A size-adaptative segmentation method for better detection of multiple sclerosis lesions

Alexandre Fenneteau, David Helbert, Pascal Bourdon, Imane M'Rabet, Christine Fernandez-Maloigne, Rémy Guillevin

▶ **To cite this version:**

Alexandre Fenneteau, David Helbert, Pascal Bourdon, Imane M'Rabet, Christine Fernandez-Maloigne, et al.. A size-adaptative segmentation method for better detection of multiple sclerosis lesions. 2022. hal-03836787

HAL Id: hal-03836787
https://hal.science/hal-03836787

Preprint submitted on 31 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A size-adaptative segmentation method for better detection of multiple sclerosis lesions

Alexandre Fenneteau[1,2,3], David Helbert[2,3,*], Pascal Bourdon[2,3], Imane M'rabet[3,5], Christine Fernandez-Maloigne[2,3], Rémy Guillevin[3,4,5]
[1]Siemens Healthineers, Saint Denis, France
[2]University of Poitiers, CNRS, XLIM, Poitiers, France
[3]I3M, Common Laboratory CNRS-Siemens, University and Hospital of Poitiers, Poitiers, France
[4]University of Poitiers, CNRS, LMA, Poitiers, France
[5]Poitiers University Hospital, CHU, Poitiers, France

Abstract.
Purpose: The automatic segmentation of multiple sclerosis lesions on magnetic resonance images is an open research task aiming to bring more reproducibility in the radiological visual assessment of the disease while reducing the burden of this time-consuming task. The development of artificial intelligence has led to significant improvements in computer-aided diagnosis tools for radiology. It exists several approaches for the segmentation of multiple sclerosis lesions using convolutional neural networks. However, the small lesions are frequently neglected by those algorithms despite their importance. We propose here an adaptable method to improve the detection of small lesions.
Approach: The problem of small lesions detection mainly comes from the under-representation of those lesions at a voxel level and the segmentation loss function. The presented method consists in weighting the lesion importance during the training of a convolutional neural network depending on lesion size to correct the impact of voxel lesion imbalances.
Results: With our method, the lesion segmentation computed with the Dice score is only slightly improved but the detection sensitivity is significantly improved at the cost of a limited augmentation of lesion false positive rate. The F1 score has been substantially improved with the correct set of parameters. The improved prediction quality of segmentation maps has been confirmed visually with the help of a radiologist.
Conclusions: The described method improves the lesion detection by giving more importance to small lesions during the multiple sclerosis lesions segmentation learning, bringing a better help for radiologists towards a better impact for the patient care.

Keywords: Artificial intelligence, multiple sclerosis, CNN, small lesions, unbalanced segmentation, detection.

*David Helbert, david.helbert@univ-poitiers.fr

1 Introduction

Multiple Sclerosis (MS) is a chronic and autoimmune disease affecting the central nervous system that causes disability and cognitive impairment. The disease has been estimated to touch more than 2,000,000 people worldwide in 2016 according to Wallin *et al.*[1] and is considered as the commonest non-traumatic disabling disease to affect young adults.[2] The disease causes inflammatory demyelination lesions in the brain and spinal cord that lead to axonal degeneration and eventually neuron death as described by Baecher *et al.*[3]

The diagnosis of the disease is based on the revised McDonald criteria[4] in which the Magnetic Resonance (MR) exam plays an important role by allowing the radiologist to assess the dissemination of the lesions in space and time. Furthermore, the MR exam is particularly used for the follow-up of the disease to spot possible new and enlarging lesions. Combined with the clinical state evaluated by the EDSS score, the MRI follow-up assess response to treatment, allowing therapeutic adjustment if necessary.[5]

The MR T2 sequence is set as the reference sequence to detect MS lesions as hyper-intensities,[4] but T2-fluid-attenuated inversion recovery (FLAIR) images are more used in practice since it suppresses the hyper-intensities caused by free water while keeping lesion-related hypersignal.

MS lesions screening is performed by radiologists by comparing visually successive MR images and is considered as a daily repetitive and time-consuming task. Moreover, this implies an obvious part of subjectivity, essentially depending on the level of expertise of the radiologist. Detection has been associated with high inter- and intraobserver variability.[6] A performant automatic segmentation tool for MS lesions would reduce valuable analysis time, bring more reproducibility and give more metrics previously inaccessible.

A large number of methods for automatic segmentation of MS lesions have been proposed in the last two decades. These methods are based on statistical models, atlases, machine learning models and more recently by making use of deep learning as described by Danelakis *et al.*[7] Deep learning has been extensively used for this segmentation problem for its performances with more than a hundred articles referenced by Shoeibi *et al.*[8] Other methods, considered less efficient, are gradually being abandoned.

It has been pointed out by Kaur *et al.*[9] that one of the challenges for MS lesion segmentation is the variability of lesion size. In particular small lesions are generally less detected or less

segmented as shown by Vang *et al.*,[10] Coronado *et al.*[11] and Valverde *et al.*[12] This means that small lesions are less segmented and detected than bigger ones because of lesion heterogeneity in representation, size and appearance. However, the problem of small lesion detection is seldom mentioned or addressed despite its impact on the lesion detection and the clinical value of those lesions. Among the top-performing methods, on the ISBI MS segmentation challenge[13] leaderboard[1] the lesion true positive rate (LTPR) or detection sensitivity is rarely above 55% for best methods with a low lesion false positive rate (LFPR) often lower than 20%. It means that, in general, models learn to favor detecting correct lesions than detecting a lot of them.

However, for the radiologist, it is more important to spot a maximum of MS lesions for computer-aided detection (CAD) tools and especially the smallest ones. As a matter of fact, big lesions are easy to see compared to small lesions that can be missed, especially when the lesion load is high. In addition, for the detection of new lesions on consecutive MR exam, new lesions are often small and difficult to spot despite their value for the diagnostic and for the treatment adjustment. It is also simpler to discard a falsely detected lesion than to spot a not detected one when using CAD.

A method for a better lesion detection sensitivity and particularly for small lesions is of interest for clinical use. To the best of our knowledge, the lesion size variability problem is taking into account during the training of a deep learning model for MS lesion segmentation only by Zhang *et al.*[14] In their method, they include a lesion-wise module in which all lesions are considered as spheres with fixed volume centered on the lesion to segment to allow the model to learn lesion detection without lesion size influence.

First we proposed a simple method for the better detection of small lesions in MS lesion seg-

[1] available at https://smart-stats-tools.org/lesion-challenge

mentation using a Convolutional Neural Network (CNN). The method consists in weighting the loss function for learning MS lesion segmentation in order to prioritize the detection of those lesions. Then, we evaluate the weighting method on an in-house dataset and show how it can improve the lesion detection for a better radiological CAD. Finally, we evaluated our method on the ISBI-2015 dataset as an external validation of the method.

2 Materials and methods

The segmentation task consists in assigning for every pixel or voxel in 3D a class: *lesion* or *not lesion* in our case. Whereas the detection suggests that the object of interest is spatially detected without a voxel-wise constraint, and is most of the time associated with bounding boxes. We can see segmentation as a particular case of detection. Therefore, the segmentation quality has been used for white matter and MS lesion to reflect the detection though it is a wrong oversimplification of the problem as pointed out by Carass *et al.*[15] A good segmentation score can be associated with a low detection although the detection is, in fact, the most radiologically valued. As a matter of fact, it is more important to spot any potential lesions than to be absolutely correct on their real volumes and edges, though both properties are of interest. It is then important to define radiologically relevant metrics to evaluate, design and train automatic algorithms for MS lesion segmentation and detection.

2.1 Metrics

For white matter and MS lesion segmentation, the Dice score metric has become a golden-standard for validating the segmentation quality.[15] However, it is a segmentation measurement at a voxel point of view that does not necessarily reflect the lesion detection. The lesion detection is generally

evaluated with the F1 score that relies on the same principles but is computed at the lesion scale instead of the voxel scale.

Both metrics come from the Sørensen formula:

$$SD = \frac{2|X \cap Y|}{|X| + |Y|}, \tag{1}$$

where X and Y are two sets and $|\cdot|$ is the cardinality operator. In the binary context, where only two classes exist, the Sørensen-Dice can be written:

$$SD = \frac{2TP}{2TP + FP + FN} = 2\frac{PPV \times TPR}{PPV + TPR}, \tag{2}$$

where TP is the number of true positives, FP the number of false positives, FN the number of false negatives, $TPR = \frac{TP}{TP+FN}$ is the true positive rate or sensitivity and $PPV = \frac{TP}{TP+FP}$ is the positive predictive value or precision.

In this paper we adopt the definition of Dice score and F1 score as Carass *et al.*[13] The segmentation Dice score is computed at voxel level whereas the F1 score is computed at a lesion level. By considering voxel-wise previously described metrics with V prefix and L prefix the lesion-wise metrics, the Dice and F1 score can be written:

$$Dice = \frac{2VTP}{2VTP + VFP + VFN} = 2\frac{VPPV \times VTPR}{VPPV + VTPR}, \tag{3}$$

$$F1 = \frac{2LTP}{2LTP + LFP + LFN} = 2\frac{LPPV \times LTPR}{LPPV + LTPR}, \tag{4}$$

where LTP counts lesion overlap between lesion masks, LFP counts predicted lesions that are

not in the ground truth mask and LFN counts lesions in the ground truth mask that are not in the predicted mask as formalized by Carass *et al.*[13] Both metrics are the harmonic mean of precision and sensitivity but computed at a voxel scale for the Dice score and at a lesion scale for the F1 score. They measure then the similarity of two sampled as a compromise between detecting a maximum of elements in the ground truth (sensitivity) and predicting a minimum of elements that are not in the ground truth (precision). A common metric used in lesion detection is the lesion false positive rate $LFPR = \frac{LFP}{LFP+LTP} = 1 - LPPV$ as described by Zhang *et al.*[14]

2.2 Lesion size weighting

When applied to MS lesion segmentation, both Dice and F1 score are expected to be maximized to have a good lesion and voxel lesion detection. However, the implementation of a differentiable loss function based on F1 score is in practice really complex and that explains why the Dice loss is generally used.[16]

The problem of Dice loss is that it is computed at a voxel level, meaning that the lesion scale is not taken into account at all in the learning. When applied on MS lesion segmentation with different lesion sizes, this leads the model to learn with more voxel from big lesions than from voxel from small lesions since their volume is negligible compared to big ones. Since big lesions are easier to spot and the most prevalent, it conducts the trained model to avoid segmenting small lesions that do not account much in the loss for a big batch though false positive are still penalized. It is then observed in most methods that the Dice score is high, the $LFPR$ is close to 0 and the $LTPR$ close to 0.5, meaning that the model segments almost every time big lesions but does not take the risk to detect more difficult one (the smallest).

