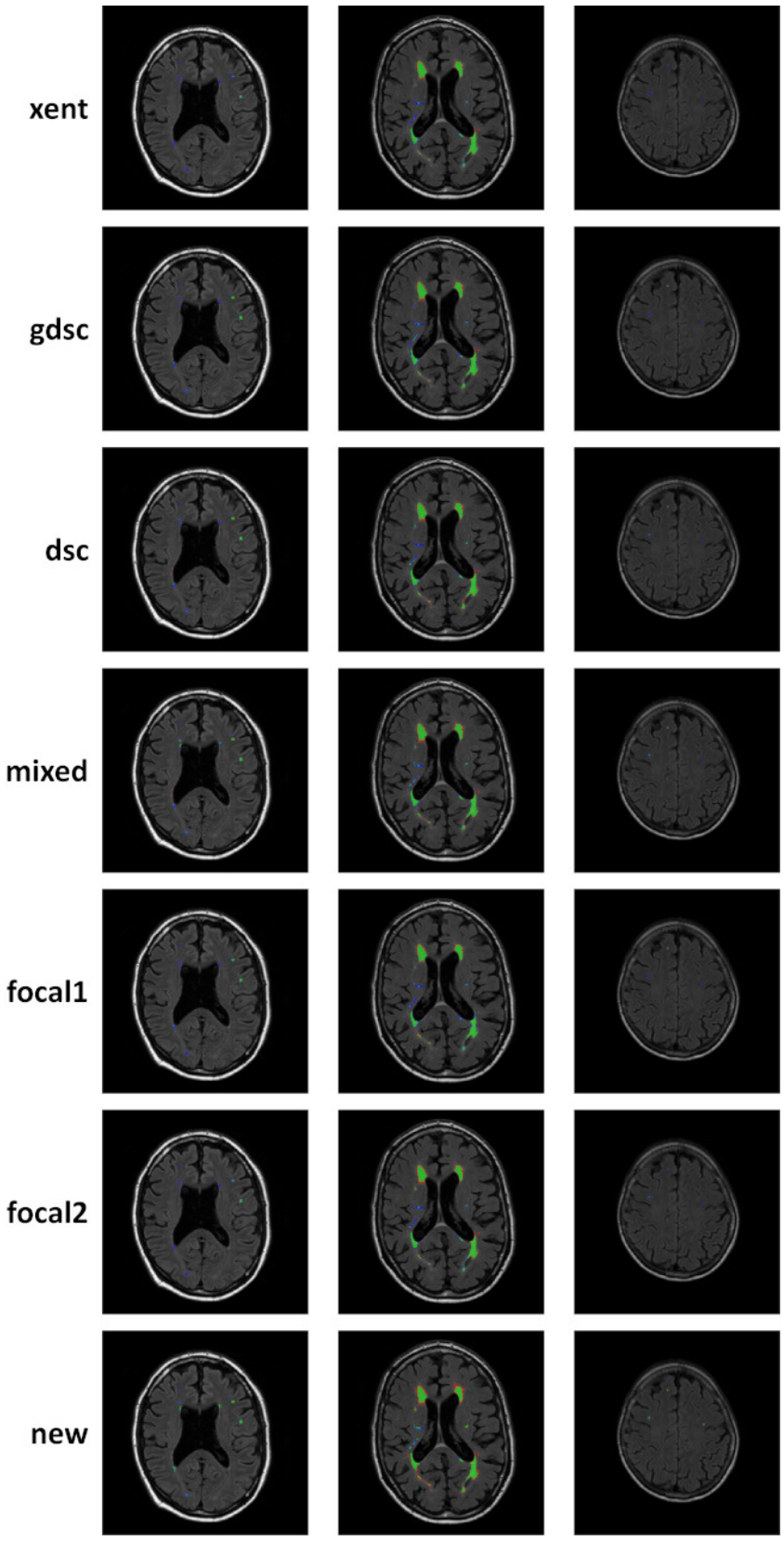

Figure 7. Qualitative examples of three subject from the WMH dataset, where white matter hyperintensities are segmented. Classified voxel are colour coded to represent true positives (green), false positives (red) and false negatives (blue).

