Scientific Reports

https://doi.org/10.1038/s41598-026-36560-9

Article in Press

Instance-level quantitative saliency in multiple sclerosis lesion segmentation

Received: 22 October 2024

Accepted: 13 January 2026

Published online: 02 February 2026

Cite this article as: Spagnolo F., Molchanova N., Cuadra M.B. *et al.* Instance-level quantitative saliency in multiple sclerosis lesion segmentation. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-36560-9

Federico Spagnolo, Nataliia Molchanova, Meritxell Bach Cuadra, Mario Ocampo-Pineda, Lester Melie-Garcia, Cristina Granziera, Vincent Andrearczyk & Adrien Depeursinge

Instance-level quantitative saliency in multiple sclerosis lesion segmentation

Federico Spagnolo[1,2,3,4], Nataliia Molchanova[4,5,6], Meritxell Bach Cuadra[5,6], Mario Ocampo-Pineda[1,2,3], Lester Melie-Garcia[1,2,3], Cristina Granziera[1,2,3], Vincent Andrearczyk[4,+], and Adrien Depeursinge[4,7,*,+]

[1]Translational Imaging in Neurology (ThINk) Basel, Department of Medicine and Biomedical Engineering, University Hospital Basel and University of Basel, Basel, Switzerland
[2]Department of Neurology, University Hospital Basel, Basel, Switzerland
[3]Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland
[4]MedGIFT, Institute of Informatics, School of Management, HES-SO Valais-Wallis University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland
[5]CIBM Center for Biomedical Imaging, Lausanne, Switzerland
[6]Radiology Department, Lausanne University Hospital (CHUV) and University of Lausanne, Lausanne, Switzerland
[7]Nuclear Medicine and Molecular Imaging Department, Lausanne University Hospital (CHUV), Lausanne, Switzerland
[*]adrien.depeursinge@hevs.ch
[+]these authors contributed equally to this work

ABSTRACT

In recent years, explainable methods for artificial intelligence (XAI) have tried to reveal and describe models' decision mechanisms in the case of classification and even for segmentation. However, XAI methods for semantic segmentation and in particular for single specific instances (e.g. one given lesion among others of the same class in medical imaging) have yet to be developed to understand what drove the detection and contouring of the latter, which is crucial for all multi-lesional diseases.
We proposed instance-level explanation maps for semantic segmentation extending both SmoothGrad and Grad-CAM++ methods and yielding quantitative instance saliency for the former. The instance-level methods were applied to the segmentation of white matter lesions (WML), a magnetic resonance imaging (MRI) biomarker in multiple sclerosis (MS). 687 patients diagnosed with MS for a total of 4023 FLAIR and MPRAGE MRI scans were collected at the University Hospital of Basel, Switzerland. WM lesion masks were annotated by four expert clinicians on baseline and follow-up imaging. Three deep learning networks —a 3D U-Net, nnU-Net, and Swin UNETR— were trained and tested on these data (test normalized Dice score, respectively of 0.71, 0.78, 0.80; true positive rate of 79%, 78%, and 85%; false discovery rate of 37%, 38%, and 36%; false negative rate of 20%, 22%, and 14%), then saliency maps were computed.
Consistent with clinical practice, the proposed instance saliency maps revealed that the model relied more on FLAIR than MPRAGE to segment WMLs, with positive saliency values inside a lesion and negative in its neighborhood. FLAIR hyperintensity combined with healthy WM around the lesion border was required for their detection. Beyond the aforementioned sanity checks, we observed that peak values of the generated saliency maps presented distributions that differ significantly between TP, FN, FP and TN predictions, suggesting that the quantitative nature of the proposed saliency could be used to identify errors.
In conclusion, we introduced two XAI methods to generate quantitative instance-level explanations in semantic segmentation. The proposed XAI maps can be applied to any architecture and could serve as a basis to (i) improve model performance (e.g. reducing FPs), (ii) optimize their internal architecture (e.g. patch size), and (iii) justify the model's decisions to the end users, which are contextualized to a specific lesion instance of interest.

Introduction

Multiple sclerosis (MS) is an autoimmune neurological disease, which affects people at a relatively young age, presenting a considerable impact on the quality of life [1]. One of the most important biomarkers in MS are white matter (WM) lesions reported on magnetic resonance imaging (MRI) [2]. The standard MRI sequences used for MS diagnosis and follow-up [3] are the fluid attenuated inversion recovery (FLAIR) and T1-weighted (T1-w) contrast, such as the magnetisation-prepared rapid gradient echo (MPRAGE) [4]. WM lesions generally appear hyperintense on FLAIR (except for cavitary lesions [5]), and a subset with greater demyelination and tissue damage appears hypointense on T1-weighted images [6].

These lesions are usually manually or semi-automatically annotated by clinicians with several years of experience, through a time-consuming process, subject to inter-observer variations. Despite many efforts to automate the process of lesion detection and segmentation with deep learning (DL) methods [7, 8, 9, 10, 11], their clinical integration is being jeopardized by two main issues:

1. The "black box" nature of the models [12]. Since these methods contain many layers and millions of parameters it is hard to interpret and explain which are the drivers for a particular decision, i.e., which voxels were more important to identify and segment a given lesion of interest.
2. Insufficient clinical validation of the models, described in Spagnolo et al. [13].

To address the first issue, research in explainable AI (XAI) may play a decisive role. XAI has the potential to support trustworthy AI via better understanding (e.g., decision rules, biases) and optimization of DL models [14]. However, XAI for semantic segmentation and in particular for detecting and contouring single instances of interest has been little studied to date. Semantic segmentation is a computer vision task, where labels are associated with every pixel of an image. In the case of WM lesions, such as depicted in Fig.1a with four distinct lesion instances, the segmented plaques appear as disconnected volumes in the MRI. In general, objects may appear as connected, and can be segmented into separate entities through instance segmentation. Instance segmentation can extract supplementary information from the image, such as the number of objects of the same class (as lesion one to four in Fig.1b). Treating separate instances would be crucial to understand the mechanisms underlying automatic detection and segmentation of a given object of interest (e.g., a lesion in medical imaging or a person in a natural image) among other objects of the same class. This also applies to the case of MS lesions, where plaques can be subdivided based on their location in the brain, or the stage of the disease [15]. Specifically, considering separate instances would be important to generate instance-specific explanations not only for diagnosis at single time-points, but also for disease monitoring in follow-ups.

Introducing instance-level explanation methods could facilitate a better understanding of AI's segmentation of single instances. In addition to the case of MS, such methods could be applicable to any pathology involving sparse lesions, beyond MRI, and possibly to any segmentation task.

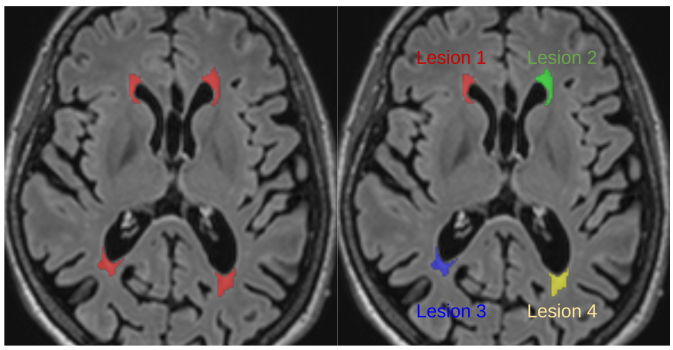

Figure 1. (a) FLAIR MRI presenting WM lesions segmented as separate entities, and (b) example of instance segmentation inspired by Varatharasan et al. [16].

An exhaustive review of XAI models and applications can be found in Saranya et al. [17]. For convolutional neural networks (CNN) in a classification scenario, a widely used ad-hoc method is pixel attribution (or saliency maps), in which pixels are

colored based on their contribution to the classification. To this end, vanilla gradients [18] is a method based on forward and backward propagation through the network. First, an input image (for instance a 3D volume) is fed forward through the network and a class score is computed. Then, the gradient of the score with respect to the input — or a layer — is calculated backwards to form a map, which represents positive and negative contributions of input voxels to the classification of the image. Maps generated with this method are easy to compute and visualize, but also noisy and sensitive to small changes in the input [19]. To partially address these problems, Smilkov et al. [20] introduced a method called SmoothGrad (SG). A more stable output is obtained by feeding multiple noisy versions of the input image to the network and averaging the obtained saliency maps. Some recent applications can be found in Goh et al. [21] and Agarwal et al. [22].

Another widely used XAI method for classification, is Grad-CAM [23]. The gradients of a class score — with respect to activation maps of a given layer — are spatially aggregated through global average pooling. This way, a weight is computed for each activation map, representing its relevance. These weights are then employed to calculate the visualization heatmap as a linear combination of the activation maps, followed by a ReLU to focus on features positively influencing the class score. However, this method presented a low accuracy when the image contained multiple instances of the same class. To overcome this, Chattopadhyay et al. [24] described Grad-CAM++, where the weights are obtained through a weighted average of the gradients.

SG and Grad-CAM++ can be seen as complementary methods. The first provides local-level information and works well in identifying the impact of multiple input channels (or modalities), since the gradients flow all the way back to the inputs. Yet, as mentioned above, it may be too sensitive to changes in the input. The second can generate more stable heatmaps, and probe specific layers of the network. However, it also bears its disadvantages: 1) the choice of the intermediate layer can be not trivial, 2) the potential presence of early skip-connections in the architecture would be disregarded, 3) stopping the backpropagation at an intermediate layer would make it impossible to determine the impact of the input on the output.