The objective of the presented method is to formalize a differentiable loss function for segmen-

tation that gets closer to the lesion detection F1 score and especially increases the lesion sensitivity or $LTPR$ for small lesions. To meet those expectation, the method aims at increasing the importance of small lesions, which are neglected by most voxelwise segmentation trained models due to the over-representation of large lesion voxels and to the loss functions.

To give more importance to small lesions, we designed a lesion voxel-weighting method depending on lesion size. The visible lesions on classical MR images vary in volume in a large range from less than $1\ mm^3$ to a few cm^3. A simple linear weighting is not relevant since the volume range is too large for such a mapping and would either not give enough importance to small lesions or to big ones. The proposed weighting method is designed:

- to be decreasing or stagnating with lesion size,
- to have a configurable decay to adjust the small and big lesions weight contrast,
- to have limited sensitivity to small size variations in the beginning nor to separate big from very big lesions from a certain threshold.

The designed weighting method is based on logistic functions for their interesting properties and "S" shape. Weighting function $\omega : \mathbb{R}^+ \mapsto \mathbb{R}^+$ is defined as:

$$\omega(v_{les}) = w_{\max} - \frac{(w_{\max} - w_{\min})}{1 + \alpha \times e^{\frac{-kv_{les}}{range}}}, \tag{5}$$

where v_{les} is the lesion volume, $w_{\max}$ is maximum weighting value, $w_{\min}$ is the minimum weighting value, α sets the x-axis translation of the curve, $range \in \mathbb{R}^+$ sets the v range between the minimum and maximum asymptotes of the curve and k defines the steepness of curve. Weighting function is illustrated in **Figure** 1.

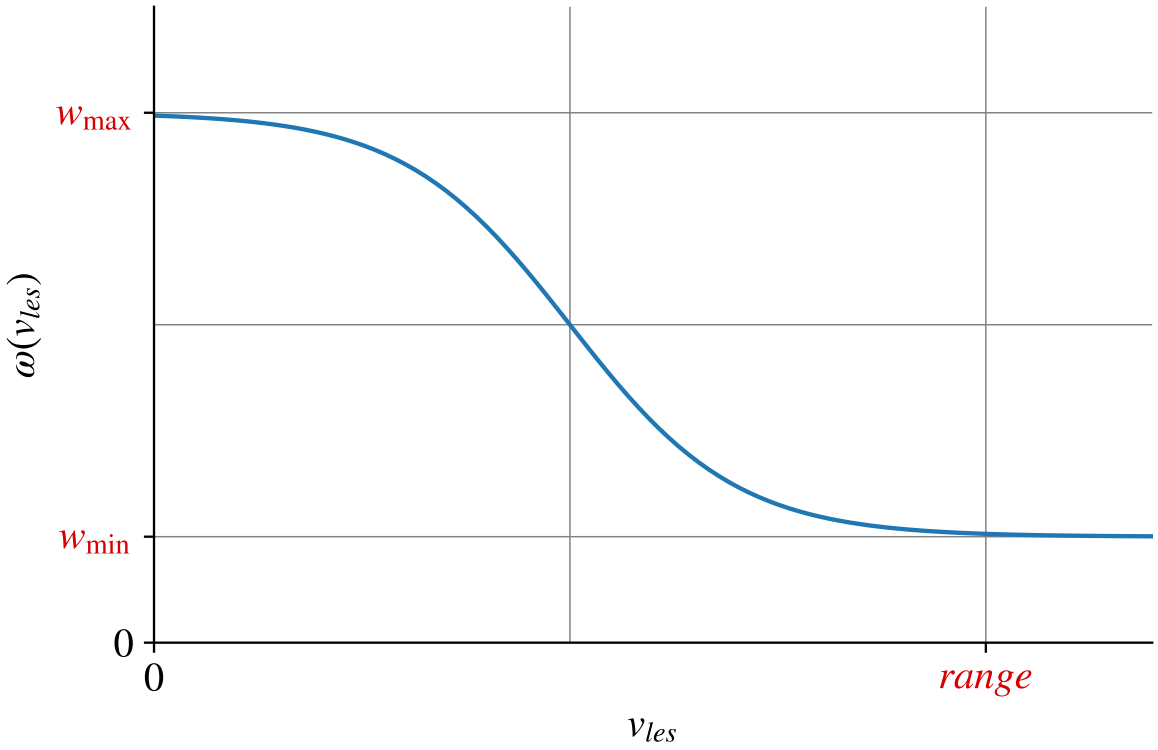


Fig 1 The lesion size weighting function $\omega(v_{les})$ with $w_{\max}$, $w_{\min}$, $range$ parameters illustrated

Parameter α is set to have the inflexion point in the middle of the $range$, which is equivalent by symmetry to solve:

$$w_{\max} - \omega(0) = \omega(range) - w_{\min} \implies \alpha = \sqrt{e^k}.A_{\max} \tag{6}$$

In practice, $w_{\max}$, k and α has been considered as constants and the two parameters of interest are:

- $w_{\min}$ which sets the weighting contrast between small and big lesions,
- $range$ that sets the decreasing speed and the threshold between small and big lesions.

In order to improve the lesion detection, and small lesion detection in particular, parameters $w_{\min}$ and $range$ of the weighting function are investigated in this study.

2.3 *Loss function*

The loss function used is basically a weighted Dice loss function calculated at a voxel level. To weight the lesion voxel depending on the lesion volume, the mapping of weighting function $\mathbf{\Omega}$ is firstly generated assigning for each lesion voxel position the value of $\omega(v_{les})$ where v_{les} is the volume of the lesion in which the voxel belongs. Weighted Dice loss function is written:

$$WD(\mathbf{P}, \hat{\mathbf{P}}, \mathbf{\Omega}) = -2\frac{\sum_{b=1}^{B}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{d=1}^{D}\mathbf{P}_{b,h,w,d}\hat{\mathbf{P}}_{b,h,w,d}\mathbf{\Omega}_{b,h,w,d}}{\sum_{b=1}^{B}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{d=1}^{D}\mathbf{P}_{b,h,w,d} + \hat{\mathbf{P}}_{b,h,w,d}}, \tag{7}$$

where $B \in \mathbb{N}^+$ is the batch size, $H \in \mathbb{N}^+$ the height, $\Omega \in \mathbb{N}^+$ the width, $D \in \mathbb{N}^+$ the depth, $\mathbf{P} \in \{0,1\}^{BHWD}$ the ground truth segmentation and $\hat{\mathbf{P}} \in [0,1]^{BHWD}$ the model prediction. The loss function, by definition, gives more importance to matched voxel with a high weight. It, then, leads the model to increase its sensitivity for high-weighted voxels that are voxels belonging to small lesions in this case.

2.4 *CNN architecture*

The U-net architecture, originally used for segmentation in electronic and photonic microscopy[17] has been extensively used for segmentation tasks and particularly in medical images. It has become a reference architecture for pixel- and voxelwise segmentation and has been successfully adapted for MS lesion segmentation by multiple research teams. The top-performing methods of MS lesion segmentation (from Zhang *et al.*,[14] Isensee *et al.*,[18] Kang *et al.*[19]) on the ISBI MS segmentation challenge,[13] use U-net like architectures.

We used the MPU-net++2CBND architecture,[20] renamed MPU-next in this article. It is a light

