NeuroImage: Clinical 31 (2021) 102707

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

icobrain ms 5.1: Combining unsupervised and supervised approaches for improving the detection of multiple sclerosis lesions

Mladen Rakić [a,c,*], Sophie Vercruyssen [a], Simon Van Eyndhoven [a], Ezequiel de la Rosa [a,d], Saurabh Jain [a], Sabine Van Huffel [b], Frederik Maes [c], Dirk Smeets [a], Diana M. Sima [a]

[a] icometrix, Leuven, Belgium
[b] KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, 3001 Leuven, Belgium
[c] KU Leuven, Department of Electrical Engineering (ESAT), Processing Speech and Images (PSI) and Medical Imaging Research Center, 3001 Leuven, Belgium
[d] Technical University of Munich, Department of Computer Science, Munich, Germany

ARTICLE INFO

Keywords:
Multiple sclerosis
Lesion segmentation
Attention-gate U-net
Combined classification
Unsupervised machine learning
Deep-learning

ABSTRACT

Multiple sclerosis (MS) is a chronic autoimmune, inflammatory neurological disease of the central nervous system. Its diagnosis nowadays commonly includes performing an MRI scan, as it is the most sensitive imaging test for MS. MS plaques are commonly identified from fluid-attenuated inversion recovery (FLAIR) images as hyperintense regions that are highly varying in terms of their shapes, sizes and locations, and are routinely classified in accordance to the McDonald criteria. Recent years have seen an increase in works that aimed at development of various semi-automatic and automatic methods for detection, segmentation and classification of MS plaques. In this paper, we present an automatic combined method, based on two pipelines: a traditional unsupervised machine learning technique and a deep-learning attention-gate 3D U-net network. The deep-learning network is specifically trained to address the weaker points of the traditional approach, namely difficulties in segmenting infratentorial and juxtacortical plaques in real-world clinical MRIs. It was trained and validated on a multi-center multi-scanner dataset that contains 159 cases, each with T1 weighted (T1w) and FLAIR images, as well as manual delineations of the MS plaques, segmented and validated by a panel of raters. The detection rate was quantified using lesion-wise Dice score. A simple label fusion is implemented to combine the output segmentations of the two pipelines. This combined method improves the detection of infratentorial and juxtacortical lesions by 14% and 31% respectively, in comparison to the unsupervised machine learning pipeline that was used as a performance assessment baseline.

1. Introduction

Multiple sclerosis (MS) is a chronic autoimmune, inflammatory neurological disease of the central nervous system (CNS). It affects the myelinated axons in the CNS, destroying the myelin and the axons to a varying degree. The evolution of the disease is unpredictable, often including reversible neurological deficits in the early stages, followed by progressive neurological deterioration over time. The cause of the disease is unknown, but it involves a combination of genetic susceptibility and a non-genetic trigger (e.g. virus, metabolism or environmental factors) (Calabresi, 2004; Hauser and Goodin, 2005; Weinshenker, 1996; Goldenberg, 2012). The diagnosis of MS is rather complex, due to the varying nature of the disease, but today's clinical practice includes a brain MRI scan, as it is the most sensitive imaging test for MS (Brust, 2018). On a brain MRI, MS is characterized by white matter lesions (or plaques) that appear hyperintense on FLAIR images and hypointense on T1w images. These plaques are varying in terms of their size, shape or location, and are classified according to the McDonald criteria (McDonald et al., 2001), based primarily on their location, into 3 classes: (*i*) infratentorial lesions, found below the cerebellar tentorium line, which separates cerebrum from cerebellum and brainstem; (*ii*) periventricular lesions, which are in direct contact with lateral ventricles; and (*iii*) juxtacortical lesions, which are found in the close proximity to the cortex. All the remaining white matter FLAIR hyperintensities that do not fall in any of the three categories, are further referred to as deep white matter lesions. Due to the highly varying appearances of these plaques, the task of their segmentation and classification is very demanding, both for the radiologists and software, and involves subtle

* Corresponding author.
E-mail address: mladen.rakic@icometrix.com (M. Rakić).

https://doi.org/10.1016/j.nicl.2021.102707
Received 19 January 2021; Received in revised form 20 May 2021; Accepted 21 May 2021
Available online 4 June 2021

considerations when detecting smaller lesions in particular. The complexity of the task increases even more when class imbalance is taken into account: periventricular plaques tend to be dominant, followed by deep white matter, while infratentorial and juxtacortical plaques are commonly underrepresented. Furthermore, MRI scans are varying in protocols and image characteristics in general, in contrast to very standardized images obtained, for instance, with computed tomography. This paper presents a combined method for the MS lesion segmentation and classification on real-world clinical data, which combines a traditional machine learning technique implemented in ico**brain ms** 5.0, with a deep-learning network that aims to improve weaker points of the former method. It has been shown that combined methods add a margin of improvement in similar tasks aimed at segmentation and/or classification of various brain pathologies (Kamnitsas et al., 2017; Valverde et al., 2017; Rakić et al., 2020). Special attention is given to making the method robust and able to generalize well, which reflects mainly in the way multi-center real-world data stratification is performed. ico**brain ms** 5.0 (Jain et al., 2015) was used as a baseline to assess the performance on the segmentation and classification task. The method was then developed in order to improve the current version, and was consequently integrated into the new version of the ico**brain ms** software, 5.1.