Both algorithms were originally designed for classification tasks, such as in Rajpurkar et al. [25]. This means that, in general, the output y of a multi-class classification network is a vector of C scalar values, where C is the number of classes. In the case of semantic segmentation, y is a set of C tensors (2D or 3D images) containing segmentation scores for each class (Fig. 2).

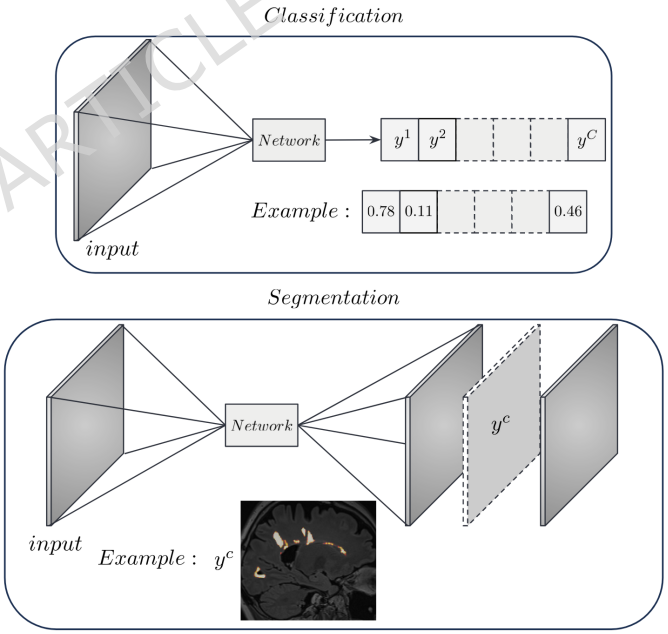

Figure 2. Input-output dimensions for a classification (top) and a semantic segmentation (bottom) task.

Indeed, computing a gradient map for all the output voxels would not be convenient or meaningful in a segmentation scenario. The number of maps per patient would be excessive, thus impractical for the end-user, especially for a clinician. Additionally, the computation time would be unnecessarily high.

A straightforward way to adapt explainable methods to semantic segmentation would be, for a given class c, to aggregate all the spatial predictions y^c into a single scalar (e.g., a summation) and compute its gradients. However, this approach yields confusing and hardly interpretable maps, where the relevance of input voxels to the segmentation scores of all output voxels (even those not segmented as part of class c) are merged together, as reported in Fig. 3. This makes it more complicated to understand and determine which part of the output segmentation is influenced by which region of the input. That is especially important when there are multiple objects belonging to the same class, so that input intensity of neighboring voxels or objects may impact the segmentation. Indeed, these results do not provide any kind of explanation to the segmentation of a particular instance of the considered class c.

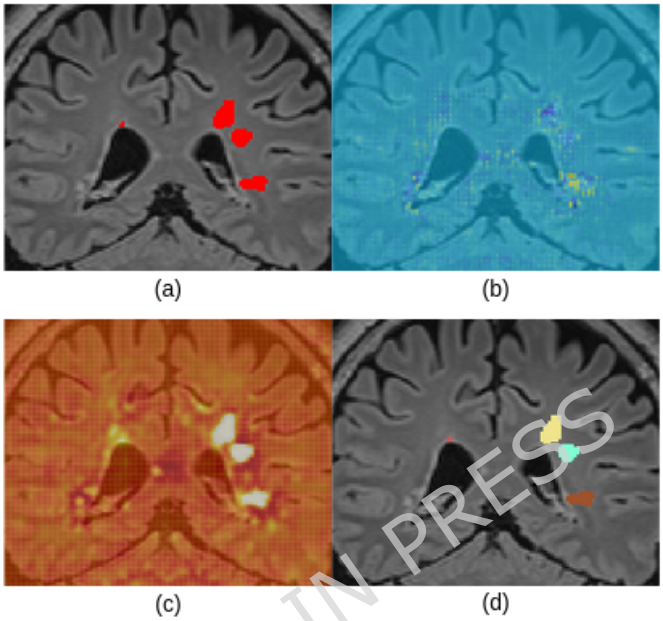

Figure 3. (a) The output of a semantic segmentation network showing several instances of the considered class. SmoothGrad (b) and Grad-CAM (c) applied to all the spatial predictions. (d) How can we explain the segmentation of a particular lesion of interest (e.g., the yellow instance)?

Recent works [26, 27] clearly state that saliency maps were not initially developed for segmentation tasks and, therefore, no explicit methods are currently available to do so. In Arun et al. [27], this statement was used to warn the readers about potential problems related to the misuse of XAI in clinical practice, and proposed pixel-level metrics to evaluate saliency maps. Singh et al. [28] applied saliency maps to classification networks, while adopting uncertainty maps to assess the confidence of their skin lesion segmentation method.

First steps towards the use of XAI in segmentation were taken in the work of Wickstrom et al. [29]. They used guided backpropagation [18] to generate saliency maps for the explanation of colorectal polyp segmentation. Their saliency maps were obtained by aggregating all the positive spatial predictions. Similarly, Vinogradova et al. [30] generated heatmaps by using the first version of Grad-CAM [23], and aggregating only output voxels segmented as part of a target class c. The authors employed a U-Net and the dataset Cityscapes [31] to perform semantic segmentation.

While these approaches lead to a possible class-level explanation for semantic segmentation, they still do not provide any instance-level information, such as which input voxels were exploited to segment a specific instance. In addition, these methods do not provide quantitative saliency maps, which means these maps do not allow to interpret their absolute values across images or objects of interest to, e.g., distinguish false positives or false negatives from true positives.

To address the aforementioned limitations, this paper presents the adaptation of SG and Grad-CAM++ to obtain quantitative and instance-level explanation maps, and their application to WM lesion segmentation in MS.

The quantitative characteristic of the saliency method described in Section Quantitative saliency maps: maximum versus average aggregation was exploited in Spagnolo et al. [32], using the same MRI data. Radiomic features extracted from the

proposed instance-level XAI maps were fed to a simple logistic regression model to refine the classification of TP and FP examples. In that study, a bootstrapping approach with 1000 iterations was used to compute the 95% confidence intervals, highlighting an F1 score relative improvement on the test set.

Methods

Dataset and models

687 patients diagnosed with MS for a total of 4023 FLAIR and MPRAGE MRI scans (age=45.2±12.2, 433 females, SIEMENS Avanto/Espree/Symphony 1.5T and Prisma/Skyra/Verio/MAGNETOM Vida 3T, 1*mm* isotropic, Expanded Disability Status Scale median of 2.5 [0-9]) were collected at the University Hospital of Basel, Switzerland [33]. The study received ethical approval by the local independent ethics committee (Ethikkommission Nordwest- und Zentralschweiz, EKNZ), all patients provided written informed consent, and all methods were carried out by relevant guidelines and regulations.

The image size of both MR contrasts is $(192 \times 240 \times 256)$, which corresponds to a volume of $(192 \times 240 \times 256)$ mm^3. Table **??** in the Supplementary Material reports data information for each MR system, such as age, sex and number of visits. For consistency reasons, patients were mainly scanned first using Avanto 1.5T and then Skyra 3T, operating other MR systems only in case of unavailability of the two mentioned.

WM lesion masks from baseline and follow-ups were semi-automatically annotated by four expert clinicians (>5 years of experience), independently and without consensus. The binary lesion masks corrected by the experts were generated by a U-Net variant described in La Rosa et al. (2020)[34] (since that variant was trained on FLAIR and MP2RAGE data, we retrained it on separate proprietary FLAIR and MPRAGE data before the inference). Data were randomly split into training, validation and test sets (containing 560, 90 and 37 patients with 3369, 553 and 101 scans, respectively; training, validation and hold-out test set's mean lesions number of 52.0 ± 36.3, 56.2 ± 35.3, and 42.3 ± 21.4 per patient) to train and evaluate a 3D U-Net [35] variant (different from the one in La Rosa et al [34]), an nnU-Net [36], and a Swin UNETR [37] for WM lesion segmentation, using patches of dimensions 96^3 to ensure the inclusion of at least part of the brain in every patch. The split was performed at patient level to ensure that images from the same patient belong to the same split. We adopted a linear combination of normalized Dice [38] and blob [39] losses to tackle instance imbalance within a class and bias towards the occurrence of positive class [40]. Pre-processing steps included the registration of FLAIR images to MPRAGE space using the *elastix* toolbox [41, 42], N4 bias field inhomogeneity correction [43] and z-score intensity normalisation.

Notations

Following Depeursinge et al. (2020) [44], we noted a discrete image as a D-dimensional function of the variable $v = (v_1, \ldots, v_D) \in \mathbb{Z}^D$, taking values $x[v] \in \mathbb{R}$. A subset Γ of $\mathbb{Z}^D$ was considered in practice for the spatial image domain with dimensions $N_1 \times \cdots \times N_D$ as possible values for the index vector $v \in \Gamma$. We also referred to the lesion domain Ω as a subset of the image domain with cardinality (i.e. number of voxels) $|\Omega|$, such that $\Omega \subset \Gamma \subset \mathbb{Z}^D$.

Input images $x[v]$ performing a forward pass through the network resulted in logits $y(x)[v] \in \mathbb{R}$, where $v \in \Gamma$ is a map of raw output values, which had the exact same dimensions as the input values $x[v]$ since we considered a segmentation task. We use the simplified notation $y[v]$ when the input x is unambiguous. This raw output was generally interpreted as a probability map (e.g., after Softmax), which was binarized using a threshold $t = 0.3$ yielding the best normalized Dice score during validation. Various thresholds t were explored, ranging from 0.1 to 0.5, and the one selected scored about 2% better than the second best. Then, each connected component of the binary map was considered as a specific WM lesion instance, forming Ω.