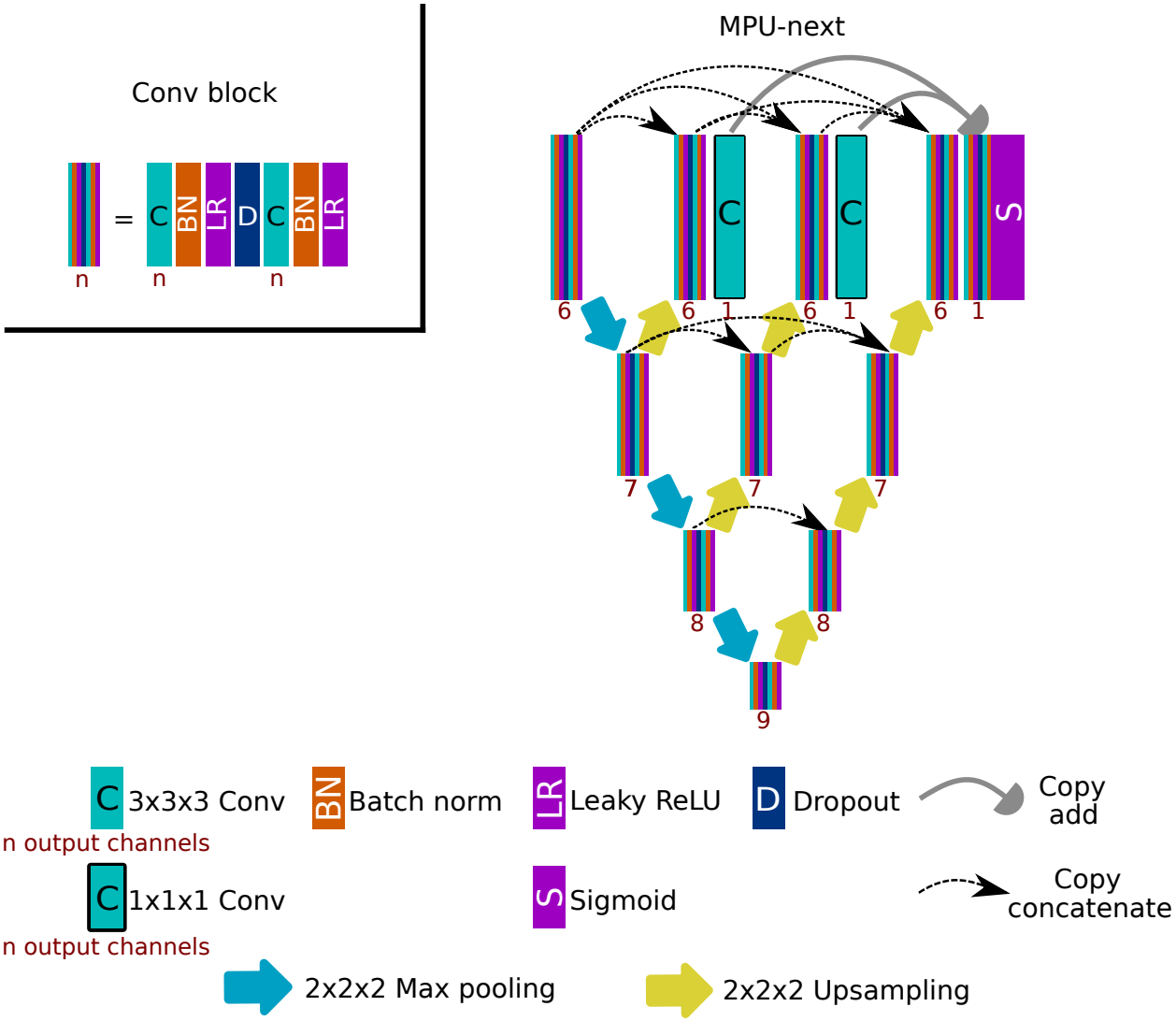


Fig 2 The architecture of the MPU-next model, also referred as MPU-net++2CBND in a previous article.[20]

U-net-like architecture illustrated in **Figure** 2. The MPU-next architecture has been designed to be very light with 22 convolutional layers and 37, 935 learning parameters only. It has been chosen for its good segmentation performance on small datasets[20] and its fast convergence because of its lightness.

2.5 Dataset

2.5.1 In-house dataset

We used an in-house dataset from the university hospital of Poitiers gathered by the I3M laboratory as dataset for the lesion size study. This dataset is used for the study since it contains high resolution FLAIR images and to be able to compare the segmentation prediction with the radiologist who annotated the images. In this way, the radiological feedback of the method comes from the annotator itself that can confront the prediction and its segmentation. The dataset contains FLAIR cerebral MR images from 35 different MS diagnosed patients of the university hospital of Poitiers. Only FLAIR images were gathered since it is the most clinically valued sequence for MS lesion detection and also the most important for learning.[21] The FLAIR images have been acquired sagittally with a resolution of $1 \times 0.5 \times 0.5\ mm^3$ resolution by a *Verio* and a *Skyra* Siemens MR scanners with a 3 Tesla magnetic field. The segmentation masks were segmented slice by slice for each exam by the same radiologist on not preprocessed FLAIR images.

The training set is composed of 20 FLAIR images and the testing set by 15 FLAIR images. The size of the training set is enough with the light architecture used since 10 images are sufficient.[20] The MR images have been carefully distributed in each set equitably regarding the lesion profile (*ie* the lesion sizes, locations, density) to make training and testing set as much representative and comparable as possible.

For the training and testing images were preprocessed by $1\ mm^3$ isometric resampling to work on cubic voxel with a good compromise between memory space and lesion size. Images are skull-

stripped with the *BrainSuite extractor* tool, histogram matched[22] and the intensities are centered and reduced to having comparable and in a small range around zero.

2.5.2 ISBI-2015 dataset

The ISBI 2015 MS segmentation challenge[1,32] dataset has been used to compare and validate the method with other existing methods thanks to the online open submission.

The ISBI training dataset is composed of 21 preprocessed MR exams from different time points (4 to 5) of five different patient. The exams consist in T1, T2, FLAIR and Proton Density (PD) images. Ground truth segmentations from two different radiologists are provided for each exam.

The testing dataset is composed by 61 exams from 14 different patients acquired in another MR scanner. For those examples, the ground truth segmentation is not provided.

All the images have been preprocessed by the challenge organization including registration steps and $1 \times 1 \times 1\ mm^3$ resampling. Note that the FLAIR image acquisition resolution is thicker than the in-house FLAIR images with voxel size of $0.8 \times 0.8 \times \{4.4, 2.2\}\ mm^3$.

2.6 Training and testing

Since the best set of parameters $\{w_{\min}, w_{\max}, range, k, \alpha\}$ is unknown, multiple configurations have been evaluated. In practice, parameter $w_{\max}$ has been set to $w_{\max} = 10$, to ensure the loss to be big enough when $w_{\min}$ is close to 0, given the rare occurrence of small lesion voxels, to avoid gradient issues due to the float precision when the model begins to be trained. The k parameter has been fixed in order to be close enough to each asymptote when $v_{les} = 0$ and $v_{les} = range$. We set $k = 10$ to be close to the asymptote at those points with a distance of ≈ 0.06. Then α is set to $\sqrt{e^{10}} = e^5$ following **Equation** 6.

[2]leaderboard available at https://smart-stats-tools.org/lesion-challenge

In total, 32 combination of weighting parameters $w_{\min}$ and $range$ have been tested, the corresponding profiles of the corresponding weighting function are provided in **Figure** 7 of the appendix.

Each learning is initiated with the exact same model with the MPU-next architecture initialized uniformly[23] with the same values in order to evaluate the trained model performances without the initialization factor. Following previous work,[20] the model is trained with $32 \times 32 \times 32$ patch with a 2, 048 batch size. The models are trained with 6, 144 patches randomly extracted in the brain volume of each training exams each epoch and are trained 40 epochs. The Adam optimizer[24] is used with a learning rate of 0.0004.

The predictions are generated per patch regularly spaced by 8 voxels in each spatial direction and spatially averaged to obtain the predicted image. The prediction map is resampled and registered to the original resolution, it is then binarized with a threshold of 0.5 before being evaluated.

Every test segmentation prediction is evaluated with the Anima segmentation performance analyzer script[3].

3 Results and analysis

The problem of detecting small lesions is not obvious in most evaluations because most segmentation are evaluated regardless of lesion size and metrics are mostly aggregated into statistical descriptors calculated by patient with truly distinct lesion profiles. We propose here to deepen the problem of lesion detection and segmentation through analysis at the patient and lesion level. In this section, we show how it is important to consider the lesion size and discuss the improvements

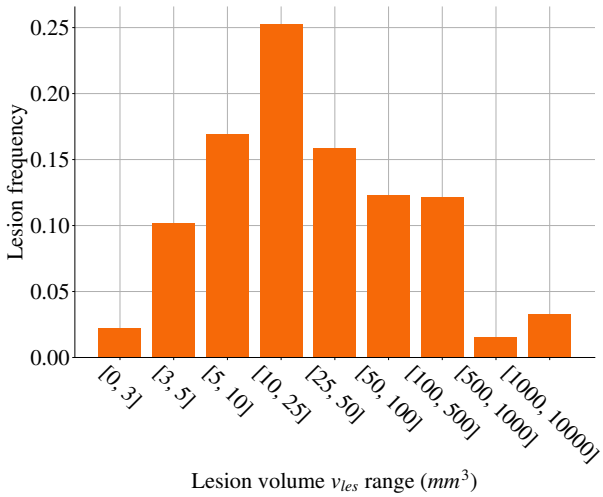
[3] https://anima.irisa.fr/

brought by our method.

3.1 The lesion size

The size-adaptative method is based on the observation that the described model trained with the common Dice loss gives an acceptable Dice score but a bad F1 score when averaged by patients. So, even if the voxel-wise segmentation is acceptable, the quality of lesion detection is low. The performances vary a lot from an exam to another and are particularly low for patients with small and few lesions.

(a) Lesion-wise



(b) Voxel-wise

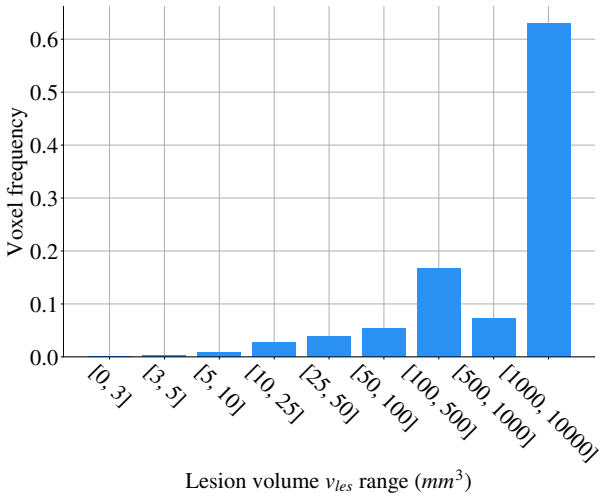


Fig 3 The frequency of lesions depending on the lesion volume in the in-house dataset in (a) and the frequency of voxels depending on the lesion volume they belong to (b). The volume range is not linear to avoid flatten representation due to a large volume range. The frequencies are calculated on the test set to be comparable with other results.

In **Figure** 3(a) 2% of lesions are less than $3mm^3$, 10% of lesions have a volume in $[3, 5]$, 17% of lesions have a volume in $[5, 10]$ and 55% of lesions have a volume inferior to 26 mm^3 in the test set. This indicates that most of the lesions are really small lesions with a volume inferior to 26 mm^3. But, compared to the voxel representation of those lesion in **Figure** 3(b), the small lesions

with a volume inferior to 26 mm^3 represents less than 4% of voxels whereas very large lesions with a volume superior to 1000 mm^3 represent 63% of lesion voxels whereas they represent only 3% of lesions. This phenomenon is a problem since segmentation losses is computed at a voxel level and results in a over-representation of big lesion voxels.

3.2 Size-adaptative method

The size adaptative method is always compared to the performances of the model learned without weighting, with the Dice loss.