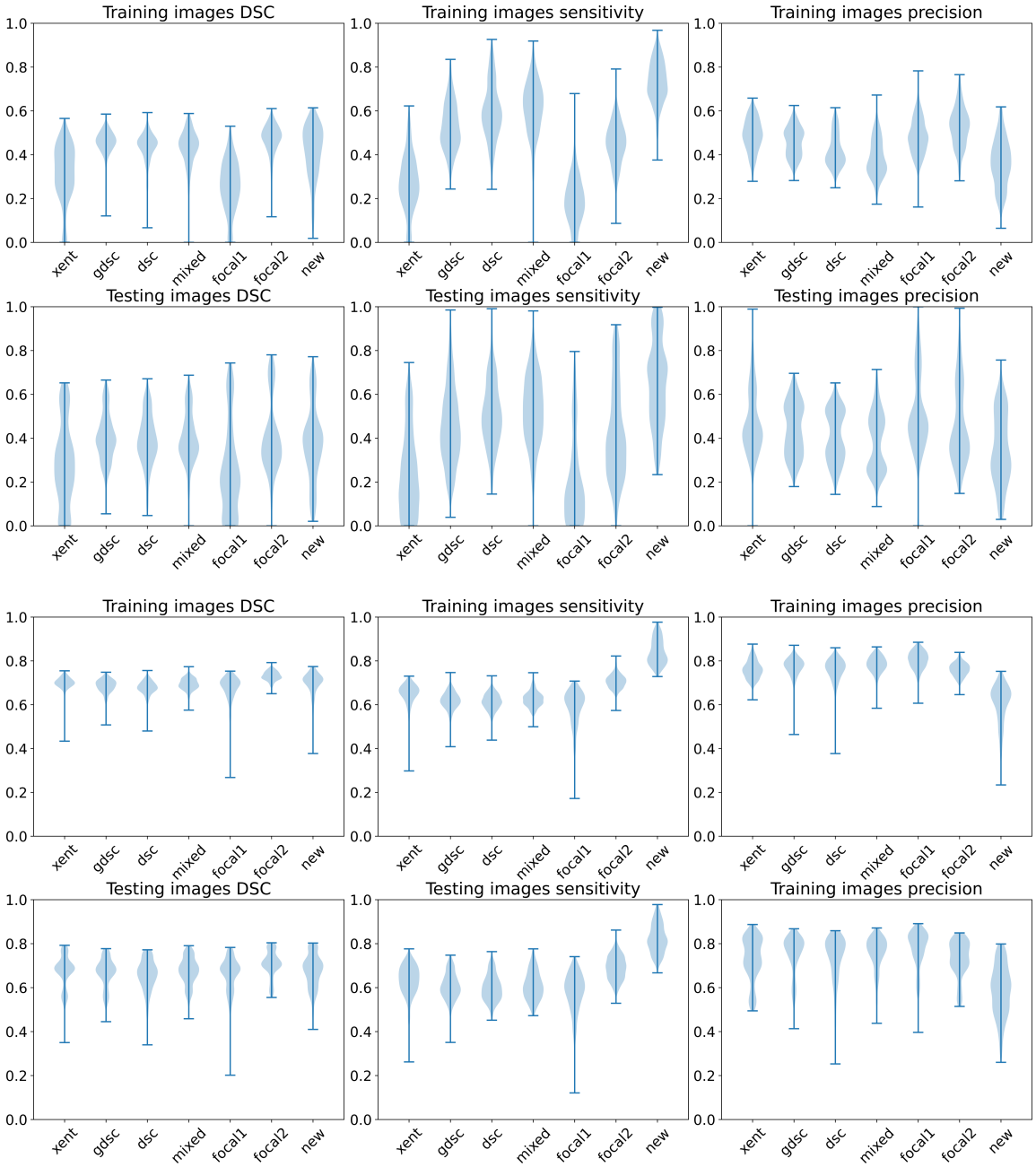

Figure 8. Violin plots for the *DSC*, *sensitivity* and *precision* metrics for all the folds and training seeds. The first row represents the results on the LIT dataset, while the second row represents the results on the WMH dataset.

4. Discussion

4.1. The Effect of Confident Errors on the Loss Function

The Dice loss is one of the most commonly used techniques to address class imbalance due to its application as a metric for segmentation [23]. Furthermore, it is usually believed that it is one of the best choices for heavily imbalanced datasets. However, the original metric was not developed for optimisation and was designed to address a set theory problems (specifically, to measure the overlap between two sets). As a consequence, there are certain unexpected side effects when pairing it with common activations for segmentation (sigmoid and softmax). One of these unexpected side effects is the effect of confidently wrong predictions (for a mathematical analysis through derivation, we refer the reader to Appendix A). This effect is evidenced by the metric curves for the Dice loss where

we found the largest discrepancy between patch-based metrics (the ones during training) and image-based metrics (the final desired output). On the first step, the Dice loss favours the prediction of positive samples, leading to a high true TP_f at the cost of a large number of false positive predictions (as evidenced by the low DSC value). Some of these erroneous predictions maintain a high confidence score throughout the whole process, leading to false positive predictions that are never corrected even with a large number of epochs. These findings align with another recent study where the effect of different loss functions on the logits of the network were analysed [9] and extends to the other two related losses (mixed and gdsc).