2. Materials and methods

2.1. ico*brain ms*

ico**brain** 5.0 is a medical device software that measures relevant volumes of brain structures to assist radiologic assessment of patients with neurological disorders (Jain et al., 2015). It is a fully automated tissue and lesion segmentation and quantification software that uses 3D T1-weighted and FLAIR MRIs. The entire method is described elsewhere (Jain et al., 2015). Briefly, the FLAIR image is rigidly co-registered to the T1-weighted image, and probabilistic anatomical priors for grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF), defined in Montreal Neurological Institute (MNI) space (Evans et al., 1992), are transferred to the T1-weighted image space using an affine registration followed by a non-rigid registration. An iterative algorithm based on probabilistic tissue priors is then used to segment the T1-weighted image into GM, WM and CSF, while maintaining spatial consistency, until convergence. Further, an iterative process that generates a map based on deviation of the intensity of each voxel in the FLAIR image from the combined tissue classes (outlier belief map) is used to segment WM lesions, which are then filled in on the bias-corrected T1-weighted image with their neighborhood WM intensities. Binary lesion segmentation is split into its individual connected components prior to labeling them according to location. Anatomic regions are used in order to establish proximity and to guide the classification of lesions based on their location. The following regions are first obtained from ico**brain** segmentations of WM, GM and CSF, with the help of a parcellation atlas defined in MNI space and brought to patient space through the non-rigid atlas to image registration: (*i*) cortical gray matter, (*ii*) lateral ventricles, (*iii*) brainstem, and (*iv*) cerebellum. The following rules for labelling each component are applied in cascade: (*i*) if the dilated component's overlap with cortical gray matter accounts for more than 10% of the component's volume, it is juxtacortical; (*ii*) else if the component is within the infratentorial region (brainstem or cerebellum, as defined by their corresponding segmentation masks), it is infratentorial; (*iii*) else if the component intersects the dilation of the lateral ventricles segmentation, it is periventricular; (*iv*) else the component is deep white matter.

2.2. Dataset

The dataset consists of real-world pre-contrast T1 and FLAIR MS patient brain scans, together with the manual delineation of four different spatial classes of white matter plaques/FLAIR hyperintensities: infratentorial, periventricular, juxtacortical and deep white matter. Manual delineation, also referred to as ground truth, was based on 2D semi-automatic delineation on axial slices, and each subject was delineated by one rater trained in manual MS lesion segmentation, after which the validation was performed by an independent reviewer. The data is a multi-center and multi-scanner sample of real-world clinical data, and was stratified into training, validation and testing sets by following a carefully designed protocol that preserves robustness and minimizes any bias towards a certain aspect of the data, such as screening site, scanner model, magnetic field strength, scan quality, slice thickness, etc. From the available pool of subjects that agreed to the use of their data for research purposes, we sampled 93 subjects for training, 15 subjects for validation, and 51 subjects for the testing set. The sampling was done in a semi-random way in order to meet the stratification criteria as closely as possible. Take for example scanner manufacturer as a criterion: the idea is to have scans coming from all manufacturers present in the full data, which were Philips, GE, Siemens and Hitachi, in all 3 sets, in order to have diverse sets; but, in the original pool of patients, scans coming from Philips and Siemens scanners were predominantly represented, and will stay dominant by randomly sampling cases for the training set, but it is ensured that scans coming from all 4 scanner manufacturers are still present in the stratified set. Afterwards, we try to mimic the same notion in the validation and testing sets – to have all 4 represented, but by following the same distribution that was determined from the original pool of subjects. This sampling procedure was employed by simultaneously accounting for the following characteristics: patient age, patient sex, lesion load (in ml), T1 slice thickness (in mm), FLAIR slice thickness (in mm), magnetic field strength (in T), scanner manufacturer and scanner model. To further minimize bias during training with respect to screening site and scanner manufacturer in this multi-center dataset, the following strategy was used: if a certain screening site/scanner manufacturer combination was sampled and present in the testing set, the same combination cannot be associated with any of the subjects selected for training and validation sets, the idea being to keep training and testing phases as independent from each other as possible. Manual delineation of all cases was performed by a panel of raters. Each case was segmented by one and reviewed by

Table 1
Summary of the dataset details after the data stratification step. Numbers specified in brackets indicate range of values. Lesion load was computed from the MNI space. (GT – ground truth, IT – infratentorial, PV – periventricular, JC – juxtacortical, DWM – deep white matter)

Criterion	Training	Validation	Testing
Patient age	[16, 81]	[17, 73]	[20, 70]
Patient sex	M: 40, F: 53	M: 4, F: 11	M: 20, F: 31
T1 slice thickness [mm]	[0.45, 1.20]	[0.50, 1.50]	[0.50, 1.20]
FLAIR slice thickness [mm]	[0.50, 3.00]	[0.50, 3.00]	[0.50, 3.00]
Lesion load from GT [ml]	[0.22, 76.54]	[2.45, 27.56]	[0.17, 69.43]
Lesion load per class [ml]	IT: [0, 4.07] 0.24	IT: [0, 0.55] 0.14	IT: [0, 1.09] 0.14
[MIN–MAX] MEAN	PV: [0.05, 74.14] 11.44	PV: [0.84, 22.30] 10.82	PV: [0.02, 66.11] 8.33
	JC: [0, 7.18] 0.77	JC: [0, 2.79] 0.86	JC: [0, 4.25] 0.66
	DWM: [0, 5.55] 1.29	DWM: [0, 4.01] 1.58	DWM: [0, 7.91] 1.03
Magnetic field strength	1.5T: 39, 3T: 54	1.5T: 4, 3T: 11	1.5T: 19, 3T: 32
Scanner manufacturer	Philips: 35	Philips: 5	Philips: 18
	GE: 10	GE: 2	GE: 17
	Siemens: 46	Siemens: 7	Siemens: 15
	Hitachi: 2	Hitachi: 1	Hitachi: 1
Number of cases	93	15	51