3.2.1 Patient-wise

Dice	F1	LTPR	LFPR
0.5990	0.4883	0.4490	0.4236

Table 1 Average performances of the model trained with the simple Dice loss for an exam of the in-house test set. LTPR is the lesion detection sensitivity and LFPR is the lesion false positive rate.

In **Figure** 4 and **Table** 1, we can see the voxel-wise segmentation performances in Dice score and detection performances in terms of F1 score, LTPR and LFPR averaged per patient.

For all metrics, we observe different behaviors when $w_{\min} \leq 0.01$. For each metric, when $w_{\min} \leq 0.01$ and $range \leq 100$, the worst performances are obtained, lower than the scores obtained without weighting in **Table** 1. This indicates that the rapid and abrupt separation of large and small lesion weights (see **Figure** 7 in appendix) leads to an overall decrease in lesion detection and segmentation quality. When the $range$ increases with the same values of $w_{\min}$, we observe an improvement which shows that by keeping the same values of weights, but spreading them more over the size range, the model learns to detect more lesions and segment them better.

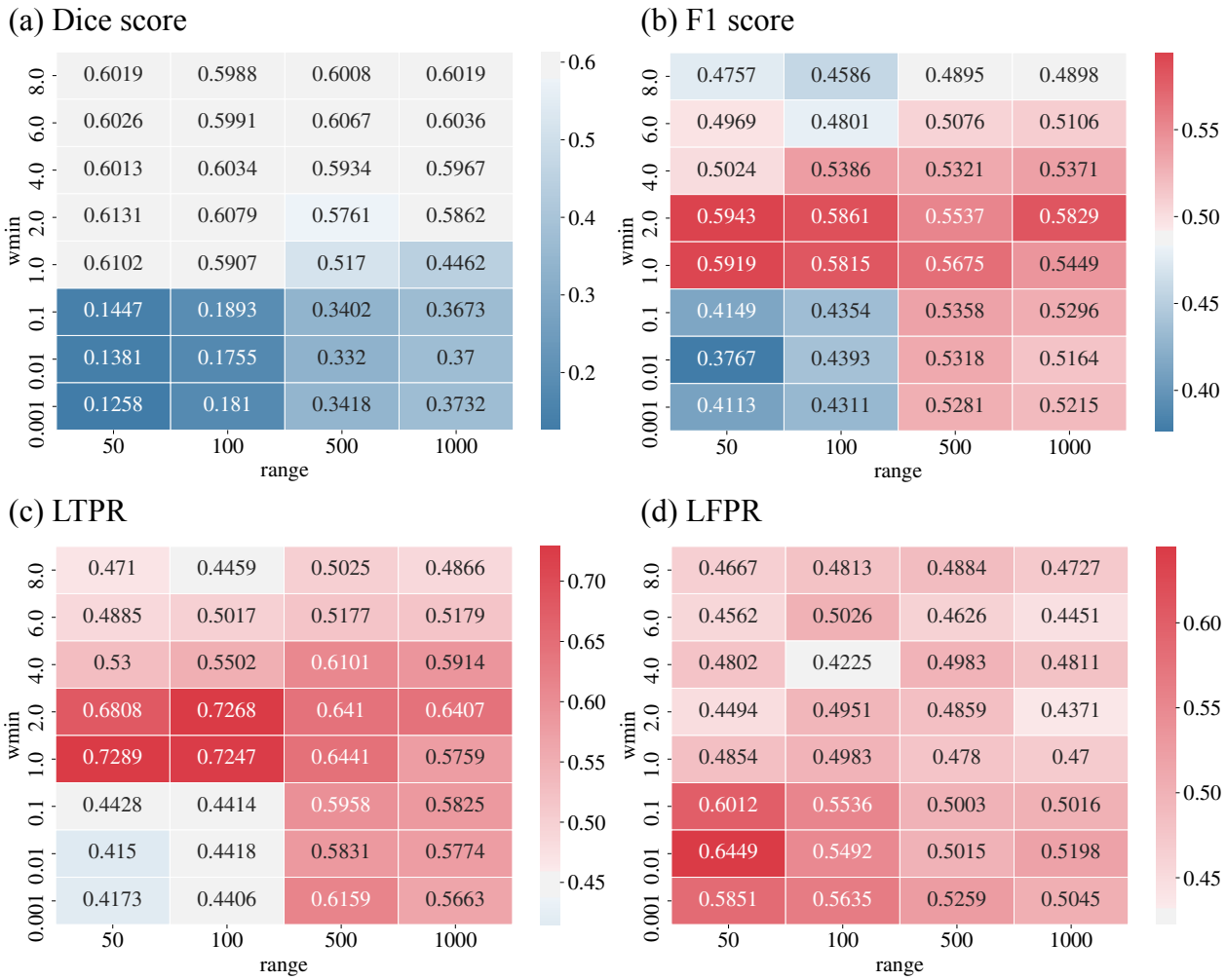


Fig 4 Average performances of the model trained with weighted Dice loss with the set of tested parameters $w_{\min}$ and $range$ for an exam of the in-house test set. Voxel-wise segmentation is measured in Dice score (a), the lesion detection is measured in F1 score (b), in LTPR (c) and in LFPR (d). For each metric, the color blue indicates that the value is inferior to the corresponding metric of the model trained without weighted Dice loss (**Table** 1) and the red color indicates that it is superior.

In **Figure** 4, we observe that by decreasing $w_{\min}$ to 2, and 1, and decreasing $range$ to 100 and 50 there is an increase in Dice score, F1 and LTPR while there is no real trend on LFPR. The LFPR is even higher than without weighting for almost every weighting function evaluated. This observation remains consistent with the weighting which is designed to improve the detection of small lesions that are generally more difficult to spot and more easily confused with healthy brain structures.

The weighting functions tested in Figure 4 only slightly improve the Dice score up to 0.6103 for

$w_{\min} = 2$ and $range = 50$ compared to 0.5990 without weighting in **Table** 1. Voxel segmentation is thus slightly improved, however, lesion detection is drastically improved with an F1 score up to 0.5943 for the same values of $w_{\min}$ and $range$ compared to an F1 score of 0.4883 without weighting. The gain in F1 score is explained by the drastic increase in LTPR brought by the method up to 0.7289 on average for $w_{\min} = 1$ and $range = 50$. On the other hand, the LFPR is slightly higher between 0.45 and 0.5 for the best F1 scores against 0.4236 without weighting. This behavior is consistent with the objective of method to detect more lesions and particularly the smallest ones. In fact, without weighting, the learned model detects less than half of the lesions with a LFPR close to 42% whereas with the right set of weights we can detect more than 68% of the lesions or even 73% in average per patient with a higher LFPR of 3% to 8%. In these cases, weighting increases drastically the lesion detection with an F1 score around 0.59 against 0.49 without the proposed method.

On average (per patient), the best segmentations and lesion detections are achieved for $w_{\min} = 2$ with $range = 50$ with better (lower) LFPR also and for $w_{\min} = 1$ with $range = 50$ for better TPR. Matching these results with the weighting model, we have the best detection performance by weighting small lesions 5 to 10 times more than large lesions with a progressive separation of lesions by volume up to $50\ mm^3$.

3.2.2 Lesion-wise

Figure 5 shows the detection capacity of the model trained with the best sets of weighting parameters *ie* $w_{\min} \in \{1, 2\}$ and $range \in \{50, 100\}$ depending on lesion size. The same analysis performed on all tested weighting functions is available in **Figures** 8, 9 and 10 of appendix. In **Figure** 5 it appears plainly that the presented method improvement comes from a significantly bet-