If we zoom out and analyse the other losses we explored in this study, we can observe that they can be easily derived from the cross-entropy loss. This loss in particular is ubiquitous in classification studies (including segmentation defined as pixel classification) due to its simplicity and its relationship to logistic regression. As opposed to the Dice loss, this function was derived to interact with sigmoid (and softmax in problems with multiple labels), leading to a linear gradient. In that sense, if we observe the curves for these losses (xent, focal1, focal2 and new), we can observe a smoother transition and increasing tendency when it comes to sensitivity over number of epochs and a higher mean precision. This suggests that false confident predictions are improved over time, as the weight updates are mostly linear with respect to the error on the prediction. Furthermore, the weights on the target class (lesions) are the most important trade-off between sensitivity and precision as observed in the differences between the focal1 and focal2 losses. Finally, even though cross-entropy is the most basic function, it is also highly reliable and can obtain similar results, even though it might have a higher number of epochs.

Overall, the positive detection of lesions voxels remains high after 25 epochs as evidenced by the qualitative examples in Figures 6 and 7. However, small differences in false positives and negatives are observed depending on the loss, with a common trade-off between them.

4.2. The Effect of Randomness during Training

To understand the effect of randomness during training, we use five random sets that affect the initial value of the layer's weights and the shuffling of the patches used for training at each epoch. To illustrate this effect, we plot the results of each epoch with a set of bands with the minimum, maximum and mean values over all seeds for each method in Section 3. One of the first conclusions that can be observed is the large overlap between all methods and metrics for image-wise results. As expected, this effect is made stronger on the testing images that were not seen during the training process. In that sense, while the mean curves can be easily distinguished between methods, it is also obvious that changing the random seed could lead to different conclusions if the lowest bound for the "baseline" methods is selected and the highest bound "seed" is used for a method we would like to highlight over the others. Furthermore, datasets with a low percentage of foreground voxels are more sensitive to that issue as exemplified by the curves on the LIT dataset shown in Figure 4.

4.3. The Discrepancy between Patch-Based Results and Image-Based Results

The most common way to analyse how the training process evolves is to monitor the training and validation losses. In the scenario where a model is trained on patches (due to memory constraints), these two curves represent the ability of the model to segment patches. Even though such measures are generally considered a good proxy for the real end goal (segmenting lesions on the whole brain), our experiments suggest that these values are not only over-optimistic but also highly misleading (especially when comparing methods). If we look at the patch-based results for both datasets, all methods have high *DSC* values over 0.7 (considered a good value for lesions) [24], with some losses reaching a value close to 1 (a perfect score). While this is not surprising because the model is trained with these same patches, the same metric drops drastically for the LIT dataset (the one with the lowest

percentage of lesion voxels) and less so for the WMH dataset. Furthermore, our proposed loss function that obtains the lowest *DSC* on the training patches by a large margin obtains comparable results when evaluated on the image segmentation results reconstructed from patches as shown in Figure 3.

On a related note, while it might seem counterintuitive that the sensitivity is also affected when comparing results between patches and images, this phenomenon might be caused by the reconstruction of the final results through averaging. In particular, boundary lesion voxels might have a decreased score when considering different predictions for the same voxel in different relative positions within the patch. In fact, previous studies on CNNs have proven that while convolutions are inherently shift-invariant, pooling and padding can lead to encoding positional information during training [25,26]. In that sense, both reconstructing image segmentations from patches or performing inference directly on the whole image would lead to different results when comparing metrics on images and patches. Moreover, if overlapping patches are used during training, the loss metric might be unrealistic image-wise, as some voxels would be counted as independent occurrences more than once.

To conclude, these results suggest a few possible explanations that have also been corroborated by previous studies focusing on the *DSC* metric and its role as a measure of overlap that provides a balanced overview in terms of true and false predictions. On one hand, it is well known that the *DSC* metric is sensitive to the size of the object being analysed [23,27]. In our results, this is reflected by the difference between patch-wise and image-wise results but also by the fact that LIT results are a decimal point lower than those obtained with the WMH dataset. On the other hand, patch-wise results focus exclusively on patches with both foreground and background voxels (as background-only patches would not contribute to improve lesion segmentation and could be detrimental). In that sense, our analysis further emphasises the importance of not relying on loss values alone when evaluating model performance and to validate with metrics that are as close as possible to the desired end goal.

5. Conclusions

In this paper, we presented an evaluation of a set of common losses used for binary lesion segmentation on heavily imbalanced scenarios. We observed that training with sampled patches from images gives an unrealistic measure of the final model when applied to whole images. In other words, while some losses might provide a better performance on the training patches, that improvement is not reflected on the training images as a whole, due to the bias introduced during sampling. Furthermore, we have observed that random initialisation and shuffling can cause a large variance on the performance metrics, giving a range of different conclusions depending on which random seed and stopping epoch we take to evaluate the losses. However, most of the analysed losses have similarly optimal performance results (if the best stopping epoch and seed are known), even though they show clearly different trends throughout the training process depending on what the loss prioritises (e.g., an increase in positive predictions for our proposal). To conclude, heavily weighting the gradient of the foreground class gives the best results in terms of true positives, while still being able to obtain a competitive DSC metric on average, with an acceptable number of false positives.