another rater. However, raters assigned to segmenting training cases did not get the testing cases assigned to them, and vice versa. Table 1 summarizes the details of the three sets, listing the associated characteristics used during the stratification step.

2.3. Attention gate U-net

In order to tackle the challenge of locating, segmenting and classifying infratentorial and juxtacortical plaques in a more robust way, we propose in this paper a deep-learning approach. We trained a 3D attention-gate U-net (Fig. 1), inspired by the works of Çiçek et al. (2016) and Oktay et al. (2018). The underlying principle of an attention-gate block is that it guides the training process so that the network focuses on learning more salient features of the input patches. All images were subjected to a common preprocessing pipeline, which includes the following steps: (*i*) each T1w and FLAIR image was affinely registered to the MNI template (Evans et al., 1992) and linearly resampled to $1 \times 1 \times 1$ mm, (*ii*) skull-stripping using the brain mask obtained with ico**brain** was performed, and (*iii*) a simple z-score intensity normalization (i.e. subtracting the mean intensity from each brain and dividing by its standard deviation) followed. The first half of the network (downward path) compresses the relevant information from the input image, which helps the second half of the network (upward path) in synthesising the segmentation information at four different resolutions. In the deeper layers of the network, the receptive field is much larger than in the shallow layers, providing coarse-grained context for the segmentation in the upward path. It is important to note that all operations specified here are performed on both T1w and FLAIR input patches (i.e. T1w and FLAIR images form a 2-channel 3D image input). In the compression/downward path, every layer has two convolutional blocks with kernel size of $3 \times 3 \times 3$ and each convolution is followed by a rectified linear unit (ReLU) activation function (Krizhevsky et al., 2017). Then, a $2 \times 2 \times 2$ max pooling operation decreases the input resolution by half. During the synthesis/upward path, a $2 \times 2 \times 2$ upsampling operation is performed, followed by concatenation with an attention-gated feature map from the downward path at the same level, as well as with the cropped feature map from the downward path at the same level, and two convolutions with ReLU activation. In the attention-gate blocks, instance normalization (Ulyanov et al., 2016) is employed after each convolution layer with a leaky ReLU activation (Maas et al., 2013). Skip connections are added from the same resolution of the compression path to provide high-resolution features before performing convolution operation. In the final layer, a $1 \times 1 \times 1$ convolution layer reduces the number of output channels to 5 classes (background, infratentorial plaques, periventricular plaques, juxtacortical plaques and deep white matter plaques) followed by a softmax function to enforce sparse segmentation. The network has around 21 million parameters and is trained with input voxel patches of size $64 \times 64 \times 64$ with 5 output channels and a batch size of 2. The output patch size is $64 \times 64 \times 64$ with a voxel size of $1 \times 1 \times 1$ mm^3 in the output segmentation. The output of the network is resampled to the input resolution by performing nearest neighbour interpolation. The network is trained by minimizing the sum of the mean (over classes) of the Dice loss and the Weighted Categorical Cross-Entropy via means of Stochastic Gradient Descent with momentum (learning rate $= 10^{-3}$, decay factor $= 0.1$, momentum $= 0.9$).