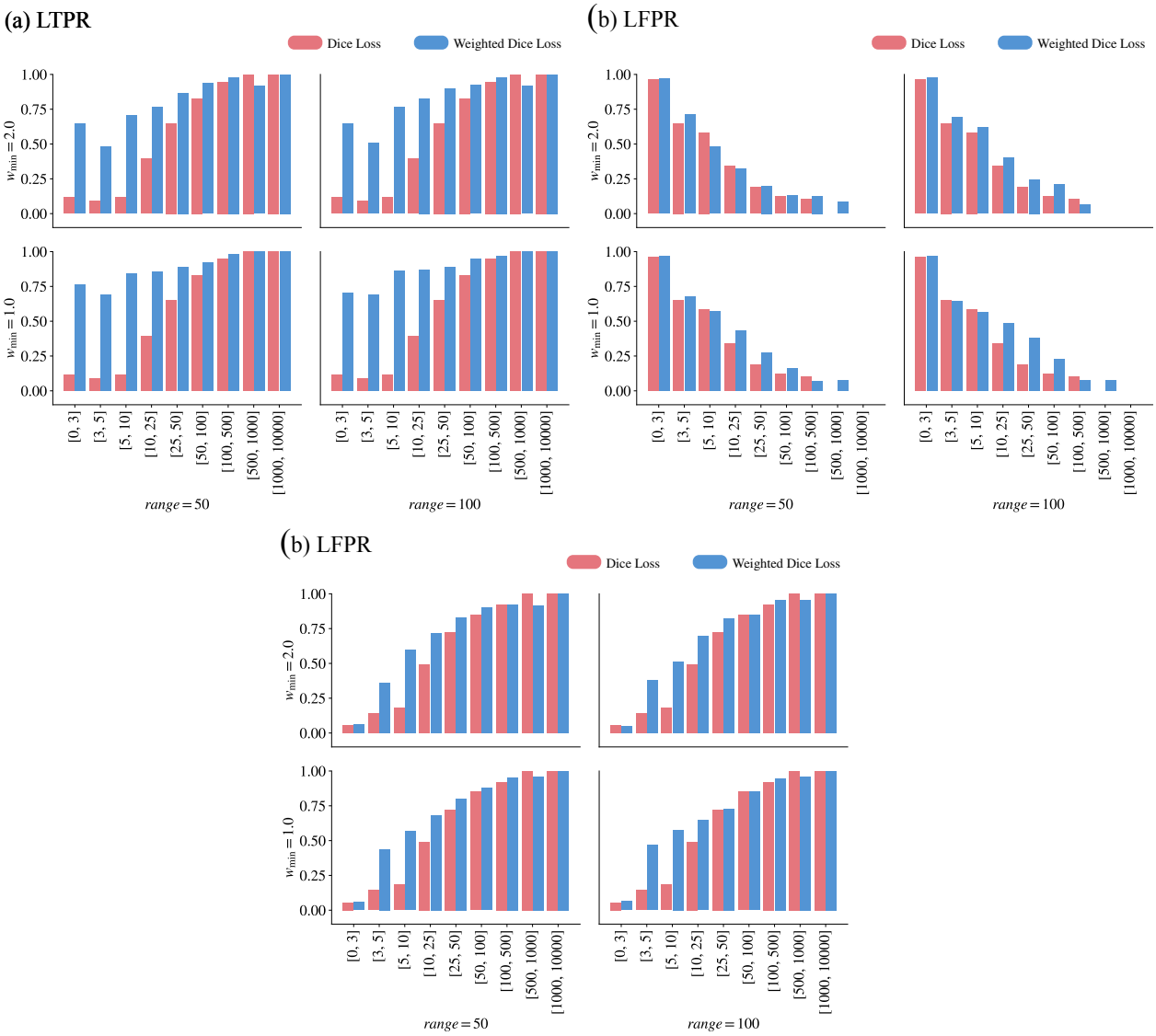


Fig 5 The lesion detection in LTPR (a), LFPR (b) and F1 score (c) depending on the lesion volume in mm^3 of the four best weighting functions in blue with $range \in \{50, 100\}$ and $w_{\min} \in \{1, 2\}$ and without weighting in red. The performances of all tested weighting functions are available in **Figures** 8, 9 and 10 of appendix. Computed on the in-house test set.

ter LTPR. It means that more small lesions are detected with the method. However, the LFPR, *ie* the rate of false lesions detected among all predicted lesions, is slightly higher in general, but not as far as the sensitivity is improved. It means that, with the weighting method, the model learns a new lesion detection compromise between sensitivity and precision and detects far more lesions at the cost of detecting some not existing ones.

The F1 score, more commonly used to assess the detection performance, is the harmonic mean of LTPR and PPV (1 - LFPR). The F1 score of best methods depending on the lesion volume is illustrated in **Figure** 5(c). In this figure, the benefit of the proposed weighting method in lesion detection is mostly for small lesions with volume in $[3, 25]$ mm^3. However, compared to the distribution of lesion size in **Figure** 3(a), the lesions with volume in $[3, 25]$ mm^3 represents more than 52% of lesions. So, the proposed method improves the detection of small lesions that represents more than half of lesions. The smallest lesions with a volume less than 3 mm^3 do not have a good F1 score since the LFPR is very high for those lesions despite the good LTPR up to 75% depending on weighting parameters $range$ and $w_{\min}$.

To summarize, the proposed method, with the right set of parameters, improves the detection of small lesions, and particularly the detection sensitivity. Small lesions are the most represented lesions and also the more difficult to spot for a radiologist. The proposed method, then, helps to improve the value of automatic MS lesion segmentation tools for radiologists where it is the most valued.

3.3 Radiological confirmation of the method

The radiological assessment of the method was performed on multiple test patients with different weighting parameters sets. For clarity, we have illustrated the observations with the example in **Figure** 6.

In this figure, we observe, for this example, at first that the unweighted segmentation (the reference) and the radiologist segmentation are very close, except for a small lesion not segmented by the model near the gray matter of the right brain cortex. Some parts are over-segmented and

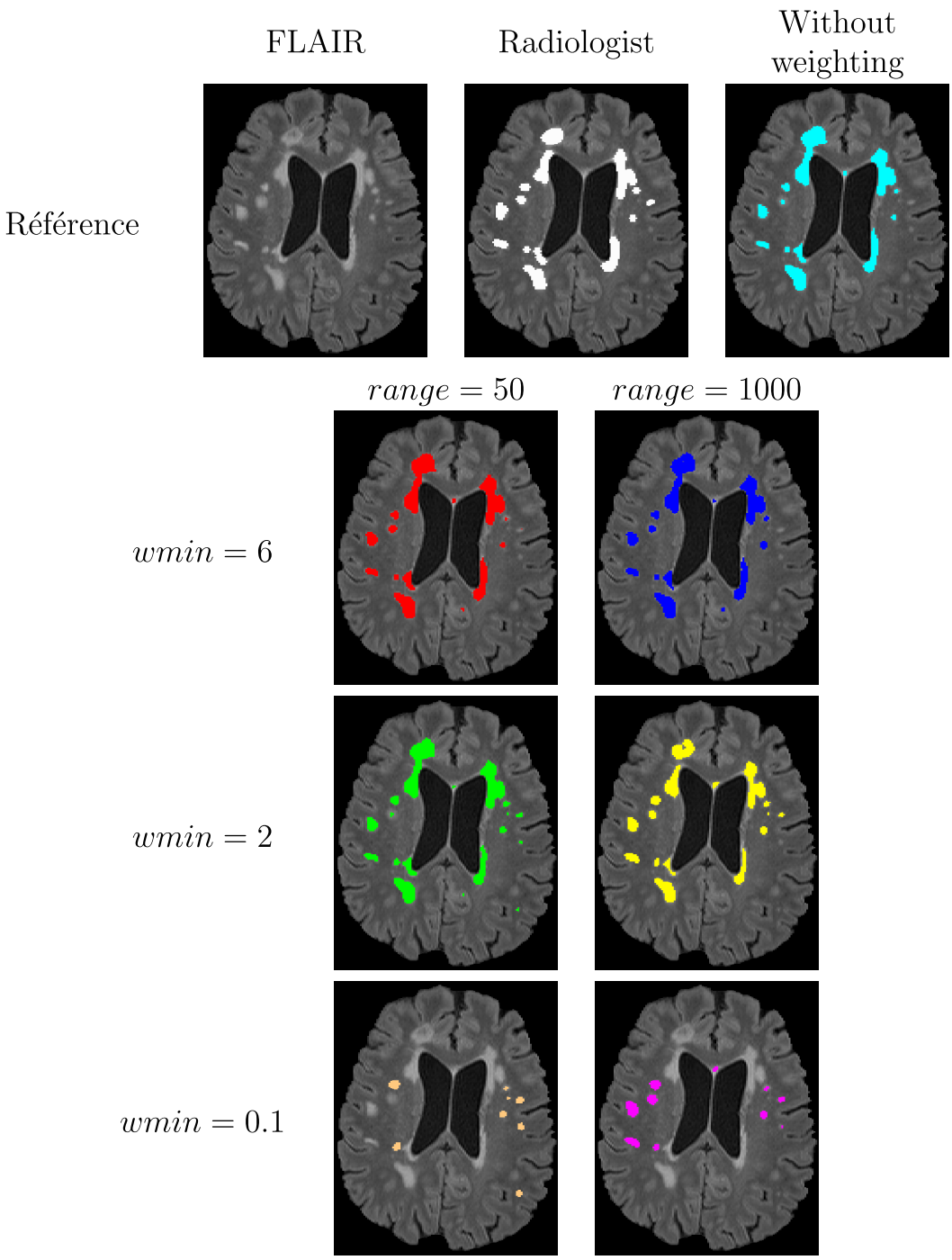


Fig 6 Visualization of ground truth segmentation and predicted segmentations on an example of the in-house dataset without weighting and with weighting functions with $range \in \{50, 1000\}$ and $w_{\min} \in \{0.1, 2, 6\}$.

others under-segmented compared to the reference segmentation. When $w_{\min} = 0.1$, we observe that the largest lesions are not segmented at all, indicating that the smallest lesions are overly favored during training. However, with $range$ augmenting from 50 to 1,000, the model tends to

segment bigger lesions which is consistent with weighting model.

The segmentation for $w_{\min} = 6$ is very close to the unweighted segmentation which is consistent with the results in **Figure** 4 and is explained by only a small difference in weighting between different lesion sizes.

For $w_{\min} = 2$, new small lesions are segmented that are not on the radiological segmentation. However, after verification, these are edges of lesions that were not segmented, very small lesions that were not seen on the radiological segmentation and one false positive lesion. With $w_{\min} = 2$, we observe that the increase of $range$ from 50 to 1000 eliminates the smallest segmented lesions, which remains consistent with the weighting function, indeed, the passage from large to small lesions is more progressive and spread out, which decreases the importance of very small lesions.

From observation of the predicted segmentations, we see that with $w_{\min}$ values approaching 0, there is a tendency to ignore large lesions and prefer only smaller ones. For $w_{\min}$ values at 2 and 1 and $range$ values at 50 and 100, we observed very good segmentations with some false negatives. Several small lesions that are not detected without weighting are detected with this weighting.

More generally, when observing the different segmentations, we found that the best performing models tend to segment more widely at the edge of lesions than the reference radiologist segmentation, thus lowering the Dice score without influencing the radiological analysis of the segmentation itself. The noisiest images were also the images on which the models performed the least well, however, the models were not trained with data augmentation which could have alleviated this problem, at least in part. The learned models segmented fewer lesions at the base of the skull. In fact, this area contains artifacts that creates local hyper-intensities that are not lesions and it contains few lesions in general. The models have therefore learned to recognize this area and to avoid to segment any hypersignals in this volume.

A few lesions have been confirmed to be missed by our radiologist but segmented by the model. This last observation points out an interesting fact: an automatic segmentation model, although imperfect, can help a practitioner by detecting lesions that are not very prominent or too small to be obvious, especially with our weighting method.

3.4 Evaluation on the ISBI-2015 challenge dataset

The method has been trained on the ISBI-2015 challenge dataset and evaluated by the organization on their test set. The models were trained with every available MR sequence provided (T1, T2, FLAIR and Proton Density) in order to stay consistent with the challenge and to allow the model to work not only with upsampled FLAIR image to grasp the brain anatomy in working resolution. Data augmentation has also been performed in order to help the model to generalize and to improve prediction on unseen data domain since the test dataset is not acquired with the same machine. The data augmentation performed consists in random flip in 3 dimensions, random blur, random gaussian noise and random anisotropy as implemented in the TorchIO library.[25]

The method is evaluated and compared on the ISBI-2015 challenge dataset on **Table** 2. Our method outperforms the other known published attempts in terms of lesion sensitivity with a LTPR of 0.623 for $w_{\min} = 1$ and $range = 100$. However, the LFPR reaches 0.457, F1 score is 0.521 and the Dice score is 0.576 which are outperformed by most of other recent approaches in **Table** 2.