Author Contributions: Conceptualisation, M.C. and Y.D.; methodology, M.C. and Y.D.; software, M.C. and Y.D.; validation, M.C. and Y.D.; formal analysis, M.C. and Y.D.; investigation, M.C. and Y.D.; resources, M.C. and Y.D.; data curation, M.C. and Y.D.; writing—original draft preparation, M.C. and Y.D.; writing—review and editing, M.C. and Y.D.; visualisation, M.C. and Y.D.; supervision, M.C. and Y.D.; project administration, M.C. and Y.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to using anonymised publicly available data.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data for this study are publicly available on the websites of the challenge organisers. The code used for this study is also publicly available at https://github.com/marianocabezas/rethinking_dsc (accessed on 1 January of 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Mathematical Analysis of the Dice Loss

Appendix A.1. The Dice Similarity Coefficient for Binary Masks

The Dice similarity coefficient was presented by Dice in 1945 to quantify the association between ecological species [22]. Generally speaking, this metric evaluates the overlap between two binary sets. Formally, given a set $A \in \{0,1\}^N$ and a set $B \in \{0,1\}^N$, this coefficient is computed with the following equation:

$$DSC = \frac{2 \cdot |A \cap B|}{|A| + |B|}, \tag{A1}$$

where N is the number of elements in both sets. By definition, this metric is bounded between 0 and 1, with 1 representing perfect overlap and 0 no overlap between the sets. Furthermore, this metric is undefined when $|A| = |B| = 0$.

This metric has become the de facto standard in medical imaging to evaluate the automatic segmentation of a tissue, lesion or structure given a previous manual or semiautomatic annotation considering the ground truth. In that sense, the binary masks from the automatic and manual segmentation can be interpreted as two sets over the voxels of the image, where 1 represents the presence of a tissue and 0 the lack of it.

Appendix A.2. The Probabilistic Dice Function

Given the previous definition of the DSC metric and the domain of A and B, Equation (A1) is equivalent to the following one:

$$pDSC = \frac{2 \cdot \sum_i^N min(a_i, b_i)}{\sum_i^N (a_i + b_i)}, \tag{A2}$$

where $A = \{a_0, ..., a_n\}$ and $B = \{b_0, ..., b_n\}$ and $a_i, b_i \in \{0,1\}$. This new interpretation of the DSC metric can be used as an extension to fuzzy sets where now $a_i, b_i \in [0,1]$. In our case, only A is a fuzzy domain, and we can consider its values as probabilities that represent the confidence of the binary segmentation. Thus, we can use the following equation which is equivalent to (A2):

$$pDSC = \frac{2 \cdot \sum_i^N a_i \cdot b_i}{\sum_i^N (a_i + b_i)}. \tag{A3}$$

This metric is still bounded between 0 and 1, with 1 still representing perfect overlap and 0 representing no overlap. We can easily generalise that metric to a multi-class generalised coefficient (gDSC) following Crum et al.'s work [28] as follows:

$$gDSC = 2 \cdot \frac{\sum_k^K w_k \cdot \sum_i^N a_i \cdot b_i}{\sum_k^K w_k \cdot \sum_i^N (a_i + b_i)}, \tag{A4}$$

where K is the total number of classes and w_k is the weight for class k. Furthermore, if we consider binary segmentation to be a two-class segmentation problem, we can simplify that equation to:

$$gDSC = 2 \cdot \frac{(w_1 + w_0) \cdot \sum_i^N a_i \cdot b_i + w_0 \cdot (N - \sum_i^N (a_i + b_i))}{(w_1 - w_0) \cdot \sum_i^N (a_i + b_i) + 2 \cdot w_0 N}, \tag{A5}$$

where w_0 and w_1 are the weights for the background and foreground class, respectively. The pDSC (Equation (A3)) is a special case of the gDSC, where $w_0 = 0$.

Appendix A.3. The Dice Loss

Deep learning methods are currently the state of the art on medical image segmentation. Letting $X = \{x_0, ..., x_N\}$ represent the input data, each voxel x_i has D features (i.e., different contrasts), $Y = \{y_0, ..., y_N\}$ represents the labels of that patch and $y_i \in \{0, 1\}$ is the real binary value of a given voxel, representing whether it belongs to the foreground or background class. Finally, we can define a deep learning model as:

$$\hat{Y} = \sigma(f_i(X, \Theta)), \tag{A6}$$

where σ is the sigmoid function that provides the final estimate of whether the voxel belongs to the foreground and $f_i(X, \Theta)$ is the output i of the network according to its weights Θ and input X.