The Weighted Categorical Cross-Entropy loss (wCXE) for the batch of

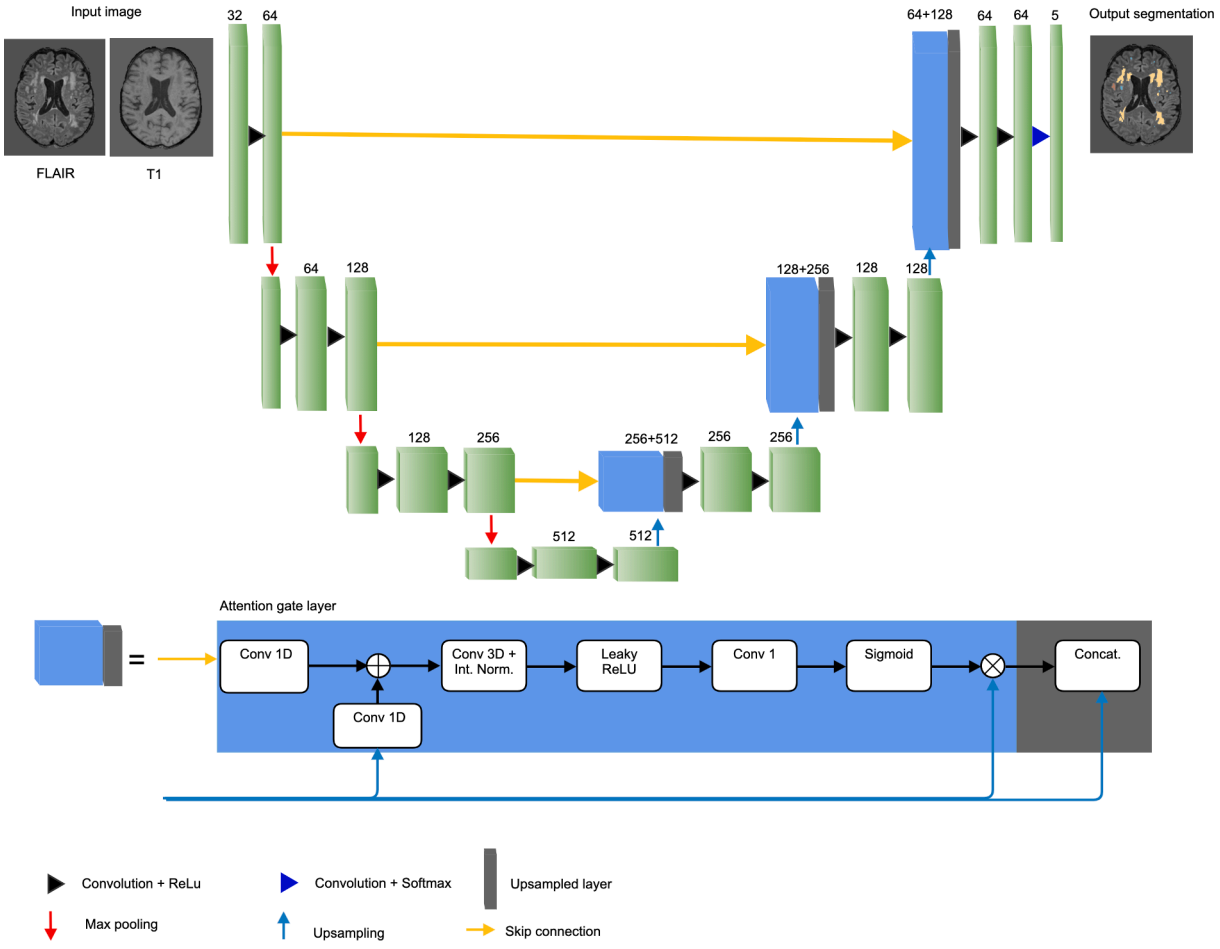


Fig. 1. Graphical representation of the attention-gate U-net used for segmentation and classification of the plaques. (Conv. – convolution, Int. Norm. – intensity normalization, Concat. – Concatenation, + – element-wise sum, × – matrix multiplication).

M images, in which the sample weight at each voxel is purely defined by its ground truth class c, is defined as:

$$J_{wCXE}(\boldsymbol{\alpha}, \boldsymbol{Y}, \boldsymbol{P}) = -\frac{1}{MI}\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{c=1}^{C}\alpha_c y_{mi}^{(c)}\log p_{mi}^{(c)} \tag{1}$$

$$= -\frac{1}{MI}\sum_{m=1}^{M}\sum_{c=1}^{C}\alpha_c\langle \mathbf{y}_m^{(c)}, \log\mathbf{p}_m^{(c)}\rangle \tag{2}$$

It is bounded between 0 (when the predictions for all ground truth positive voxels approach 1) and ∞ (when the predictions for all ground truth positive voxels approach 0). In the formula above, M is the batch size, I is the number of voxels in each patch, while C is the number of classes. Further, $\mathbf{y}_m^{(c)} \in \{0,1\}^I$ is the ground truth for class c in image m, $\mathbf{p}_m^{(c)} \in \mathbb{R}^I$ is the softmax prediction for class c in image m and $\boldsymbol{\alpha} \in \mathbb{R}^C$ are the class importance weights. These weights are 0.35 for infratentorial, 0.1 for periventricular, 0.35 for juxtacortical and 0.25 for deep white matter class, as well as 0.05 for the background.

The Mean Dice loss is the average of the class-wise Dice losses.

$$J_{MDL}(\boldsymbol{Y}, \boldsymbol{P}) = 1 - \frac{1}{C}\sum_{c=1}^{C}\sum_{m=1}^{M}\frac{2\sum_{i=1}^{I} y_{mi}^{(c)} p_{mi}^{(c)}}{\sum_{i=1}^{I}(y_{mi}^{(c)} + p_{mi}^{(c)})} \tag{3}$$

The Mean Dice loss is bounded between 0 (when the predictions for all ground truth positive voxels approach 1, and the predictions for all ground truth negative voxels approach 0) and 1 (when the predictions for all ground truth positive voxels approach 0).

In order to increase diversity of the dataset and minimize overfitting, simple data augmentation, which includes left/right flipping of the patches and addition of white Gaussian noise, was applied. To compensate for the class imbalance, namely the underrepresentation of infratentorial and juxtacortical plaques, we consider class-weighted cost function, as previously stated. Generally, each voxel was assigned the label corresponding to the class with the highest softmax probability in the network's output, but only if this probability exceeded a set threshold (0.5 for infratentorial and juxtacortical lesions, 0.75 for periventricular and deep white matter lesions) – else it was considered background. These thresholds were tuned and chosen with a goal of approximately maximizing the lesion detection performance on the validation dataset. Subsequently, several refinement steps were conducted as part of the postprocessing pipeline. These steps relate to labelling restrictions that can be formulated as follows: (*i*) all lesions segmented in the infratentorial region of the brain (which was defined by overlaying the cerebellum and brainstem mask) were strictly assigned infratentorial label, and vice versa, lesions segmented in the supratentorial region as infratentorial were assigned the label based on the second highest probability; (*ii*) only lesions touching the lateral ventricles of the brain could be assigned periventricular label; and (*iii*) in case a segmented group of connected voxels contained more than one assigned label, the most frequent label present in the segmentation was kept for the entire connected component.