Gradient-based saliency maps

During the backward pass and for each voxel of $y(x)[v]$, the gradients with respect to all the voxels of input values $x[v]$ can be computed. These gradients can be visualized as an image with same dimensions as the input and output, constituting a first method to construct saliency maps. This approach is usually referred to as vanilla gradients [18].

The SG algorithm [20] tackled the problem of instability of vanilla gradients maps: the gradient of logits $y(x_n)$ is computed N times based on artificially noised versions of the input $x_n[v]$. The authors demonstrated that the average M of these maps is more stable than vanilla gradients maps:

$$M[v] = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial y(x_n)}{\partial x_n[v]}. \tag{1}$$

Instance-level saliency (gradients)

This approach was introduced in a classification paradigm. However, with segmentation models, there are predictions for each output voxel v. As a result, the visualisation of many heatmaps for a single output voxel is neither convenient nor meaningful. Hence, we adapted the original method to the segmentation task by aggregating these heatmaps. For a given lesion, the implementation consists of:

1. Injecting a Gaussian noise $\mathcal{N}(0,\sigma)$ with standard deviation σ to obtain a noisy version of the input,
2. Computing a collection of saliency maps for all output voxels in the domain Ω of the lesion,
3. Determining the average map from this collection of maps,
4. Repeating steps 1-3 and combining $N = 50$ saliency maps ($\sigma = 0.05$) from N noisy versions to obtain a single one.

Steps 2 and 3 are illustrated in Fig. 4. Eq.(2) details the computation of lesion-level saliency maps $M_\Omega^{\text{gradient}}[v] \in \mathbb{R}$ combining gradients calculated from each output voxel of a lesion. A separate saliency map is generated for each input modality (two in our case) allowing us to investigate their respective contribution.

$$M_\Omega^{\text{gradient}}[v] = \frac{1}{N|\Omega|}\sum_{n=1}^{N}\sum_{v'\in\Omega}\frac{\partial y(x_n)[v']}{\partial x_n[v]} \tag{2}$$

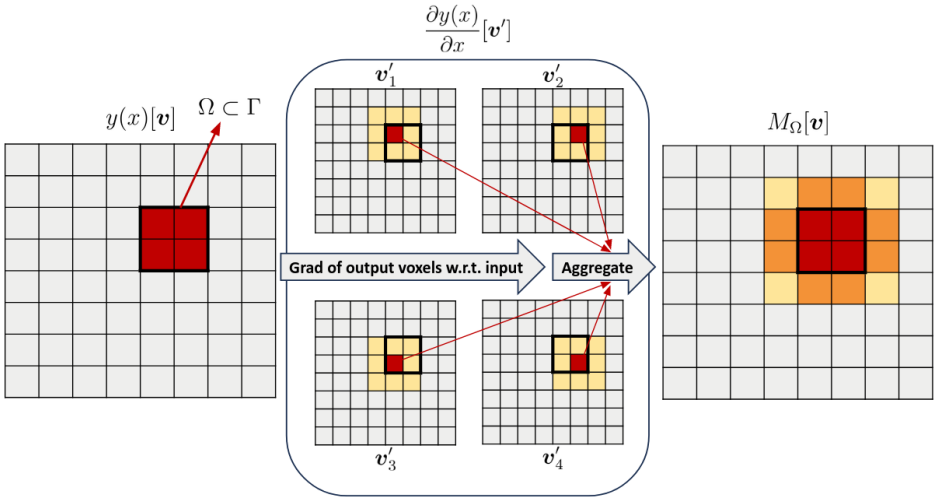

Figure 4. Overview of the proposed adaptation of SG to segmentation.

Quantitative saliency maps: maximum versus average aggregation

The advantage of $M_\Omega^{\text{gradient}}[v]$ is that it shows whether voxels outside the lesion domain Ω impact the prediction of voxels belonging to Ω. However, the lesion domain's dimensions may range from a few voxels to a considerable volume. The greater the lesion size, the more extended the potential distance between two voxels $p, q \in \Omega$. A long-distance means that the saliency map generated for the voxel p will present a low gradient value for q. The same principle applies to regions of the input far away from a lesion: their contribution to the prediction is low. As a consequence, the average over the lesion domain Ω in Eq.(2) will cause gradient values in $M_\Omega^{\text{gradient}}[v]$ for extensive lesions to be systematically lower. Thus, following this method, the voxel values in saliency maps generated for different Ω (lesions) would not be comparable.

To this end, we propose Eq.(3), a slightly modified version of Eq.(2) where the average of saliency maps generated from each element of Ω is replaced by the voxel-wise maximum with sign:

$$M_\Omega^{\text{gradient}}[v] = \frac{1}{N}\sum_{n=1}^{N} D^n_{argmax_{v'}|D^n_{v'}|}, \text{ where } D^n_{v'} = \frac{\partial y(x_n)[v']}{\partial x_n[v]} \tag{3}$$

Saliency maps based on Grad-CAM++

The second proposed method is based on Grad-CAM++ [24]. We generated, for a given layer of a network, a heatmap M as a linear combination between weights $\{\omega^k\}_{k=1}^K$ and activation maps $\{A^k\}_{k=1}^K$. The activation maps of the selected layer may have different dimensions. In this case, the final heatmap is upsampled to the input dimensions. For the sake of simplicity, we defined the domain of the activation maps and that of the input image to be the same (i.e. activations from the last layer since we have a segmentation architecture): Γ, with values $A^K[v] \in \mathbb{R}$. Following Vinogradova et al. [30], to compute the gradients, we considered y' as the sum of logits $y[v]$ higher than a threshold t, as in Eq.(5). Then, each weight ω^k is computed using the gradients of y' - with respect to the k^{th} activation map A^k - and a coefficient $\alpha^k[v]$. We have

$$M^{\text{GradCAM}}[v] = Relu\left(\sum_k \omega^k \cdot A^k[v]\right), \tag{4}$$

$$y' = \sum_{v|y[v]>t} y[v], \tag{5}$$

$$\omega^k = \sum_{v\in\Gamma} \alpha^k[v] \cdot Relu\left(\frac{\partial y'}{\partial A^k[v]}\right), \tag{6}$$

$$\alpha^k[v] = \frac{\frac{\partial^2 y'}{\partial \left(A^k[v]\right)^2}}{2\cdot\frac{\partial^2 y'}{\partial \left(A^k[v]\right)^2} + \sum_{v'\in\Gamma}\left(A^k[v']\cdot\frac{\partial^3 y'}{\partial \left(A^k[v]\right)^3}\right)}, \tag{7}$$

where v and v' correspond to a different indexing over the domain Γ. The above eqs. (4) to (7) are derived from Chattopadhyay et al. (2018) [24]. A graphical representation of this method is reported in Fig. 5.

Instance-level saliency (Grad-CAM++)

The simple adaptation of Grad-CAM++ to segmentation presented above generated a class-level explanation for semantic segmentation by merging contributions to different lesion instances. The impact of input regions on different parts of the output was combined by:

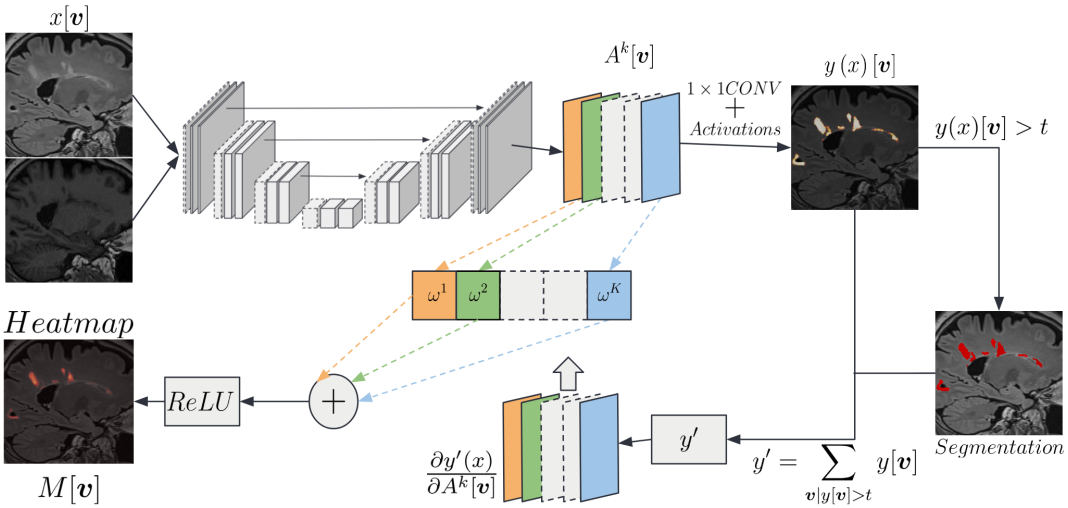

Figure 5. Overview of Grad-CAM++, generating a class-level explanation heatmap, similarly to Vinogradova et al. [30], adapted for segmentation.

- Considering the gradients of a subset y' of logits $y[v]$,
- Assigning a single weight ω^k for each feature map A^k.