These models rely on gradient descent techniques to update their weights according to a loss function that we want to minimise. A common loss for imbalanced datasets (i.e., small lesions inside the brain) is the so-called DSC loss. Given Equation (A5) for the gDSC, the generalised DSC loss presented by Sudre et al. [12] can be defined as:

$$\mathcal{L}_{DSC}(\hat{Y}, Y) = 1 - 2 \cdot \frac{(w_1 + w_0) \cdot \sum_i^N \hat{y}_i \cdot y_i + w_0 \cdot (N - \sum_i^N (\hat{y}_i + y_i)}{(w_1 - w_0) \cdot \sum_i^N (\hat{y}_i + y_i) + 2 \cdot N w_0}, \tag{A7}$$

where $\hat{Y}$ is the estimated segmentation from Equation (A6) and Y is the real segmentation. By definition, the gDSC is maximal when there is a perfect overlap; thus, we need to negate the original equation for a minimisation approach.

Following the chain rule, the derivative of Equation (A7) with respect to the network parameters Θ is given by:

$$\frac{\partial \mathcal{L}_{DSC}(\hat{Y}, Y)}{\partial \Theta} = \frac{\partial \mathcal{L}_{DSC}(\hat{Y}, Y)}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial \Theta} = \sum_i^N \frac{\partial \mathcal{L}_{DSC}(\hat{Y}, Y)}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial \Theta}. \tag{A8}$$

Since we are only interested in how the loss affects the gradient of any network, we will only analyse formally its derivative with respect to $\hat{y}_i$:

$$\mathcal{F}(\hat{Y}) = (w_1 + w_0) \cdot \sum_i^N \hat{y}_i \cdot y_i + w_0 \cdot (N - \sum_i^N (\hat{y}_i + y_i)) \tag{A9}$$

$$\mathcal{G}(\hat{Y}) = (w_1 - w_0) \cdot \sum_i^N (\hat{y}_i + y_i) + 2 \cdot N w_0 \tag{A10}$$

$$\mathcal{L}_{DSC}(\hat{Y}, Y) = 1 - 2 \cdot \mathcal{F}(\hat{Y}) / \mathcal{G}(\hat{Y}) \tag{A11}$$

$$\frac{\partial \mathcal{L}_{DSC}(\hat{Y}, Y)}{\partial \hat{y}_i} = -2 \cdot \frac{\frac{\partial \mathcal{F}(\hat{Y})}{\partial \hat{y}_i} * \mathcal{G}(\hat{Y}) - \frac{\partial \mathcal{G}(\hat{Y})}{\partial \hat{y}_i} * \mathcal{F}(\hat{Y})}{\mathcal{G}(\hat{Y})^2} \tag{A12}$$

$$\begin{aligned}\frac{\partial \mathcal{F}(\hat{Y})}{\partial \hat{y}_i} * \mathcal{G}(\hat{Y}) = \\ ((w_1 + w_0)y_i - w_0)((w_1 - w_0) \cdot \sum_j^N (\hat{y}_j + y_j) + 2 \cdot N w_0) = \\ (w_1^2 - w_0^2) \cdot y_i \cdot \sum_j^N (\hat{y}_j + y_j) + \\ 2 \cdot N w_0 \cdot (w_1 + w_0) \cdot y_i - \\ w0 \cdot (w_1 - w_0) \cdot \sum_j^N (\hat{y}_j + \ y_j) - \\ 2 \cdot N w_0^2\end{aligned} \tag{A13}$$

$$\begin{aligned}\frac{\partial \mathcal{G}(\hat{Y})}{\partial \hat{y}_i} * \mathcal{F}(\hat{Y}) = -(w_1 - w_0)((w_1 + w_0) \cdot \sum_j^N \hat{y}_j y_j + w_0 \cdot (N - \sum_j^N (\hat{y}_j + y_j))) = \\ -(w_1^2 - w_0^2) \cdot \sum_j^N \hat{y}_j y_j + \\ w_0 \cdot (w_1 - w_0) \cdot \sum_j^N (\hat{y}_j + y_j) - \\ N w_0 \cdot (w_1 - w_0)\end{aligned} \tag{A14}$$

$$\begin{aligned}\frac{\partial \mathcal{L}_{DSC}(\hat{Y}, Y)}{\partial \hat{y}_i} = \\ -2 \cdot (w_1 + w_0) \frac{(w_1 - w_0)\left[y_i \cdot \sum_j^N (\hat{y}_j + y_j) - \sum_j^N (\hat{y}_j y_j)\right] + N w_0 \cdot (2y_i - 1)}{\left[(w_1 - w_0) \cdot \sum_j^N (\hat{y}_j + y_j) + 2 \cdot N w_0\right]^2}.\end{aligned} \tag{A15}$$

This equation is independent of the prediction $\hat{y}_i$. Moreover, the denominator is the same for all possible network outputs. Therefore, from now on, we will refer to it as the normalising factor $\mathcal{N}(\hat{Y}, Y) = \frac{2 \cdot (w_1 + w_0)}{\left[(w_1 - w_0) \cdot \sum_j^N (\hat{y}_j + y_j) + 2 \cdot N w_0\right]^2}$ to simplify the analysis.