2.4. Combined method

We use a label fusion approach to combine outputs of ico**brain ms** 5.0 and the attention-gate U-net presented in Section 2.3. Each case is thus first segmented with both pipelines. Since the attention-gate U-net was designed to focus on the underrepresented juxtacortical and infratentorial classes, we hypothesize that the U-net segmentation outperforms the ico**brain ms** 5.0 pipeline for these classes only. Therefore, we kept periventricular and deep white matter lesions detected with ico**brain ms** 5.0, and infratentorial and juxtacortical plaques coming from the outputs of attention-gate U-net in the final segmentation. A rule-based approach is conducted as a post-processing step, in order to resolve potential labelling conflicts that can occur when merging the resulting segmentation outputs from the two separate pipelines. A detailed graphical explanation of this conflict resolving step is provided in Fig. 2. It is important to note the fact that detected lesions smaller than 0.005 ml were removed from the final segmentation. This threshold was chosen by evaluating different values ranging up to 0.01 ml on the validation dataset and by visual inspection. Even though it might be clinically relevant to catch such small lesions, we found that keeping those would result in too many false positives. Furthermore, we believe that such small lesions are potentially heavily impacted by partial volume effects, so we do not expect meaningful training and inference to be possible at that scale. The combined approach, together with the baseline ico**brain ms** 5.0, was also evaluated on the publicly available training set containing 15 cases from 2016 MICCAI MS Lesion Segmentation Challenge (Commowick et al., 2018).

2.5. Performance metrics

We use lesion-wise Dice score (LWDS) computed over the entire test set as a metric that encompasses both detection of the plaques, as well as their classification, for it allows simple comparison when evaluating results coming from different methodological approaches. It is a metric that quantifies the detection rate and correct classification rate of different types of plaques, by comparing labels of connected components in ground truth with associated network output.

$$LWDS = \frac{2TP + \epsilon}{2TP + FN + FP + \epsilon} \tag{4}$$

In the formula above, TP, FN and FP signify counts of detected lesions (i.e. connected components) that are true positives, false negatives and false positives, respectively. ϵ is a small smoothing factor, commonly used to avoid division by zero.

We define a lesion as detected if at least one labelled voxel in a connected component inside the segmented volume matches the corresponding voxel's label in the ground truth. All the connected components in the output segmentation have a single label assigned to each, so the issue of having a single voxel randomly assigned a different label inside a connected component, and thus by chance get matched with a

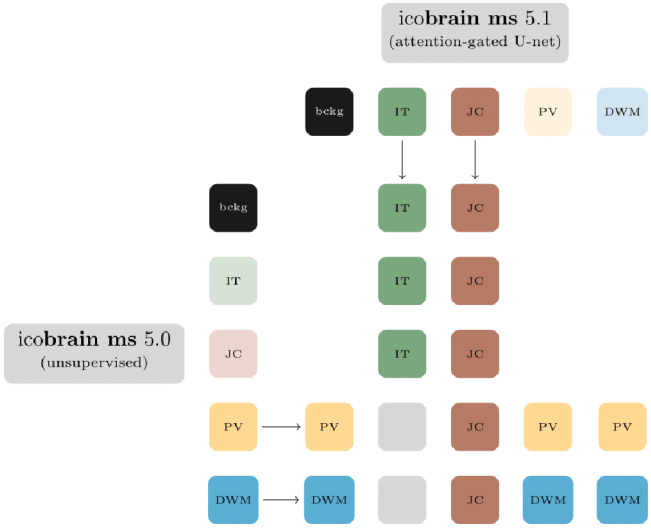


Fig. 2. Potential labelling conflicts are resolved as indicated in the matrix above. A vivid color indicates that the label provided by one of the pipelines should be kept (i.e., periventricular (PV) and deep white matter (DWM) from ico**brain ms** 5.0, and infratentorial (IT) and juxtacortical (JC) from the combined method, ico**brain ms** 5.1), while a light color indicates that the label class should be discarded. In case of conflicts, the matrix shows which label prevails. Gray squares signify that conflicts are not possible for IT lesions, since the anatomical constraints for assigning the IT label enforce mutual exclusion with other labels.

corresponding component in ground truth, is avoided. Note that in order to define and label connected components, we consider the 26-connected voxels neighborhood in 3-dimensional space.