Compared to the unweighted method, the weighted MPU-next with $w_{\min} = 1$ and $range = 100$ allows a substantial gain in Dice score (+0.060), F1 score (+0.106), LTPR (+0.301) for a loss of 0.215 in LFPR.

Surprisingly, the other evaluated weighting parameters ($\{w_{\min} = 1, range = 50\}, \{w_{\min} = 2, range = 50\}, \{w_{\min} = 2, range = 100\}$) did not perform as well as in **Figure** 4 computed

Method	Dice	F1	LTPR	LFPR
Weighted MPU-next $w_{\min} = 1, range = 100$ (ours)	0.576	0.521	**0.623**	0.457
Multi-view CNN[26]	0.627	0.533	0.568	0.498
All-Net[14]	0.639	**0.663**	0.533	0.122
nn-Unet[18]	**0.679**	0.645	0.523	0.159
Weighted MPU-next $w_{\min} = 2, range = 100$ (ours)	0.561	0.505	0.495	0.356
Ensembling Models[27]	0.622	0.645	0.491	0.151
Multi-Dimensional GRU[28]	0.629	0.605	0.487	0.201
Geo-loss[29]	0.643	0.618	0.480	0.132
Weighted MPU-next $w_{\min} = 1, range = 50$ (ours)	0.463	0.524	0.476	0.293
Location-aware CNN[30]	0.501	0.426	0.429	0.577
Imagine focal[31]	0.584	0.569	0.414	**0.087**
Multi-branch CNN[32]	0.611	0.556	0.410	0.139
Cascaded 3D CNN[33]	0.630	0.512	0.367	0.153
Weighted MPU-next $w_{\min} = 2, range = 50$ (ours)	0.449	0.451	0.365	0.255
MPU-net[34]	0.632	0.429	0.347	0.347
MPU-next without weighting	0.516	0.415	0.322	0.242

Table 2 Comparison of average performances on the ISBI-2015 challenge dataset with known published attempts and our method with weighting functions with $range \in \{50, 100\}$ and $w_{\min} \in \{1, 2\}$ and without weighting ordered by LTPR.

on the in-house dataset used for the rest of the study. This can be explained by the fact that the in-house dataset and the ISBI-dataset have a really different lesion size distribution and so the optimal weighting setting is different. As a matter of fact, the in-house dataset has been segmented on FLAIR images acquired with a fine resolution of $1 \times 0.5 \times 0.5\ mm^3$ against an acquisition resolution of $0.8 \times 0.8 \times \{4.4, 2.2\}\ mm^3$ for the ISBI-2015 training FLAIR images which results in augmenting the minimal discernible lesion size.

4 Discussion

The automatic MS lesion segmentation can be included into CAD systems for helping radiologists in a daily, tiring and time-consuming task. Given, the context of the disease, it is radiologically more valuable to detect any lesion suspicion than to be sure on prediction. Indeed, it is easier to discard a false detected lesion (such as artifacts or non pathological thin hyperintensities in con-

tact with the ventricles) than to spot undetected lesions. In clinical practice, it is essential for the follow-up not to miss any new lesion, even if it is very small, because that may change the medical care. Moreover, the detection is crucial when assessing the spatial and temporal dissemination of lesions that are required for the diagnostic. However, most performing methods for MS lesion segmentation, rely on voxel-wise classification losses that give more importance to big lesion voxels even if small ones have a better radiological value. It results, most of the time, in a competitive voxel-wise segmentation with a limited capacity to detect small lesions. In this work, we focused on giving more importance to small lesions, during model training, in order to improve small lesion detection for a better radiological value.

We showed that, with our weighting method, we can significantly improve the small lesion detection and the lesion detection in general since small lesions are the most frequent. The lesion detection is improved in terms of sensitivity with a LTPR improvement of 0.28 on our in-house dataset and of 0.30 on the ISBI-2015 challenge test set. This improvement is accompanied by an improved Dice score of about 0.01 on our in-house dataset and 0.06 on the ISBI-2015 test set and an increased F1 score from 0.49 to 0.59 on our in-house dataset and from 0.415 to to 0.521 on the ISBI test. The LFPR, however, increases up to 0.8%.

Small lesions are the more difficult to discover, especially when the lesion load is important. Furthermore, the new lesions on successive MR exams are the most crucial to detect in spite of their often limited size. The presented method has an important radiological significance given the applied context, since less perceptible lesions are more spotted and the overall sensitivity is improved. The high LFPR on smallest lesions can be partially explained by the existing difficulty to detect those lesions by radiologists and consequently by the absence of some of these small lesions in the segmented ground truth. Cases of missing segmented lesions in our in-house dataset

have been observed resulting in wrong false positive leading to a biased increase of LFPR. In the same way, in the ISBI-2015 dataset, some lesions, especially small ones, were spotted by only one radiologist segmentation over the two available pointing out the difficulty of spotting small lesions.

The weighting method has, however, to be adapted depending on the dataset and acquisition resolution. More the resolution is fine more the small lesions have to account in the segmentation loss of the learning model. The proposed weighting can be easily reused with other models and methods in order to improve the lesion detection and the radiological contribution for a better adoption of such tools for a better patient care.

Appendix A: Additional figures and results

The appendix contains the profile of all tested weighted function ω in Figure 7. It also gather the performances in terms of LTPR (Figure 8), LFPR (Figure 9) and F1 score (10) when training with those weighted functions depending on the lesion size on the in-house dataset.

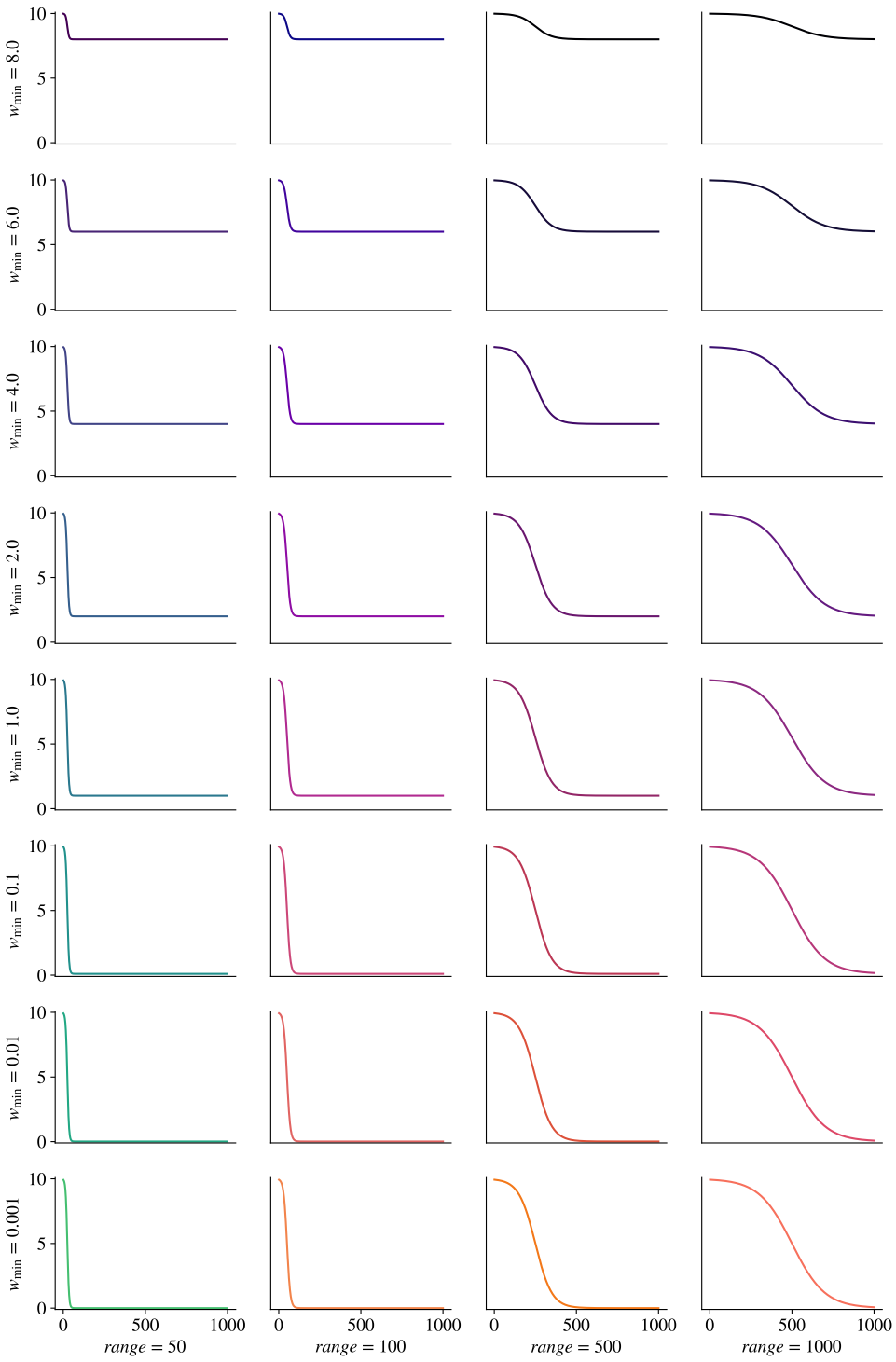


Fig 7 Profiles of the tested weighting functions ω. $w_{\max}$ is set to 10 and only $w_{\min}$ and $range$ are varying. $range$ sets the distance between the maximal and minimal asymptotes and $w_{\min}$ gives the minimum weighting value associated to the largest lesions.