If we go back to Equation (A15), we can observe three different terms in the numerator: a global one $\frac{\partial \mathcal{G}_{DSC}}{\partial \hat{y}_i} = -(w_1 - w_0)\left[\sum_j^N (\hat{y}_j y_j)\right]$ that is applied to the gradient of all voxels, a term which only affects the foreground voxels $\frac{\partial \mathcal{F}_{DSC}}{\partial \hat{y}_i} = (w_1 - w_0)\left[\sum_j^N (\hat{y}_j + y_j)\right]$ and a mixed term that depends on the label $\frac{\partial \mathcal{M}_{DSC}}{\partial \hat{y}_i} = N w_0 \cdot (2y_i - 1)$. We will analyse each of them independently with respect to a given weight Θ, and afterwards we will group them together.

Given a true binary segmentation Y, we can define $Y^0 = \{i : y_i = 0\}$ and $Y^1 = \{i : y_i = 1\}$ as the sets of voxels from the background and foreground class, respectively. By definition, $|Y^0| + |Y^1| = N$. Now, starting with the global term, we can define the gradient with respect to the output $\hat{y}_i$ as follows:

$$\frac{\partial \mathcal{G}_{DSC}}{\partial \Theta} = -(w_1 - w_0)\left[|Y^1| \mathop{\mathbb{E}}_{i \in Y^1}[\hat{y}_i]\right]\left[|Y^0| \mathop{\mathbb{E}}_{i \in Y^0}\left[\frac{\partial \hat{y}_i}{\partial \Theta}\right] + |Y^1| \mathop{\mathbb{E}}_{i \in Y^1}\left[\frac{\partial \hat{y}_i}{\partial \Theta}\right]\right], \tag{A16}$$

where $\mathop{\mathbb{E}}_{i \in Y^0}$ represents the expectation of the background voxels and $\mathop{\mathbb{E}}_{i \in Y^1}$ the expectation of the foreground voxels. Continuing with the foreground term:

$$\frac{\partial \mathcal{F}_{DSC}}{\partial \Theta} = (w_1 - w_0)\left[|Y^1| + |Y^0| \mathop{\mathbb{E}}_{i \in Y^0}[\hat{y}_i] + |Y^1| \mathop{\mathbb{E}}_{i \in Y^1}[\hat{y}_i]\right]\left[|Y^1| \mathop{\mathbb{E}}_{i \in Y^1}\left[\frac{\partial \hat{y}_i}{\partial \Theta}\right]\right]. \tag{A17}$$

Finally, we can express the third term as:

$$\frac{\partial \mathcal{M}_{DSC}}{\partial \Theta} = N w_0\left[|Y^1| \mathop{\mathbb{E}}_{i \in Y^1}\left[\frac{\partial \hat{y}_i}{\partial \Theta}\right] - |Y^0| \mathop{\mathbb{E}}_{i \in Y^0}\left[\frac{\partial \hat{y}_i}{\partial \Theta}\right]\right]. \tag{A18}$$

If we now group all three terms together again, the final gradient with respect to Θ is given by:

$$\frac{\partial \mathcal{L}_{DSC}}{\partial \Theta} = -\mathcal{N}(\hat{Y}, Y) \cdot \left[\frac{\partial \mathcal{G}_{DSC}}{\partial \Theta} + \frac{\partial \mathcal{F}_{DSC}}{\partial \Theta} + \frac{\partial \mathcal{M}_{DSC}}{\partial \Theta}\right]. \tag{A19}$$

If we ignore the normalising factor to focus on the relationship between background and foreground voxels, we can simplify that equation as:

$$\begin{aligned} \frac{\partial \mathcal{L}_{DSC}}{\partial \Theta} \propto -(\frac{\partial \mathcal{G}_{DSC}}{\partial \Theta} + \frac{\partial \mathcal{F}_{DSC}}{\partial \Theta} + \frac{\partial \mathcal{M}_{DSC}}{\partial \Theta}) = \\ |Y^1|(w_1 - w_0)\left[|Y^0| \mathop{\mathbb{E}}_{i \in Y^1}[\hat{y}_i] \mathop{\mathbb{E}}_{i \in Y^0}\left[\frac{\partial \hat{y}_i}{\partial \Theta}\right] - (|Y^1| + |Y^0| \mathop{\mathbb{E}}_{i \in Y^0}[\hat{y}_i]) \mathop{\mathbb{E}}_{i \in Y^1}\left[\frac{\partial \hat{y}_i}{\partial \Theta}\right]\right] + \\ N w_0\left[|Y^0| \mathop{\mathbb{E}}_{i \in Y^0}\left[\frac{\partial \hat{y}_i}{\partial \Theta}\right] - |Y^1| \mathop{\mathbb{E}}_{i \in Y^1}\left[\frac{\partial \hat{y}_i}{\partial \Theta}\right]\right]. \end{aligned} \tag{A20}$$