The only problematic situation occurs when one connected component in ground truth overlaps with more connected components in the automated segmentation, or vice versa. To tackle this, we propose the following strategy of counting overlapping lesions: first, we loop over all the connected components in the ground truth segmentation and we look for matching components in the method's segmentation. If a certain lesion gets matched, and is thus considered as true positive, then the loop over all components is interrupted. Otherwise, if at the end of the loop there is no match for a certain connected component, 1 is added to the false negative count for the corresponding class. Secondly, we loop over all connected components in the method's segmentation and perform a reverse comparison with components in ground truth segmentation. In case there is a component that has no match (i.e. no overlapping voxels with any component in ground truth), 1 is added to the false positive count for the associated class. This strategy also allows for creation of confusion matrices associated with each of the described methods, which contain the number of correctly classified plaques, as well as counts of missclassified ones, illustrating different class mixing combinations that occurred. Each row corresponds to a certain class and is linked to the experts' delineation (i.e. ground truth), whereas each column corresponds to the different classes, but associated with the method output. Hence, a confusion matrix with high counts on the main diagonal (true positives) and low counts elsewhere (class mixing, false positives, false negatives) is indicative of high lesion detection performance.

Additionally, we use the volumetric computations to determine the differences between the estimated volumes and the ones calculated from the ground truth. These volumetric comparisons offer further insights into the performance of a method and were conducted on two levels: on a patient level, considering the lesions as a single class, and on a lesion level, where we distinguish among the 4 defined classes of lesions. Besides the estimated volumes, we also present voxel-wise Dice scores obtained on a patient level, and how these scores relate to the volumetric lesion load from the ground truth.

3. Results

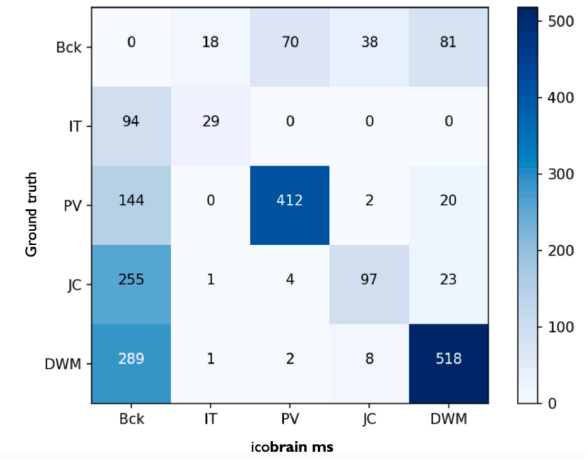
Table 2 illustrates the lesion-wise Dice scores obtained for each of the three methods (i.e. two separate pipelines and one combined approach) evaluated on the test set which contains 51 cases; the scores are split into 4 values associated with each of the 4 classes of interest.

Figs. 3a–c show the confusion matrices of ico**brain ms** 5.0, the attention-gate U-net and the combined method, respectively.

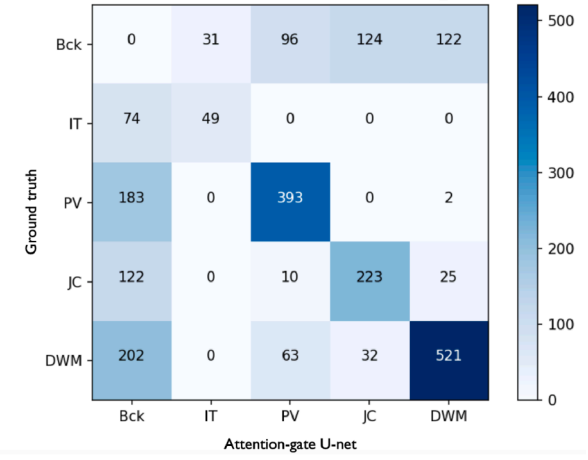
The combined approach has fewer false negatives (first column of the matrices) than ico**brain ms** 5.0. Similarly, there is a drop in false positive counts (first row of the matrices) in the combined approach, compared to attention-gate U-net method. More importantly, there is an increase in true positive counts in the combined method, especially for infratentorial and juxtacortical plaques, compared to the original ico**brain ms** 5.0 results.

Table 2
Mean lesion-wise Dice scores for different types of multiple sclerosis plaques evaluated on the test set (51 cases) using 3 different methods as segmentation techniques. (AG – attention-gate, IT – infratentorial, PV – periventricular, JC – juxtacortical, DWM – deep white matter)

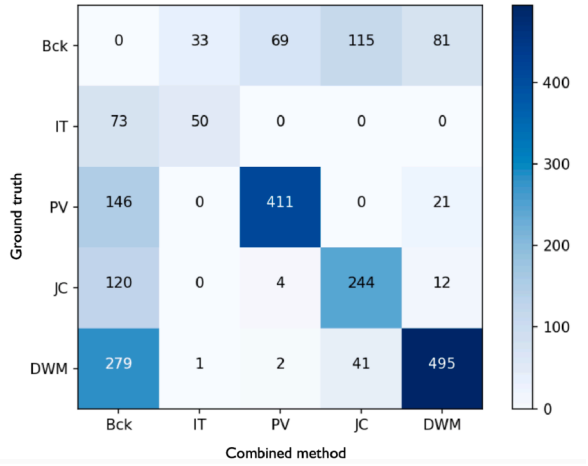
	ico**brain ms** 5.0	AG U-net	Combined
IT	0.34	0.48	0.48
PV	0.77	0.69	0.77
JC	0.31	0.59	0.62
DWM	0.71	0.70	0.69
Mean	0.57	0.62	0.64



(a) ico**brain ms** 5.0 has a very good performance for PV and DWM, but falls short in detecting JC and IT lesions.