However, it would be useful to know which input voxels influenced the segmentation of a given instance (e.g., a lesion). To adapt the algorithm to an instance-level explanation, we considered two steps. First, the gradients of y were computed from the domain Ω of one lesion, as in Eq.(9). Then, the summation in Eq.(6) over Γ to compute weights was removed to retain a weight for each element of a feature map. This was needed to prevent the activation of other instances to emerge, and to select only the activation in Ω. Eq.(8) and Eq.(10) represent the proposed heatmap M_l^{GradCAM} provided by the modified Grad-CAM++ method, and the weights $\omega^k[v]$:

$$M_\Omega^{\text{GradCAM}}[v] = Relu\left(\sum_k \omega^k[v] \cdot A^k[v]\right), \tag{8}$$

$$y' = \sum_{v \in \Omega} y[v], \tag{9}$$

$$\omega^k[v] = \alpha^k[v] \cdot Relu\left(\frac{\partial y'}{\partial A^k[v]}\right), \tag{10}$$

where $\alpha^k[v]$ were obtained as in Eq. (7). An overview of the proposed adaptation of Grad-CAM++ to segmentation is illustrated in Fig. 6. Comparing Figs 5 and 6, it can be observed that the output map of the first is a valuable explanation for the entire lesion class, combining the attribution of voxels from different lesions in the same map. Conversely, the output of the second separates the attribution of voxels from single lesions to obtain a more meaningful map, in cases where each instance should be treated independently.

The code for the computation of explainable maps through both methods is publicly available at the following GitHub repository: https://github.com/federicospagnolo/IES.git.

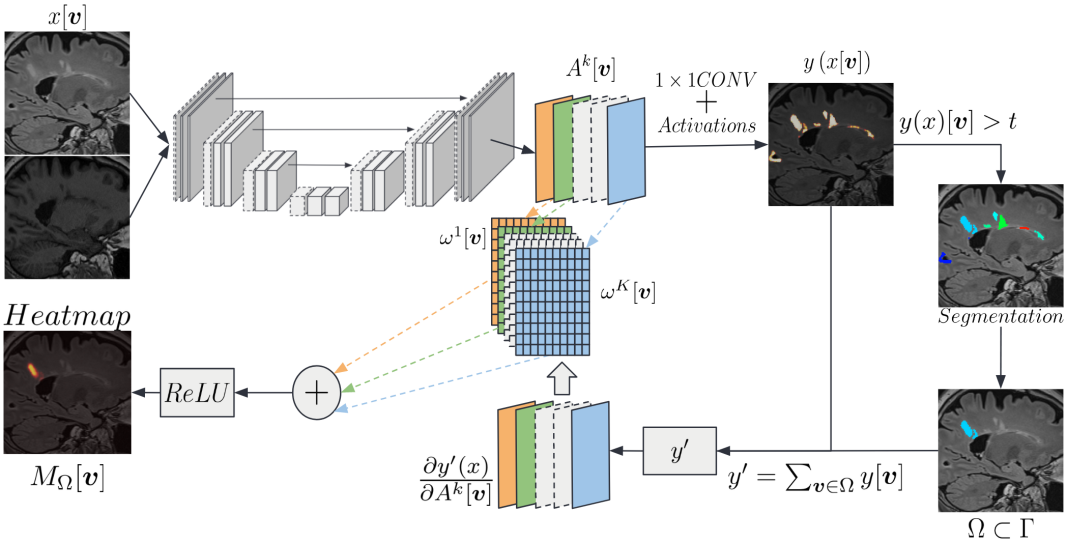

Figure 6. Overview of the proposed adaptation of Grad-CAM++, generating an instance-level explanation heatmap.

Experiments

The described methodologies were applied to all the three selected architectures (U-Net, nnU-Net and Swin UNETR), obtaining a collection of saliency maps. The XAI maps presented below were computed using the U-Net for a total of 3050 true positive (TP), 1818 false positive (FP), 789 false negative (FN), and 1010 true negative volumes (TN). Examples of saliency maps derived from nnU-Net and Swin UNETR are included in the Supplementary Material. TP and FP predictions were defined as having, respectively, a non-zero and zero overlap with ground truth (GT). FN predictions were considered as GT segmentations with zero overlap with the predicted lesion mask. TN examples were generated by randomly sampling ten spherical volumes ($93mm^3$) inside each patient's brain and skull, excluding volumes intersecting GT and prediction masks. The size of the TN volumes was decided based on the average lesion volume in the GT masks of the test set. In this study, when mentioning TP, FP, FN and TN we refer to examples, i.e. volumes, not voxels. While TN voxels represent most of the image, we tackle class imbalance by undersampling, randomly selecting ten spherical volumes per patient.

As for the minimum considered lesion size, the McDonald's criteria recommend a minimum in-plane diameter of 3 *mm* based on old 2D sequences [45, 46], and in literature there is no clear consensus. As a trade off between limiting partial volume effects and the number of false positives [47], a minimum volume size of $5mm^3$ was set for each connected component, using a connectivity of 18. Several publications have selected a similar minimum lesion volume, for example: 1) De Rosa et al. (2024)[48] considered 0.005 *ml* (5 mm^3); 2) Fartaria et al. (2019)[47] used 0.006 *ml*; 3) Jain et al. (2016)[49] removed candidates with a volume smaller than 0.005 *ml*.

The information exploited by the U-Net during inference for the segmentation of specific WM lesions was assessed with several experiments. Throughout these tests, the max aggregation method (see Section Quantitative saliency maps: maximum versus average aggregation) was selected to be used, due to its ability to provide quantitative information. This only applies to the saliency based on SG.

First (results in Section Assessing the contributions of input MR sequences using gradient-based saliency maps), the distribution of positive and negative values in saliency maps was observed to reveal potential patterns concerning recurring locations of positive/negative values with respect to Ω. An analysis of these distributions allowed a statistical comparison between gradient values computed with respect to FLAIR and MPRAGE. As a consequence, it was possible to isolate the contribution of both input sequences to the prediction of single lesions. In this analysis, we discarded gradient values between -0.1 and 0.1 in order to focus on voxels with a higher attention level.

Second (results in Section Statistical distribution of gradient values for TPs, FPs, FNs and TNs), we quantitatively compared the distribution of maximum and minimum saliency maps' values, for all predictions categories (i.e. TP, FP, FN and TN) to investigate if absolute saliency values can be used to flag potential detection errors. Furthermore, a two-sided Mann Whitney U

test [50] was run to statistically compare these groups.

Third (results in Section Sanity checks), we tested the saliency method's behavior in two cases: (a) positioning a domain Ω in a region presenting no MS lesions (i.e. healthy WM); (b) considering a domain Ω of a single voxel, located at the center of mass of a true lesion. For both cases, we show the generated instance-level saliency and its range of values.

To understand the U-Net's level of specificity in segmenting MS lesions in the WM, we designed three additional qualitative experiments on a batch of 10 patients, in which we observed the U-Net's behavior on synthetic lesions (results in Section Sanity checks). A first experiment was conducted, moving a clearly visible WM lesion to a different part of the WM, which originally presented no lesions. A similar approach was followed when inserting the same lesion outside the skull. A third experiment consisted of a compromise: the lesion, along with part of its surrounding healthy tissue (3*mm* from lesion border), was moved outside the skull. For all these cases, the U-Net's prediction and saliency maps were examined.

In light of the results observed during the described tests, we designed a more specific experiment on the entire test set, using all the three trained networks: the analysis of the needed amount of contextual information surrounding a lesion to obtain a segmentation (results in Section Experiment on contextual information). We selected lesions with size close to the average of the entire dataset, i.e. from 90 to 120mm^3, obtaining a total of 191 (U-Net), 173 (nnU-Net), and 161 (Swin UNETR) TP lesions. Initially, all voxel intensities were set to zero, but those of the lesion: this step was called iteration 0. Then, we gradually reassigned the original intensity to surrounding voxels through morphological 3D dilation, iterating this process 35 times. With the final iteration, the models were seeing a volume of surrounding tissue at a maximum of 35*mm* distance from the lesion's edges. At each iteration, the following metrics were recorded:

- The average and standard deviation across lesions of the mean prediction score (after Softmax) in Ω,
- The number of segmented lesions.

In this experiment, each lesion was considered detected when at least one voxel in Ω reached a prediction score above the threshold $t = 0.3$.

Results

The trained lesion segmentation models —U-Net, nnU-Net, Swin UNETR— achieved, respectively, a test Dice score of 0.60, 0.62, 0.66, a normalized Dice score of 0.71, 0.78, 0.80, a true positive rate of 79%, 78%, and 85%, a false discovery rate of 37%, 38%, and 36%, a false negative rate of 20%, 22%, and 14%. The count of TP, FP and FN examples was 3050, 1818, 789 for the U-Net, 3112, 1954, 880 for the nnU-Net, and 3516, 2018, 598 for the Swin UNETR.

Maximum versus average aggregation for quantitative saliency maps

Considering only saliency based on SG, we compared the saliency maps obtained with either the average (Eq.(2)) or the maximum (Eq.(3)) aggregation method. As reported for an example lesion in Fig. 7, the saliency maps generated with the maximum presented values that were comparable to those obtained for smaller lesions while preserving the proportion of positive and negative values according to the lesion properties. In the case of the method based on SG, the results in the following sections are obtained using the max aggregation method.