Since $\hat{y}_i = \sigma(f_i(X, \Theta))$ as defined by Equation (A6), its derivative with respect to a given weight is given by:

$$\frac{\partial \hat{y}_i}{\partial \Theta} = \hat{y}_i(1 - \hat{y}_i)\frac{\partial f_i(X, \Theta)}{\partial \Theta}, \tag{A21}$$

and so:

$$\mathbb{E}\left[\frac{\partial \hat{y}_i}{\partial \Theta}\right] = \mathbb{E}\left[\hat{y}_i(1 - \hat{y}_i)\frac{\partial f_i(X, \Theta)}{\partial \Theta}\right]. \tag{A22}$$

Analysing that expectation, it becomes 0 when all the predictions have a value of either 0 or 1, independently of the real label. Furthermore, the maximum gradient for a voxel is given when its probability is 0.5. This is usually the case after initialisation. However, as the weights are updated and that probability moves away from 0.5, that gradient will decrease whether the voxel is well-classified or not.

Taking that into account, the gradient from Equation (A19) becomes 0 when all voxels (including background and foreground) have a prediction of 0 or 1. That means that extreme false positives and false negatives will not have an effect on updating the network parameters using gradient descent. Furthermore, in the extreme case where all the foreground voxels have been misclassified as background, they will not be taken into account while updating the network weights. In an extremely unbalanced dataset, this risk is important since the network might not be able to recover from that and miss all positive examples.

Now, if we analyse Equation (A20) with respect to the expectations of $\hat{y}_i$, we can see how $\mathop{\mathbb{E}}_{i \in Y^1}[\hat{y}_i]$ represents the true positive rate for foreground voxels, while $\mathop{\mathbb{E}}_{i \in Y^0}[\hat{y}_i]$ can be

regarded as a the rate of misclassified background voxels. In that sense, if we analyse the first term, the gradient of the background voxels increases as the true positives increase, while the gradient of the foreground voxels decreases as the false positive predictions decrease. On the other hand, the second term only depends on the number of positive and negative voxels. In an extremely unbalanced case, the gradient of the background voxels would dominate that second term since $|Y^0| \ggg |Y^1|$. Furthermore, in order for the weight of the first term ($|Y^1|(w_1 - w_0)$) to be larger than the weight of the second term (Nw_0), $w_1 > \frac{Nw_0}{|Y^1|} + w_0$.

Finally, in the special case where $w_0 = 0$, the most common use of the Dice loss originally presented by Milletari et al. [11], the second term becomes 0 and the normalising factor is simplified to $\mathcal{N}(\hat{Y}, Y) = \frac{2}{\left[\sum_j^N (\hat{y}_j + y_j)\right]^2}$. In that case, the derivative (and loss function) are undefined when all predictions are set to 0 and all the voxels belong to the background. Furthermore, the issues from the first term and the derivative of the sigmoid still apply to that function.