(b) the attention-gate U-net improves the detection of IT and JC lesions, compared to ico**brain** 5.0.



(c) the combined method takes the best of both individual methods in order to improve the performance overall.

Fig. 3. Confusion matrices obtained using three different methods evaluated on the test set. (Bck – background, IT – infratentorial, PV – periventricular, JC – juxtacortical, DWM – deep white matter).

We conducted Wilcoxon signed-rank tests per lesion type to assess the difference in lesion-wise Dice scores attained by the different methods accross patients. Obtained p-values lead us to the following conclusions: (*i*) periventricular lesion detection and segmentation was significantly better in the combined method in comparison to the attention-gate U-net model ($p < 0.05$), (*ii*) juxtacortical lesion performance was significantly better in the combined method in comparison to the ico**brain ms** 5.0 ($p < 0.05$), (*iii*) there was no significant difference observed among the methods when it comes to deep white matter lesions, (*iv*) there was no significant improvement observed in infratentorial lesion detection and segmentation in the combined method compared to ico**brain ms** 5.0, even though Table 2 shows an average improvement of 14%.

The example images in Fig. 7 illustrate the new method's capabilities of better detection and segmentation of juxtacortical and infratentorial plaques, in comparison to the original ico**brain ms** 5.0 version.

As an additional insight into the obtained results from the combined method in particular, Fig. 4 summarizes the volumetric estimations obtained on a patient level, considering all the lesion classes as one, whereas Fig. 5 breaks down these estimates according to the 4 defined classes of MS lesions. Furthermore, Fig. 6 shows the values of voxel-wise Dice scores calculated on a patient level. The results reported on Figs. 4–6 are all evaluated on the test set, and all the displayed volumes are computed from the MNI space.

When it comes to evaluating the methods on the 2016 MICCAI MS Lesion Segmentation Challenge training dataset, the mean obtained lesion-wise Dice scores were 0.61 for ico**brain ms** 5.0, 0.68 for attention-gate U-net and 0.66 for the combined method, which is relatively in line with the values reported in Table 2. Even though the highest value obtained with this analysis corresponds to the attention-gate U-net, it also resulted in the highest number of false positive lesions. On the other hand, ico**brain ms** 5.0 yielded the highest number of false negative lesions, while the combined method balanced those numbers out. This showcases the generalisation capabilities and robustness of the proposed method when evaluated on an independent dataset.

Additionally, we provide the voxel-wise Dice scores obtained on the 2016 MICCAI MS Lesion Segmentation Challenge training dataset when using our three methods. The obtained average values per patient were 0.58 for both ico**brain ms** 5.0 and 5.1 and 0.62 for the attention-gate U-net. These values are further looked into in the next section.

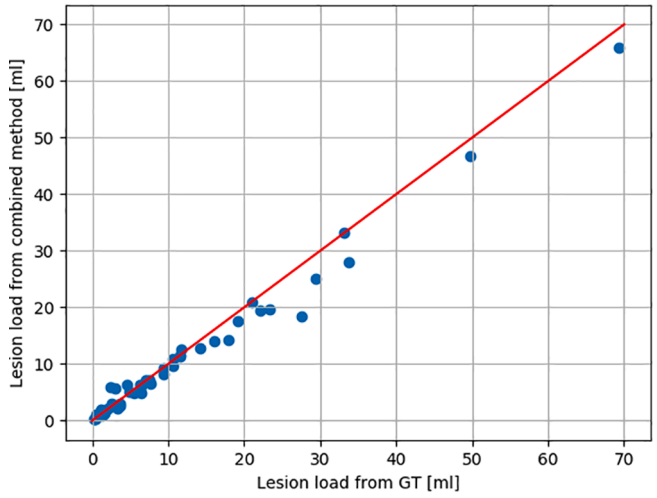


Fig. 4. The combined method shows a good trend when it comes to volumetric estimation of lesions in general (Pearson's correlation coefficient $r = 0.99$), displaying the volumetric intersection points in relatively close proximity to the identity line shown in red. Each dot represents one subject from the test set.