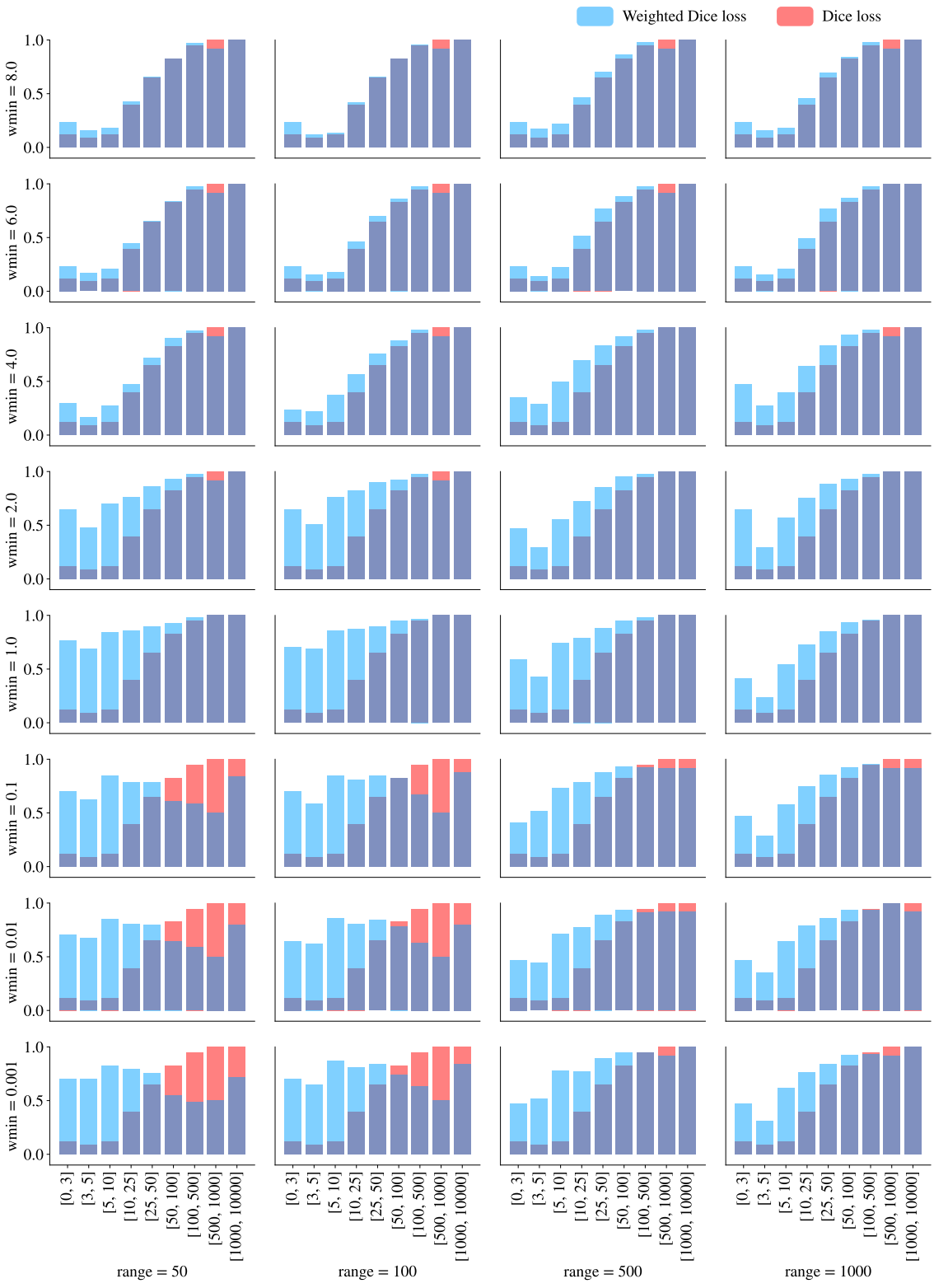


Fig 8 The lesion detection sensitivity (LTPR) depending on the lesion size with all tested weighting functions ω with varying $w_{\min}$ and $range$ in blue and without weighting in red. Computed on the in-house test set.

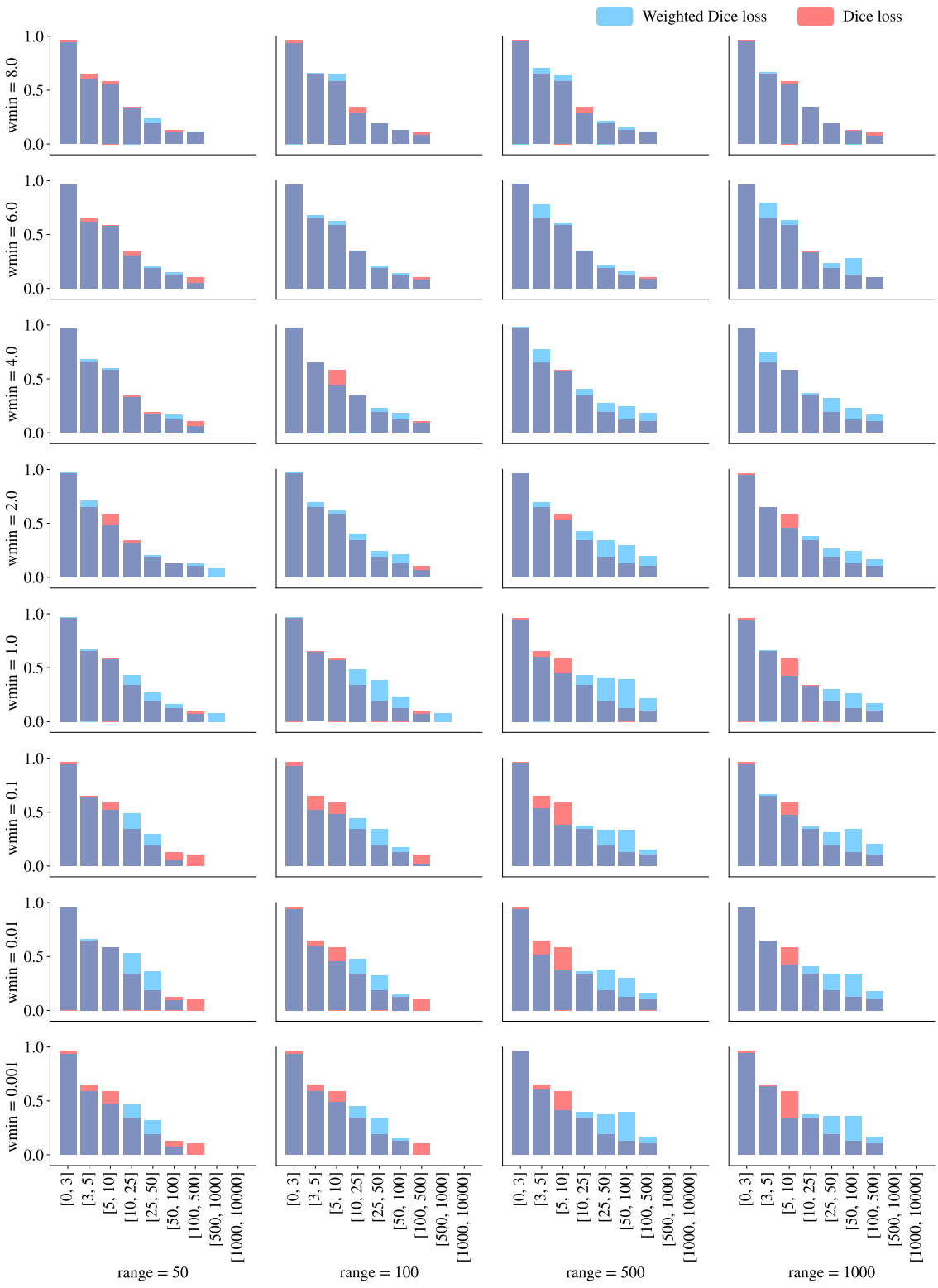


Fig 9 The lesion detection false positive rate (LFPR) depending on the lesion size with all tested weighting functions ω with varying $w_{\min}$ and $range$ in blue and without weighting in red. Computed on the in-house test set.