References

1. Lesjak, Ž.; Pernuš, F.; Likar, B.; Špiclin, Ž. Validation of White-Matter Lesion Change Detection Methods on a Novel Publicly Available MRI Image Database. *Neuroinformatics* **2016**, *1*, 403–420. [CrossRef] [PubMed]
2. Commowick, O.; Istace, A.; Kain, M.; Laurent, B.; Leray, F.; Simon, M.; Pop, S.C.; Girard, P.; Améli, R.; Ferré, J.C.; et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Nat. Sci. Rep.* **2018**, *8*, 13650. [CrossRef]
3. Vanderbecq, Q.; Xu, E.; Ströer, S.; Couvy-Duchesne, B.; Diaz Melo, M.; Dormont, D.; Colliot, O. Comparison and validation of seven white matter hyperintensities segmentation software in elderly patients. *Neuroimage Clin.* **2020**, *27*, 102357. [CrossRef] [PubMed]
4. Kuijf, H.; Biesbroek, J.; Bresser, J.; Heinen, R.; Andermatt, S.; Bento, M.; Berseth, M.; Belyaev, M.; Cardoso, M.; Casamitjana, A.; et al. Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge. *IEEE Trans. Med. Imag.* **2019**, *38*, 2556–2568. [CrossRef] [PubMed]
5. Zhou, S.K.; Greenspan, H.; Davatzikos, C.; Duncan, J.S.; Van Ginneken, B.; Madabhushi, A.; Prince, J.L.; Rueckert, D.; Summers, R.M. A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies with Progress Highlights, and Future Promises. *Proc. IEEE* **2021**, *109*, 820–838. [CrossRef] [PubMed]
6. Bernal, J.; Kushibar, K.; Asfaw, D.S.; Valverde, S.; Oliver, A.; Martí, R.; Lladó, X. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: A review. *Artif. Intell. Med.* **2019**, *95*, 64–81. [CrossRef]
7. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef] [PubMed]
8. Valverde, S.; Cabezas, M.; Roura, E.; González-Villà, S.; Pareto, D.; Vilanova, J.C.; Ramió-Torrentà, L.; Rovira, A.; Oliver, A.; Lladó, X. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *Neuroimage* **2017**, *155*, 159–168. [CrossRef] [PubMed]
9. Li, Z.; Kamnitsas, K.; Glocker, B. Analyzing Overfitting under Class Imbalance in Neural Networks for Image Segmentation. *IEEE Trans. Med. Imag.* **2021**, *40*, 1065–1077. [CrossRef]
10. Valverde, S.; Salem, M.; Cabezas, M.; Pareto, D.; Vilanova, J.C.; Ramió-Torrentà, L.; Rovira, À.; Salvi, J.; Oliver, A.; Lladó, X. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *Neuroimage Clin.* **2019**, *21*, 101638. [CrossRef]
11. Milletari, F.; Navab, N.; Ahmadi, S. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
12. 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 Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support Workshop—MICCAI, Québec City, QC, Canada, 14 September 2017; pp. 240–248.
13. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
14. Ma, J.; Chen, J.; Ng, M.; Huang, R.; Li, Y.; Li, C.; Yang, X.; Martel, A.L. Loss odyssey in medical image segmentation. *Med. Image Anal.* **2021**, *71*, 102035. [CrossRef] [PubMed]
15. Kononenko, I. Bayesian neural networks. *Biol. Cybern.* **1989**, *61*, 361–370. [CrossRef]
16. Hernández-Lobato, J.M.; Adams, R. Probabilistic backpropagation for scalable learning of bayesian neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 6–11 July 2015; pp. 1861–1869.
17. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv* **2017**, arXiv:1612.01474v3.

18. Izmailov, P.; Vikram, S.; Hoffman, M.D.; Wilson, A.G.G. What Are Bayesian Neural Network Posteriors Really Like? In Proceedings of the 38th International Conference on Machine Learning, Online, 18–24 July 2021; pp. 4629–4640.
19. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, 17–21 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 424–432.
20. Cabezas, M.; Luo, Y.; Kyle, K.; Ly, L.; Wang, C.; Barnett, M. Estimating lesion activity through feature similarity: A dual path Unet approach for the MSSEG2 MICCAI challenge. In Proceedings of the MSSEG-2 Challenge Proceedings—MICCAI 2021, Strasbourg, France, 2 September 2021; pp. 107–110.
21. Barnett, M.; Wang, D.; Beadnall, H.; Bischof, A.; Brunacci, D.; Butzkueven, H.; Brown, J.W.L.; Cabezas, M.; Das, T.; Dugal, T.; et al. A real-world clinical validation for AI-based MRI monitoring in multiple sclerosis. *NPJ Digit. Med.* **2023**, *6*, 196. [CrossRef] [PubMed]
22. Dice, L.R. Measures of the amount of ecologic association between species. *Ecology* **1945**, *26*, 297–302. [CrossRef]
23. Ostmeier, S.; Axelrod, B.; Isensee, F.; Bertels, J.; Mlynash, M.; Christensen, S.; Lansberg, M.G.; Albers, G.W.; Sheth, R.; Verhaaren, B.F.; et al. USE-Evaluator: Performance metrics for medical image segmentation models supervised by uncertain, small or empty reference annotations in neuroimaging. *Med. Image Anal.* **2023**, *90*, 102927. [CrossRef] [PubMed]
24. Ma, Y.; Zhang, C.; Cabezas, M.; Song, Y.; Tang, Z.; Liu, D.; Cai, W.; Barnett, M.; Wang, C. Multiple Sclerosis Lesion Analysis in Brain Magnetic Resonance Images: Techniques and Clinical Applications. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 2680–2692. [CrossRef] [PubMed]
25. Kayhan, O.S.; Gemert, J.C.v. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14274–14285.
26. Islam, M.A.; Kowal, M.; Jia, S.; Derpanis, K.G.; Bruce, N.D. Global pooling, more than meets the eye: Position information is encoded channel-wise in CNNs. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 793–801.
27. Carass, A.; Roy, S.; Gherman, A.; Reinhold, J.C.; Jesson, A.; Arbel, T.; Maier, O.; Handels, H.; Ghafoorian, M.; Platel, B.; et al. Evaluating White Matter Lesion Segmentations with Refined Sørensen-Dice Analysis. *Sci. Rep.* **2020**, *10*, 8242. [CrossRef] [PubMed]
28. Crum, W.; Camara, O.; Hill, D. Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis. *IEEE Trans. Med. Imaging* **2006**, *25*, 1451–1461. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.