Assessing the contributions of input MR sequences using gradient-based saliency maps

Saliency maps (based on SG) generated with respect to FLAIR for a true positive (TP) lesion are presented in Fig. 8. Positive gradient values appeared to accumulate inside the targeted lesion domain Ω and its edges, while negative values populated its neighborhood. Following the same procedure with respect to MPRAGE, we observed an opposite trend (Fig. 9): negative values in Ω and positive around its borders. In both cases, the values dropped to zero when looking at a distance of $\sim 44mm$ from Ω's borders, and presented close-to-zero values for other lesions in the same brain regions as Ω. The latter was also observed in the

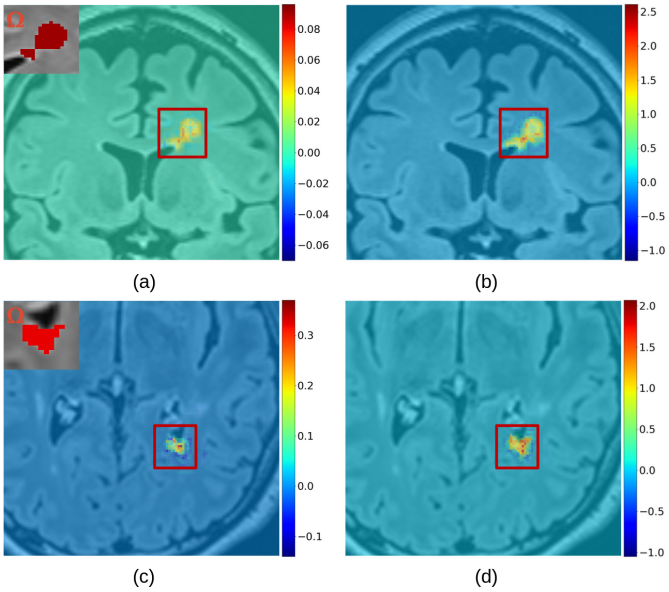

Figure 7. Saliency maps (computed for U-Net) based on SG generated for an extensive lesion, following the average (a) and the maximum (b) methodology. Saliency maps generated for a smaller lesion, following the average (c) and the maximum (d) methodology. The gradients intensity difference between (a,c) and (b,d) can be observed from the color scales range next to each map, increasing approximately from (-0.1, 0.3) to (-1, 2.5). The thumbnails illustrate the lesion domains Ω used during the computation.

instance-level explanation method based on Grad-CAM++ (see Fig. 10). For TP cases, we observed that positive values of gradients (median and 95% confidence interval (CI) of 0.50140 ± 0.00072) computed with respect to FLAIR were consistently greater (in absolute value) than negative values for MPRAGE (median and 95% CI of -0.19584 ± 0.00031).

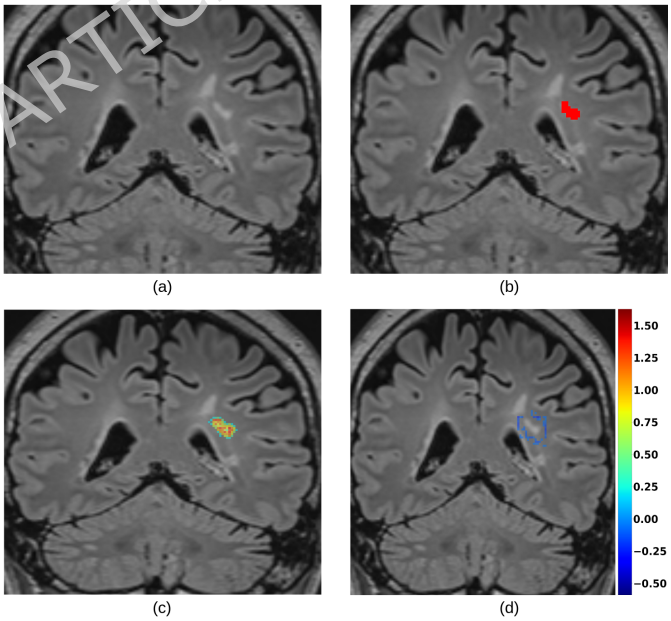

Figure 8. Example of a FLAIR image (a), the lesion domain Ω (b), and the corresponding saliency map (computed for U-Net) obtained with the proposed adaptation of SmoothGrad isolating positive (c) and negative (d) gradients. Values in [-0.05, 0.2] are not displayed to focus on most significant saliency.

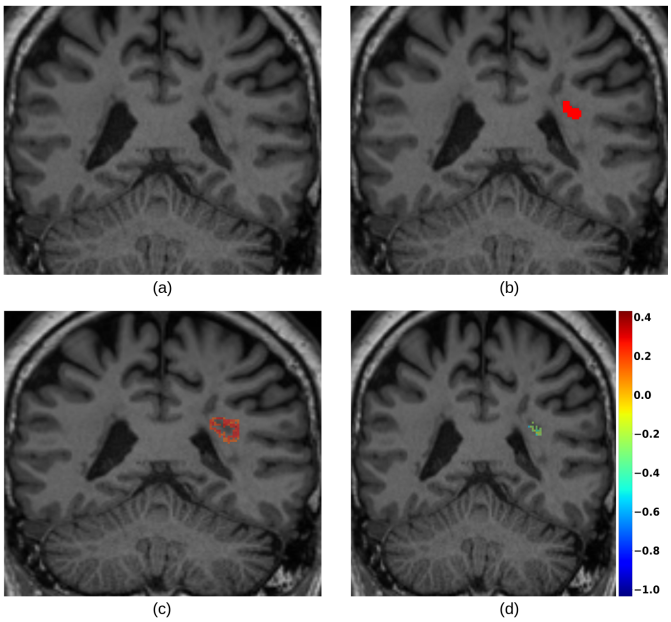

Figure 9. Example of an MPRAGE image (a), the lesion domain Ω (b), and the corresponding saliency map (computed for U-Net) obtained with the proposed adaptation of SmoothGrad isolating positive (c) and negative (d) gradients. Values in [-0.1, 0.1] are not displayed to focus on the most significant saliency.

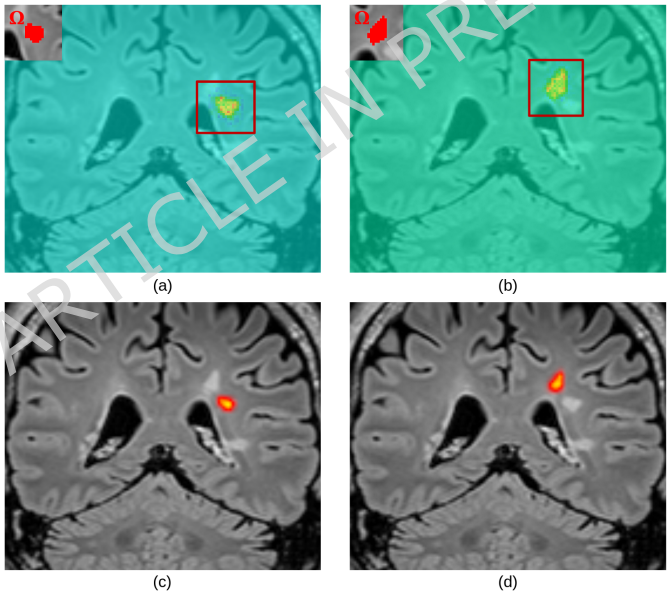

Figure 10. Saliency maps (computed for U-Net) obtained with the proposed adaptation of SmoothGrad for two close lesions (a) and (b). Heatmaps obtained with the proposed adaptation of Grad-CAM++ for the same two lesions (c) and (d). The thumbnails illustrate the lesion domains Ω used during the computation.

Statistical distribution of gradient values for TPs, FPs, FNs and TNs

Fig. 11 reports the distribution of maximum (a) and minimum (b) gradients values — with respect to FLAIR — for TN, FN, FP and TP volumes. The median value and 95% CI for positive values in each group of volumes were, respectively: 0.465 ± 0.025 (TN), 1.202 ± 0.036 (FN), 1.630 ± 0.019 (FP), and 1.997 ± 0.016 (TP). The median value and 95% CI for negative values were, respectively: -0.245 ± 0.008 (TN), -0.359 ± 0.010 (FN), -0.522 ± 0.008 (FP), and -0.639 ± 0.009 (TP). The Mann Whitney U tests run on all pairs of groups reported a p-value < 0.001.

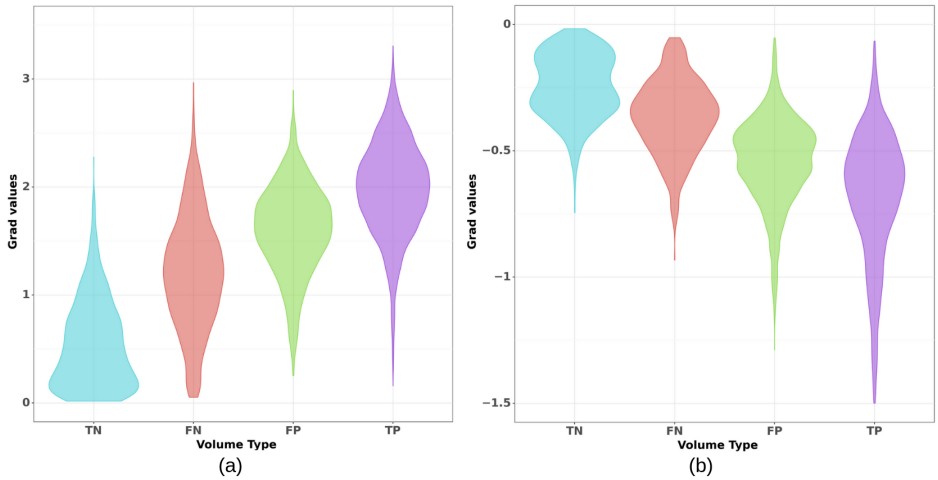

Figure 11. Violin plots representing the distribution of saliency maps (computed for U-Net) maximum (a) and minimum (b) values. The four distributions refer to TN, FN, FP and TP volumes.

Sanity checks

The saliency map generated from a volume (in the WM) with no lesions resulted in gradient values in the range of TN examples (Fig. 11), that is about 5 times smaller than the average one obtained in TP cases. An example of this finding is illustrated in Fig. 12.