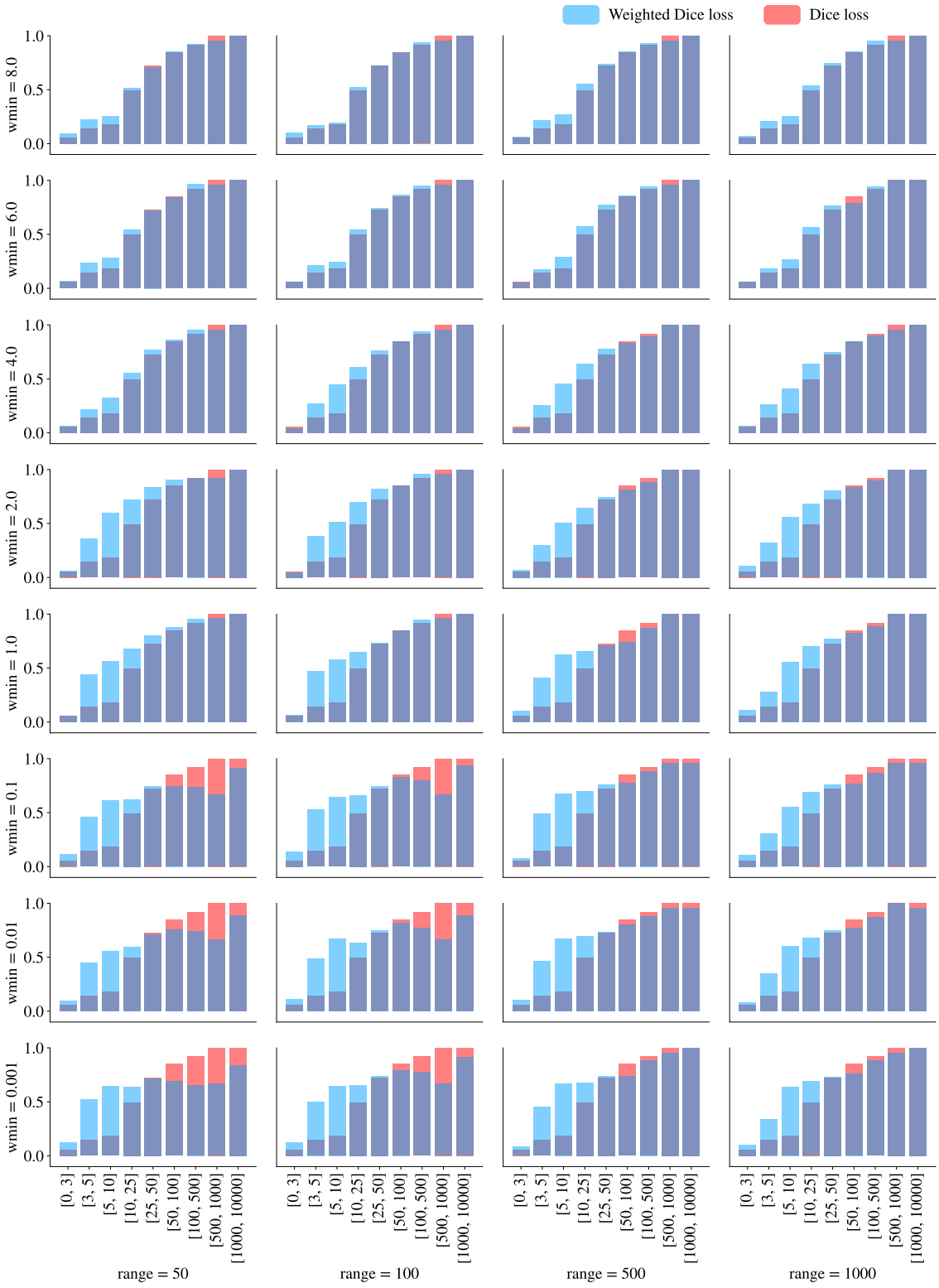


Fig 10 The lesion detection F1 score depending on the lesion size with all tested weighting functions ω with varrying $w_{\min}$ and $range$ in blue and without weighting in red. Computed on the in-house test set.

Disclosures

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

This work will not have emerged without the implication of Christophe Habas who passed away in May 2022 and we thank him for his contribution on the whole work built together.
We thank Siemens Healthineers that supported this work.

References

1 M. T. Wallin, W. J. Culpepper, E. Nichols, *et al.*, "Global, regional, and national burden of multiple sclerosis 1990–2016: a systematic analysis for the global burden of disease study 2016," *The Lancet Neurology* **18**(3), 269–285 (2019).

2 R. Dobson and G. Giovannoni, "Multiple sclerosis a review," *European journal of neurology* **26**(1), 27–40 (2019).

3 C. Baecher-Allan, B. J. Kaskow, and H. L. Weiner, "Multiple sclerosis: mechanisms and immunotherapy," *Neuron* **97**(4), 742–768 (2018).

4 A. J. Thompson, B. L. Banwell, F. Barkhof, *et al.*, "Diagnosis of multiple sclerosis: 2017 revisions of the mcdonald criteria," *The Lancet Neurology* **17**(2), 162–173 (2018).

5 J.-C. Brisset, S. Kremer, S. Hannoun, *et al.*, "New ofsep recommendations for mri assessment of multiple sclerosis patients: special consideration for gadolinium deposition and frequent acquisitions," *Journal of Neuroradiology* **47**(4), 250–258 (2020).

6 M. Cabezas, J. Corral, A. Oliver, *et al.*, "Improved automatic detection of new t2 lesions in multiple sclerosis using deformation fields," *American Journal of Neuroradiology* **37**(10), 1816–1823 (2016).

7 A. Danelakis, T. Theoharis, and D. A. Verganelakis, "Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging," *Computerized Medical Imaging and Graphics* **70**, 83–100 (2018).

8 A. Shoeibi, M. Khodatars, M. Jafari, *et al.*, "Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review," *arXiv preprint arXiv:2105.04881* (2021).

9 A. Kaur, L. Kaur, and A. Singh, "State-of-the-art segmentation techniques and future directions for multiple sclerosis brain lesions," *Archives of Computational Methods in Engineering* , 1–27 (2020).

10 Y. S. Vang, Y. Cao, P. D. Chang, *et al.*, "Synergynet: A fusion framework for multiple sclerosis brain mri segmentation with local refinement," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 131–135, IEEE (2020).

11 I. Coronado, R. E. Gabr, and P. A. Narayana, "Deep learning segmentation of gadolinium-enhancing lesions in multiple sclerosis," *Multiple Sclerosis Journal* , 1352458520921364 (2020).

12 S. Valverde, M. Salem, M. Cabezas, *et al.*, "One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks," *NeuroImage. Clinical* , 101638 (2018).

13 A. Carass, S. Roy, A. Jog, *et al.*, "Longitudinal multiple sclerosis lesion segmentation: resource and challenge," *NeuroImage* **148**, 77–102 (2017).

14 H. Zhang, J. Zhang, C. Li, *et al.*, "All-net: Anatomical information lesion-wise loss function integrated into neural network for multiple sclerosis lesion segmentation," *NeuroImage: Clinical* , 102854 (2021).

15 A. Carass, S. Roy, A. Gherman, *et al.*, "Evaluating white matter lesion segmentations with refined sørensen-dice analysis," *Scientific Reports* **10**(1), 1–19 (2020).

16 R. A. Kamraoui, V.-T. Ta, J. V. Manjon, *et al.*, "New ms lesion segmentation with lesion-wise metrics learning," *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure* , 29 (2021).

17 O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, *LNCS* **9351**, 234–241, Springer (2015). (available on arXiv:1505.04597 [cs.CV]).

18 F. Isensee, P. F. Jaeger, S. A. Kohl, *et al.*, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods* , 1–9 (2020).

19 G. Kang, B. Hou, Y. Ma, *et al.*, "Acu-net: A 3d attention context u-net for multiple sclerosis lesion segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1384–1388, IEEE (2020).

20 A. Fenneteau, P. Bourdon, D. Helbert, *et al.*, "Cnn for multiple sclerosis lesion segmentation: How many patients for a fully supervised method?," in *International Conference on Advances in Biomedical Engineering*, (2021).

21 Y. Feng, H. Pan, C. H. Meyer, *et al.*, "A self-adaptive network for multiple sclerosis lesion segmentation from multi-contrast mri with various imaging protocols.," *CoRR* **abs/1811.07491** (2018).

22 L. G. Nyúl, J. K. Udupa, and X. Zhang, "New variants of a method of mri scale standardization," *IEEE transactions on medical imaging* **19**(2), 143–150 (2000).

23 X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256 (2010).

24 D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2014).

25 F. Pérez-García, R. Sparks, and S. Ourselin, "TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *arXiv:2003.04696 [cs, eess, stat]* (2020). arXiv: 2003.04696.

26 A. Birenbaum and H. Greenspan, "Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks," in *Deep Learning and Data Labeling for Medical Applications*, 58–67, Springer (2016).

27 T. Ma, H. Zhang, H. Ong, *et al.*, "Ensembling low precision models for binary biomedical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 325–334 (2021).

28 S. Andermatt, S. Pezold, and P. C. Cattin, "Automated segmentation of multiple sclerosis lesions using multi-dimensional gated recurrent units," in *International MICCAI Brainlesion Workshop*, 31–42, Springer (2017).

29 H. Zhang, J. Zhang, R. Wang, *et al.*, "Geometric loss for deep multiple sclerosis lesion segmentation," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 24–28, IEEE (2021).

30 M. Ghafoorian, N. Karssemeijer, T. Heskes, *et al.*, "Deep multi-scale location-aware 3d convolutional neural networks for automated detection of lacunes of presumed vascular origin," *NeuroImage: Clinical* **14**, 391–399 (2017).

31 S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, *et al.*, "Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection," *IEEE Access* **7**, 1721–1735 (2019).

32 S. Aslani, M. Dayan, L. Storelli, *et al.*, "Multi-branch convolutional neural network for multiple sclerosis lesion segmentation," *NeuroImage* **196**, 1–15 (2019).

33 S. Valverde, M. Cabezas, E. Roura, *et al.*, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach," *NeuroImage* **155**, 159–168 (2017).

34 A. Fenneteau, P. Bourdon, D. Helbert, *et al.*, "Investigating efficient cnn architecture for multiple sclerosis lesion segmentation," *Journal of Medical Imaging* **8**(1), 014504 (2021).

List of Figures

3 The frequency of lesions depending on the lesion volume in the in-house dataset in (a) and the frequency of voxels depending on the lesion volume they belong to (b). The volume range is not linear to avoid flatten representation due to a large volume range. The frequencies are calculated on the test set to be comparable with other results.

4 Average performances of the model trained with weighted Dice loss with the set of tested parameters $w_{\min}$ and $range$ for an exam of the in-house test set. Voxel-wise segmentation is measured in Dice score (a), the lesion detection is measured in F1 score (b), in LTPR (c) and in LFPR (d). For each metric, the color blue indicates that the value is inferior to the corresponding metric of the model trained without weighted Dice loss (**Table** 1) and the red color indicates that it is superior.

5 The lesion detection in LTPR (a), LFPR (b) and F1 score (c) depending on the lesion volume in mm^3 of the four best weighting functions in blue with $range \in \{50, 100\}$ and $w_{\min} \in \{1, 2\}$ and without weighting in red. The performances of all tested weighting functions are available in **Figures** 8, 9 and 10 of appendix. Computed on the in-house test set.

6 Visualization of ground truth segmentation and predicted segmentations on an example of the in-house dataset without weighting and with weighting functions with $range \in \{50, 1000\}$ and $w_{\min} \in \{0.1, 2, 6\}$.

7 Profiles of the tested weighting functions ω. $w_{\max}$ is set to 10 and only $w_{\min}$ and $range$ are varying. $range$ sets the distance between the maximal and minimal asymptotes and $w_{\min}$ gives the minimum weighting value associated to the largest lesions.

List of Tables