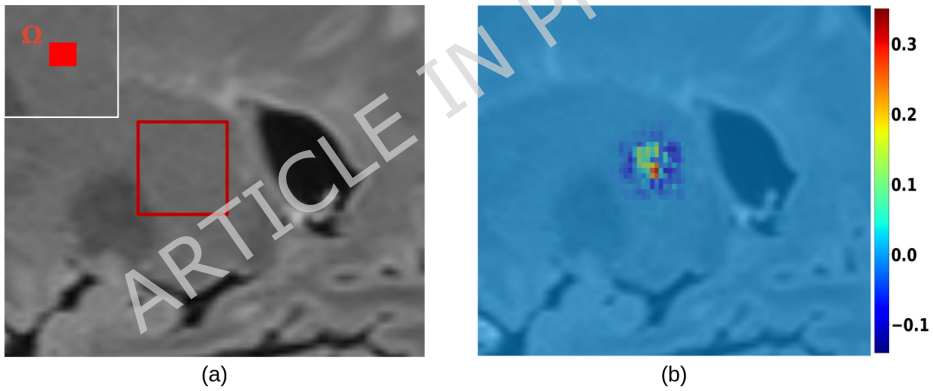

Figure 12. A region in FLAIR with healthy white matter (a) and the corresponding instance-level saliency map (computed for U-Net) obtained with the proposed adaptation of SmoothGrad (b). The color scale range is (-0.1, 0.3), well below the one in Figures 7b and 7d. The thumbnail illustrates the domain Ω used during the computation.

The case of a FLAIR with a single voxel domain Ω, is shown in Fig. 13. Saliency values suggest the prediction of Ω is attributed to input voxels which are in the vicinity, and within the lesion. It appears clear that lesion voxels, in the input, which are farther from Ω present a lower contribution (as it was noticed when using the average aggregation method). However, not only the input voxel corresponding to Ω has an influence.

When inserting a synthetic lesion in a healthy region of the WM, we obtained scores higher than 0.3, enough to trigger its segmentation. In this case, the saliency map (based on SG) resembled those obtained for true WM lesions. However, when the same lesion was placed outside the skull, the lesion domain presented scores below the threshold after the Softmax activation and, thus, the lesion was not detected. In the saliency map, we observed a few positive peak values, but the rest of the lesion volume had negative or close-to-zero gradients. Similarly to the second case, when the lesion and part of its surrounding tissue were moved outside the skull, we noticed low prediction scores. The saliency map, however, presented higher peak values than the second experiment. The prediction scores after the Softmax and the saliency maps for all these three cases were reported in Fig. 14.

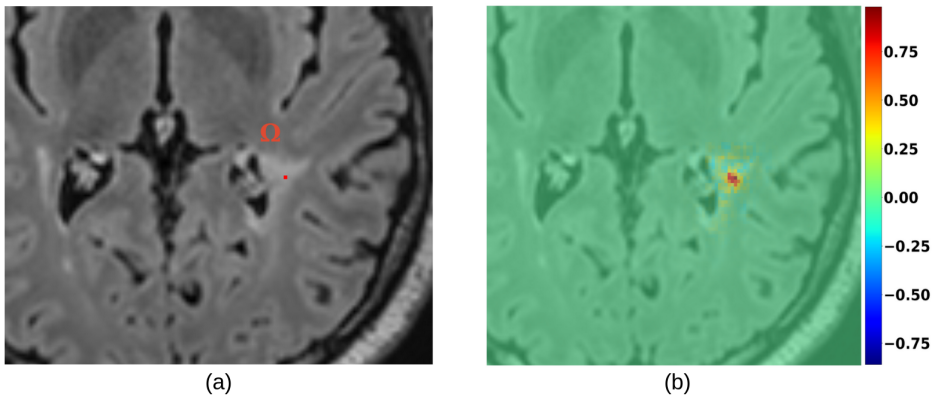

Figure 13. Axial view of an MS lesion, highlighting in red its center of mass used as domain Ω (a), and the corresponding voxel-level saliency map (computed for U-Net) obtained with the proposed adaptation of SmoothGrad (b).

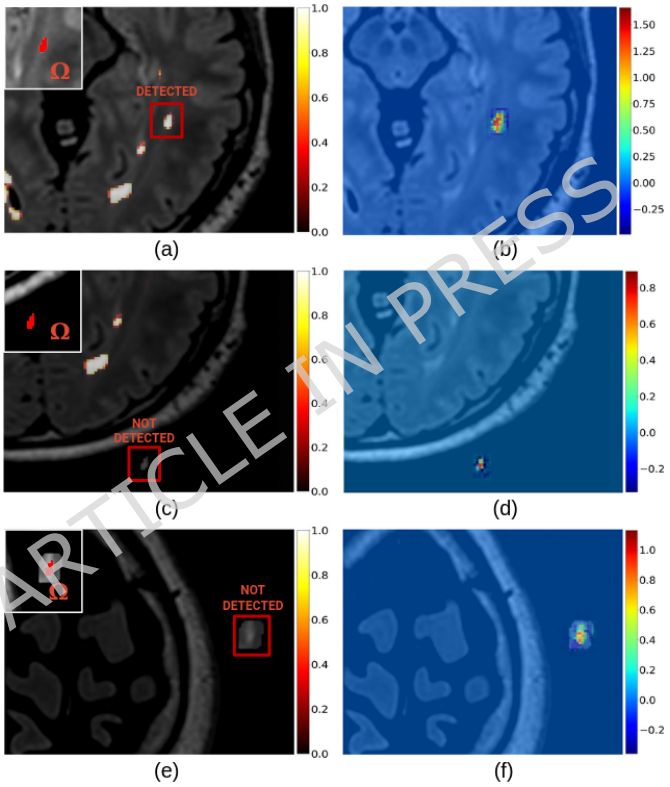

Figure 14. The prediction score before the Softmax for the case of a lesion artificially placed in the WM (a), in the background without (c) and with (e) a part of the surrounding WM. The corresponding saliency maps (b), (d) and (f), respectively. Only the synthetic lesion in the WM triggered a detection.

Experiment on contextual information

The experiment on contextual information showed that, compared to a lesion without any surroundings (i.e. constant background values outside the lesion mask), the prediction score for a lesion increased when including voxels belonging to its perilesional healthy tissue. For U-Net and nnU-Net, the prediction score reached a similar plateau when adding tissue distant $12 - 15mm$ from the lesion border. A minimum of $7mm$ of healthy perilesional tissue allowed to correctly detect all the TP lesions, as reported in Fig. 15. The standard deviation reached a peak at $3mm$ distance from lesion's border, meaning that some lesions already presented high scores while others had not yet been segmented. For Swin UNETR, the prediction score reached a plateau past $15mm$ from the lesion border, presenting a ripple around $6mm$. All the TP lesions were detected only when at least

13*mm* of healthy tissue were seen by the network.

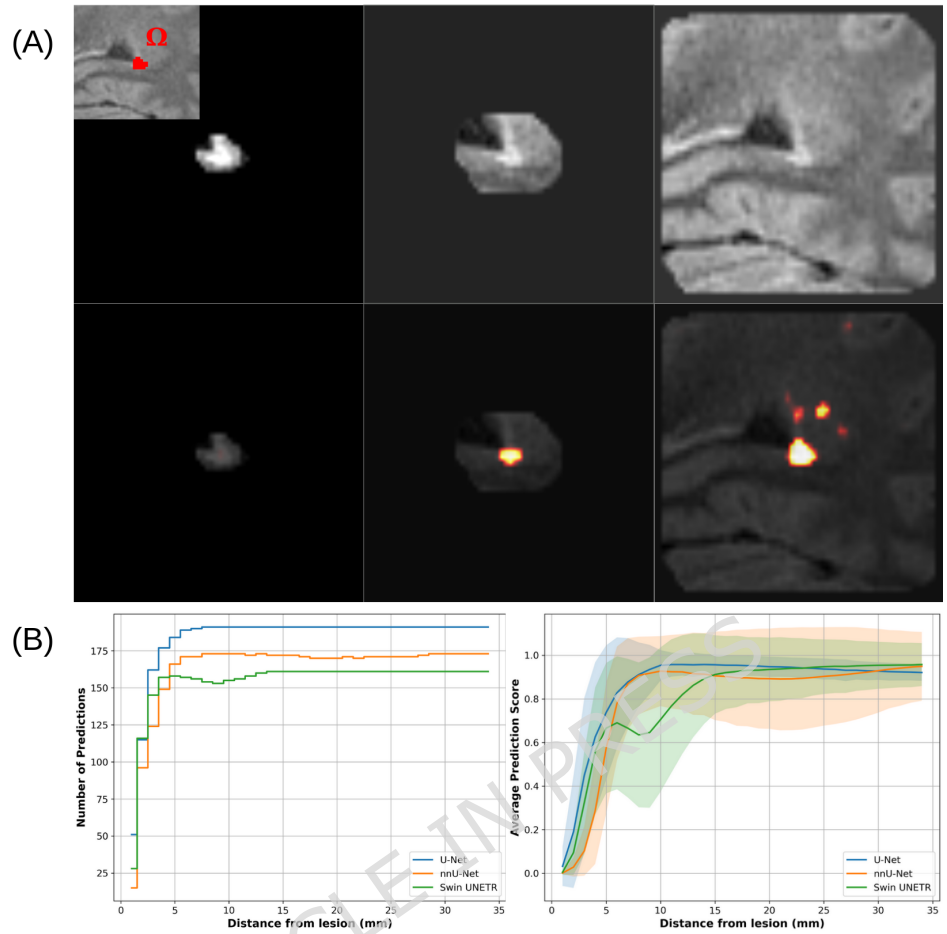

Figure 15. (A): FLAIR masked out with dilation steps 1, 5 and 24 (top), and the corresponding output probability maps computed for U-Net (bottom). (B): Plots representing the number of segmented lesions (left) and the average across patients of the mean prediction score (right) at each dilation step for three tested networks (the transparency for each line represents its standard deviation).

Discussion

We introduced novel methods to generate quantitative instance-level saliency maps, pushing the explanation capabilities of basic classification and segmentation saliency methods (SG and GradCAM++) to single instance segmentation. In medical image analysis, the latter is crucial to reveal the information used by a model to detect and segment a specific lesion of interest among other lesions of the same class, which is a very important task in daily radiology to assess disease development and response to treatment for all multi-lesional diseases (e.g., MS, metastatic cancer). In addition to providing explanations for a models' decision, the proposed quantitative maps can be used to optimize a network's architecture and to boost a semantic segmentation models' performance. This is shown in Spagnolo et al. [32], where features from saliency maps were used to increase detection F1 score and positive predictive value.

Although the test Dice scores of U-Net and nnU-Net are below that of Swin UNETR and of other state of the art methods [51, 52], there are some aspects of our approach one should consider: (i) the target metric for the loss was the normalized Dice score, which was 0.71 at test (ii) the imaging data do not contain synthetic lesions, and are not skull stripped (iii) possibly, the volume of false positive predictions is relatively small and (iv) the computational cost of saliency maps for Swin UNETR is higher. The false positive predictions of the model have been visually examined, and are mostly located in cerebral sulci and gyri. The use of brain extraction tools on our data would also have contributed to the removal of some false

positives, since a few of them were hyperintensities located in non-cerebral tissue.

The distribution of values in the saliency maps generated with SG showed that FLAIR imaging had a more significant contribution to the segmentation of lesions, compared to MPRAGE. This finding allowed us to check whether the model's behavior is coherent with clinical practice. Indeed, it resonates with the fact that FLAIR offers a better contrast for common MS lesions in the WM, such as periventricular and sub-cortical, since the signal from the cerebrospinal fluid is suppressed. Only a part of these lesions are clearly visible in T1-weighted images, such as MPRAGE, which are preferred to detect cortical plaques [53, 54].

For a lesion domain Ω, saliency maps (based on SG) with respect to FLAIR presented positive and negative gradients, respectively distributed inside Ω and in its neighborhood (Figs. 8 and 9). Indeed, positive gradients indicate that an increase in their intensities in the MRI would suggest the presence of a lesion in Ω. Conversely, voxels with negative saliency values indicate that the presence of a lesion in Ω would be suggested by a decrease in their intensities in the MRI. The antithetic distribution of saliency values in MPRAGE is due to the fact that WM lesions appear as hyperintense in FLAIR and as hypointense in MPRAGE, compared to the healthy WM.

Both of our XAI methods showed that voxels far from a lesion domain Ω did not appear to impact the prediction of voxels belonging to Ω (Fig. 10). Since we employed a CNN, each network unit did not depend on the entire input, but on a region called receptive field. Luo et al. [55] described that the effective area of a receptive field (effective receptive field, or ERF) starts as small, and then grows during training. Furthermore, the same study described skip-connections, part of a U-Net, as a cause of the reduction of receptive field's size. In our particular case, one of the possible interpretations could be that useful features to segment a lesion were close to the lesion itself and its neighborhood, so the learned receptive field is small. For instance, it is likely that re-training our model by including the images of the contextual experiment (e.g., until dilation step 25) and labeling them as background, would lead the model to have a larger ERF. Saliency maps, after such re-training, may show a higher attention to neighbouring tissue at a greater distance from a lesion border.

Peak values of saliency maps generated for the four groups of volumes (TP, FP, FN and TN) presented distributions that are significantly different from each other (Fig. 11). This suggests that, even if not segmented, FN volumes captured the model's attention notably more than TN volumes during inference. A similar conclusion can be drawn for TP and FP volumes. In both cases, our quantitative saliency maps could help increase the sensitivity and specificity of the model. However, in the case of FN volumes, some external input would be needed to select the lesion domain for computation. Such input could be that of a neurologist (for brain lesions), or could be derived from other maps, perhaps based on prediction's uncertainty.

Experiments on synthetic lesions suggested that the location of a lesion in the WM was not as important as the intensity of voxels within the lesion and its neighborhood (Fig. 14). The described behavior is expected, since we used a patch-based network. This, along with FLAIR's importance over MPRAGE, would support the hypothesis that the model's predictions rely predominantly on voxel intensities inside lesions in FLAIR.

However, the last experiment on contextual information revealed that high and stable prediction scores were related to the amount of contextual healthy brain tissue from the perilesional volume (Fig. 15). For U-Net and nnU-Net, enriching the context around the lesion in the input resulted in an increase of prediction scores up to a distance of $12-15mm$ from the lesion border. Past this distance, additional voxels no longer impact the prediction as seen by the plateau in Fig. 15(B). A possible conclusion is that a patch size of $96mm^3$ could have been unnecessarily large for most lesions. Furthermore, it would be interesting to test if the patch size during training would have a potential influence on the plateau's position of Fig. 15(B). We might expect that bigger patches would make the model require more contextual information to segment a lesion. The trajectory of prediction scores from Swin UNETR followed the same overall trend as the other two networks. However, certain architectural characteristics — such as the use of partitioned tiles ($8mm$ wide) applied to input patches — may make the model more sensitive to our experimental setup.

Regarding the potential of using the proposed XAI to improve model performance, the maximum and minimum values of XAI maps in Fig. 11 alone did not show enough discriminatory power between the different groups. However, a simple linear model relying on saliency radiomics allowed to reduce FPs, which is demonstrated in Spagnolo et al. (2025)[32] with a clear example of how our method can be relevant to the development and optimization of a segmentation model, and ultimately to the clinics.

Conclusion

We proposed novel XAI methods to provide quantitative instance-level explanations for segmentation, which we applied to the specific case of MS lesion segmentation. The analysis of the explanation maps and additional tests revealed fundamental insights into the decision mechanism of a deep neural network. The explanation maps were useful to understand the importance of the perilesional volume and improve the network's classification performances. The following experiment on the contextual information exploited by the network can guide architecture choices, such as patch size. The acquired new knowledge is crucial for AI engineers and clinical researchers, which constitutes an important step in facilitating AI integration into clinical practice. The proposed methods can potentially be applied to various segmentation architectures and tasks outside the medical imaging field, such as autonomous driving and robotics. Future research could leverage uncertainty maps to generate XAI maps of false negative examples, and further boost performance. Additionally, the pruning of false predictions could be studied separately for different areas of the brain.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Kołtuniuk, A., Pawlak, B., Krówczyńska, D. & Chojdak-Łukasiewicz, J. The quality of life in patients with multiple sclerosis – Association with depressive symptoms and physical disability: A prospective and observational study. *Front.* **13**, 1068421, DOI: 10.3389/fpsyg.2022.1068421 (2023).

2. Yang, J. *et al.* Current and Future Biomarkers in Multiple Sclerosis. *Int. journal molecular sciences* **23**, DOI: 10.3390/ijms23115877 (2022).

3. Thompson, A. *et al.* Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology* **17**, DOI: 10.1016/S1474-4422(17)30470-2 (2017).

4. Hemond C, C. & Bakshi, R. Magnetic Resonance Imaging in Multiple Sclerosis. *old Spring Harbor perspectives in medicine* **8**, 5, DOI: 10.1101/cshperspect.a028969 (2018).

5. Ayrignac, X. *et al.* Brain magnetic resonance imaging helps to differentiate atypical multiple sclerosis with cavitary lesions and vanishing white matter disease. *Eur. journal neurology* **23**, DOI: 10.1111/ene.12931 (2016).

6. Thaler, C. *et al.* T1- Thresholds in Black Holes Increase Clinical-Radiological Correlation in Multiple Sclerosis Patients. *PLOS ONE* **10**, e0144693, DOI: 10.1371/journal.pone.0144693 (2015).

7. Alrabai, A., Echtioui, A. & Hamida, A. Multiple Sclerosis Segmentation using Deep Learning Models : Comparative Study. In *2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 1–6, DOI: 10.1109/ATSIP55956.2022.9805983 (2022).

8. Zeng, C., Gu, L., Liu, Z. & Zhao, S. Review of Deep Learning Approaches for the Segmentation of Multiple Sclerosis Lesions on Brain MRI. *Front. Neuroinformatics* **14**, DOI: 10.3389/fninf.2020.610967 (2020).

9. Ma, Y. *et al.* Multiple Sclerosis Lesion Analysis in Brain Magnetic Resonance Images: Techniques and Clinical Applications. *IEEE J. Biomed. Heal. Informatics* **PP**, 1–1, DOI: 10.1109/JBHI.2022.3151741 (2022).

10. Diaz-Hurtado, M. *et al.* Recent advances in the longitudinal segmentation of multiple sclerosis lesions on magnetic resonance imaging: a review. *Neuroradiol.* **64**, DOI: 10.1007/s00234-022-03019-3 (2022).

11. Commowick, O., Combès, B., Cervenansky, F. & Dojat, M. Editorial: Automatic methods for multiple sclerosis new lesions detection and segmentation. *Front. Neurosci.* **17**, DOI: 10.3389/fnins.2023.1176625 (2023).

12. Baselli, G., Codari, M. & Sardanelli, F. Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way? *Eur. Radiol. Exp.* **4**, DOI: 10.1186/s41747-020-00159-0 (2020).

13. Spagnolo, F. *et al.* How far ms lesion detection and segmentation are integrated into the clinical workflow? a systematic review. *NeuroImage Clin.* **39**, 103491, DOI: 10.1016/j.nicl.2023.103491 (2023).

14. Kobayashi, K. & Alam, S. B. Explainable, interpretable, and trustworthy ai for an intelligent digital twin: A case study on remaining useful life. *Eng. Appl. Artif. Intell.* **129**, 107620, DOI: 10.1016/j.engappai.2023.107620 (2024).

15. Jonkman, L. *et al.* Can MS lesion stages be distinguished with MRI? A postmortem MRI and histopathology study. *J. neurology* **262**, DOI: 10.1007/s00415-015-7689-4 (2015).

16. Varatharasan, V., Shin, H.-S., Tsourdos, A. & Colosimo, N. Improving Learning Effectiveness For Object Detection and Classification in Cluttered Backgrounds. In *2019 Workshop on Research, Education and Development of Unmanned Aerial Systems (RED UAS)*, 78–85, DOI: 10.1109/REDUAS47371.2019.8999695 (2019).

17. Saranya, A. & Subhashini, R. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decis. Anal. J.* **7**, 100230, DOI: 10.1016/j.dajour.2023.100230 (2023).

18. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *2nd Int. Conf. on Learn. Represent. ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Work. Track Proc.* (2013).

19. de Vries, B. *et al.* Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. *Front. Medicine* **10**, DOI: 10.3389/fmed.2023.1180773 (2023).

20. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. SmoothGrad: removing noise by adding noise. *CoRR* (2017).

21. Goh, G. S. W., Lapuschkin, S., Weber, L., Samek, W. & Binder, A. Understanding Integrated Gradients with SmoothTaylor for Deep Neural Network Attribution. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 4949–4956, DOI: 10.1109/ICPR48806.2021.9413242 (2021).

22. Agarwal, S. *et al.* Towards the unification and robustness of perturbation and gradient based explanations. In *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 110–119, DOI: 10.48550/arXiv.2102.10618 (2021).

23. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626, DOI: 10.1109/ICCV.2017.74 (2017).

24. Chattopadyhay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 839–847, DOI: 10.1109/WACV.2018.00097 (2018).

25. Rajpurkar, P. *et al.* CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv* DOI: 10.48550/arXiv.1711.05225 (2017).

26. Mahapatra, D., Poellinger, A. & Reyes, M. Interpretability-Guided Inductive Bias For Deep Learning Based Medical Image. *Medical Image Analysis* **81**, 102551, DOI: 10.1016/j.media.2022.102551 (2022).

27. Arun, N. *et al.* Assessing the (Un)Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiol.* **3**, DOI: 10.1148/ryai.2021200267 (2021).

28. Singh, R. K., Gorantla, R., Allada, S. & Narra, P. SkiNet: A deep learning framework for skin lesion diagnosis with uncertainty estimation and explainability. *PloS one* **17**, e0276836, DOI: 10.1371/journal.pone.0276836 (2022).

29. Wickstrøm, K., Kampffmeyer, M. & Jenssen, R. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Medical Image Analysis* **60**, 101619, DOI: 10.1016/j.media.2019.101619 (2020).

30. Vinogradova, K., Dibrov, A. & Myers, G. Towards Interpretable Semantic Segmentation via Gradient-weighted Class Activation Mapping. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)*, vol. 34, 13943–13944, DOI: 10.1609/aaai.v34i10.7244 (2020).

31. Cordts, M. *et al.* The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3213–3223, DOI: 10.1109/CVPR.2016.350 (2016).

32. Spagnolo, F. *et al.* Exploiting XAI Maps to Improve MS Lesion Segmentation and Detection in MRI. In Celebi, M. E., Reyes, M., Chen, Z. & Li, X. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 Workshops*, 121–131, DOI: 10.1007/978-3-031-77610-6_12 (Springer Nature Switzerland, Cham, 2025).

33. Disanto, G. *et al.* The Swiss Multiple Sclerosis Cohort-Study (SMSC): A Prospective Swiss Wide Investigation of Key Phases in Disease Evolution and New Treatment Options. *PLoS One* **11(3)**, DOI: 10.1371/journal.pone.0152347 (2016).

34. La Rosa, F. *et al.* Multiple sclerosis cortical and WM lesion segmentation at 3T MRI: a deep learning method based on FLAIR and MP2RAGE. *NeuroImage: Clin.* **27**, 102335, DOI: 10.1016/j.nicl.2020.102335 (2020).

35. Çiçek, O., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *arXiv* DOI: 10.48550/ARXIV.1606.06650 (2016).

36. Isensee, F., Jaeger, P., Kohl, S., Petersen, J. & Maier-Hein, K. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 1–9, DOI: 10.1038/s41592-020-01008-z (2021).

37. Hatamizadeh, A. *et al.* Unetr: Transformers for 3d medical image segmentation. In *Winter Conference on Applications of Computer Vision (WACV)*, DOI: 10.1109/WACV51458.2022.00181 (2022).

38. Raina, V. *et al.* Tackling Bias in the Dice Similarity Coefficient: Introducing NDSC for White Matter Lesion Segmentation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 1–5, DOI: 10.1109/ISBI53787.2023.10230755 (2023).

39. Kofler, F. *et al.* Blob loss: instance imbalance aware loss functions for semantic segmentation. *arXiv* DOI: 10.48550/ARXIV.2205.08209 (2022).

40. Maier-Hein, L. *et al.* Metrics reloaded: Pitfalls and recommendations for image analysis validation. *arXiv.org* DOI: 10.48550/arXiv.2206.01653 (2022).

41. Klein, S., Staring, M., Murphy, K., Viergever, M. & Pluim, J. Elastix: A Toolbox for Intensity-Based Medical Image Registration. *IEEE transactions on medical imaging* **29**, 196–205, DOI: 10.1109/TMI.2009.2035616 (2009).

42. Shamonin, D. *et al.* Fast Parallel Image Registration on CPU and GPU for Diagnostic Classification of Alzheimer's Disease. *Front. neuroinformatics* **7**, 50, DOI: 10.3389/fninf.2013.00050 (2014).

43. Tustison, N. *et al.* N4ITK: improved N3 bias correction. *Med. Imaging, IEEE Transactions on* **29**, 1310 – 1320, DOI: 10.1109/TMI.2010.2046908 (2010).

44. Depeursinge, A. *et al.* Standardised convolutional filtering for radiomics. *arXiv* DOI: 10.48550/arXiv.2006.05470 (2020).

45. Polman, C. H. *et al.* Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria". *Annals of Neurology* **58**, 840–846, DOI: 10.1002/ana.20703 (2005).

46. Grahl, S. *et al.* Evidence for a white matter lesion size threshold to support the diagnosis of relapsing remitting multiple sclerosis. *Multiple Scler. Relat. Disord.* **29**, 124–129, DOI: 10.1016/j.msard.2019.01.042 (2019).

47. Fartaria, M. J. *et al.* Partial volume-aware assessment of multiple sclerosis lesions. *NeuroImage: Clin.* **18**, 245–253, DOI: https://doi.org/10.1016/j.nicl.2018.01.011 (2018).

48. De Rosa, A. *et al.* Consensus of algorithms for lesion segmentation in brain MRI studies of multiple sclerosis. *Sci. Reports* **14**, DOI: 10.1038/s41598-024-72649-9 (2024).

49. Jain, S. *et al.* Two time point ms lesion segmentation in brain mri: An expectation-maximization framework. *Front. Neurosci.* **10**, DOI: 10.3389/fnins.2016.00576 (2016).

50. McKnight, P. E. & Najab, J. *Mann-Whitney U Test* (John Wiley & Sons, Ltd, 2010).

51. Zhang, H. *et al.* Multiple sclerosis lesion segmentation with tiramisu and 2.5d stacked slices. In Shen, D. *et al.* (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, 338–346 (Springer International Publishing, 2019).

52. Commowick, O., Cervenansky, F., Cotton, F. & Dojat, M. MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure. In *MICCAI 2021 - 24th International Conference on Medical Image Computing and Computer Assisted Intervention*, 126 (2021).

53. Trip, S. A. & Miller, D. H. Imaging in multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry* **76**, 11–18, DOI: 10.1136/jnnp.2005.073213 (2005).

54. Nelson, F., Poonawalla, A., Hou, P., Wolinsky, J. & Narayana, P. 3D MPRAGE Improves Classification of Cortical Lesions in Multiple Sclerosis. *Multiple sclerosis (Houndmills, Basingstoke, England)* **14**, 1214–9, DOI: 10.1177/1352458508094644 (2008).

55. Luo, W., Li, Y., Urtasun, R. & Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*, 839–847, DOI: 10.1109/WACV.2018.00097 (2016).

Acknowledgements

This work was supported by the Hasler Foundation with the project MSxplain number 21042, the Swiss National Science Foundation (SNSF) with the project 205320_219430, and the Swiss Cancer Research foundation with the project TARGET (KFS-5549-02-2022-R). We acknowledge access to the expertise of the CIBM Center for Biomedical Imaging, a Swiss research center of excellence founded and supported by CHUV, UNIL, EPFL, UNIGE and HUG.

Author contributions statement

F.S., V.A. and A.D. conceived the experiments, F.S. conducted the experiments, L.M-G, M.O.P. and F.S. processed and organized the data, F.S. and N.M. created the software used in the work, V.A., A.D. and C.G. supervised the work. F.S. led the writing of the manuscript together with V.A. and A.D. All authors reviewed the manuscript.

Additional information

Competing interests

The University Hospital Basel (USB) and the Research Center for Clinical neuroimmunology and Neuroscience (RC2NB), as the employers of Cristina Granziera, have received the following fees which were used exclusively for (research support from Siemens, GeNeuro, Genzyme-Sanofi, Biogen, Roche. They also have received advisory board and consultancy fees from Actelion, Genzyme-Sanofi, Novartis, GeNeuro, Merck, Biogen and Roche; as well as speaker fees from Genzyme-Sanofi, Novartis, GeNeuro, Merck, Biogen and Roche.