4. Discussion

The principal motivation of this research was to come up with a robust way of segmenting and classifying MS lesions coming from real-world clinical data, regardless of the scanner model and imaging protocols used or volumetric lesion load observable from the scan. The performance of ico**brain ms** 5.0 is very good when it comes to periventricular and deep white matter lesions, but the detection of infratentorial and juxtacortical lesions is its main weak point. To address this shortcoming, we trained the attention-gate U-net. It is important to note that we also trained a standalone U-net without the attention gates, but obtained unsatisfactory results. The combination of both methods subsequently improved the performance of either method in isolation. Even though the Wilcoxon signed-rank test showed no significant improvement when it comes to infratentorial lesions, a plausible explanation is that because the lesion-wise Dice scores are computed on a per patient basis, and with the sparse occurrence of lesions often end up being 0 or 1, the statistical test was expected to have low power for this class. While the results presented in Figs. 3 and 4 showcase promising results in terms of the detection rates and the estimation of lesion load per patient, it is important to address some subtleties and draw conclusions from the results shown in Figs. 5 and 6. Even though the combined method demonstrates favorable trends in terms of volumetric estimation for certain lesion types, mainly periventricular and deep white matter, there are still cases where the estimation of underrepresented lesion types (infratentorial and juxtacortical) is subpar. Further, in case of a patient with a very low total lesion load, there is a chance that the segmentation fails in terms of Dice coefficient, which is especially observable in Fig. 6. Over the course of development of this method, special consideration was given to architectural choices for the design of the attention-gate U-net. Several loss functions, including Dice loss, Generalised Dice loss, Mean Dice loss, Weighted Categorical Cross-Entropy, and various combinations of the losses commonly present in state-of-the-art, have been tested in order to select the one that yields improved network training (Crum et al., 2006; Sudre et al., 2017; Isensee et al., 2018). The chosen loss function is based on the summation of a Weighted Categorical Cross-Entropy and a Mean Dice loss, and thus combines a per-pixel loss with a loss which considers spatial context or overlap. We found that the differential weighting of the classes in the cost function was sufficient to compensate for class imbalance, and that little or no improvement could be achieved beyond this point, via additional methods. We also considered different patch sizes, with underlying assumption being that larger patches allow the network to learn more contextual information. However, patches of $64 \times 64 \times 64$ (in batch size of 2) were the largest that were possible under the memory limitations. The research was conducted on 8 vCPUs with 16 GB memory. Regarding the computational time, the U-net processing lasts on average less than a minute per case, whereas the total computational time of the combined method is somewhere in the range of 15 to 20 min. Comparing our method with the current state-of-the-art is a highly complex task for numerous reasons. First of all, there is a high inter- and intra-expert variability when it comes to manual delineation of the plaques. Several studies have investigated this variability and reported values that go as high as 68% of volume differences among expert raters (Grimaud et al., 1996; Styner et al., 2008; Zijdenbos et al., 2002). To tackle this challenge, many studies focused on developing semiautomatic segmentation methods (Filippi et al., 1995; Grimaud et al., 1996; Udupa et al., 1997; Ashton et al., 2003; Parodi et al., 2002; Johnston et al., 1996), which often suffer from being too time consuming in everyday clinical practice. When it comes to automatic methods, which on the other side aim to exclude human interaction (and consequently reduce the variability), the biggest obstacle is the validation of a certain method prior to its integration in clinical practice (García-Lorenzo et al., 2013). Numerous studies used either synthetic data, which does not provide sufficient proof of robustness in real-world scenarios (Bricq et al., 2008; Forbes et al., 2010; Shiee et al., 2010; Younis et al., 2007). Others considered

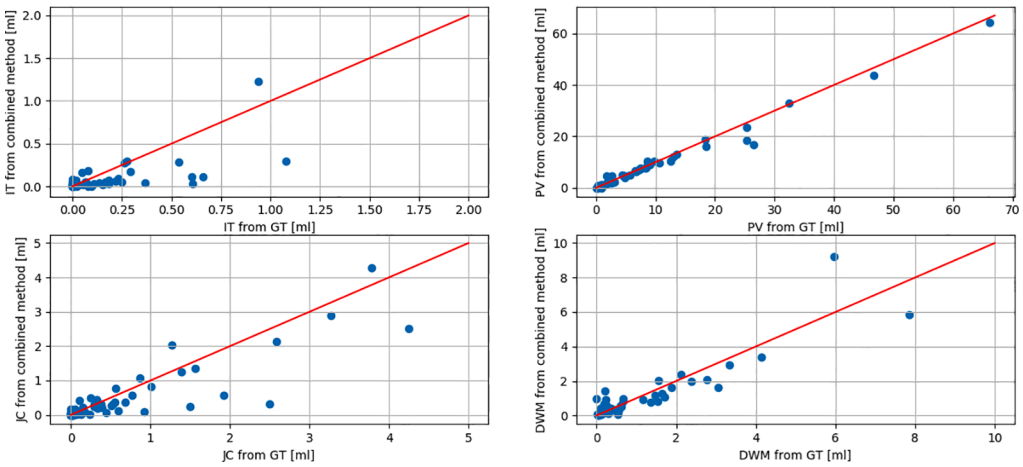


Fig. 5. The combined method demonstrates a favorable trend in volumetric estimation of PV and DWM classes, displaying the volumetric intersection points in relatively close proximity to the identity line shown in red. Each dot represents one subject from the test set. Pearson's correlation coefficients for each class are 0.66 for IT, 0.99 for PV, 0.86 for JC and 0.90 for DWM (IT – infratentorial, PV – periventricular, JC – juxtacortical, DWM – deep white matter, GT – ground truth).

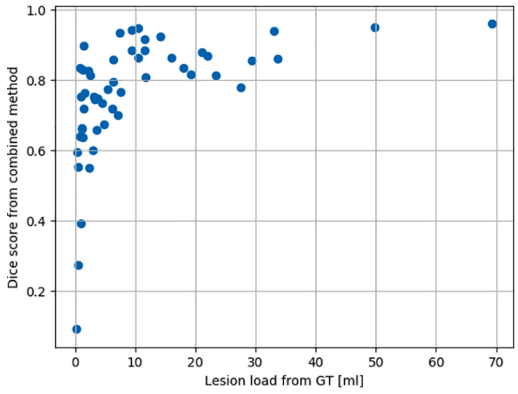


Fig. 6. Apart from several subjects with low lesion loads that are poorly segmented in terms of Dice score, there is an observable trend of relatively consistent voxel-wise Dice score values across varying lesion volumes from the ground truth, especially for lesion loads greater than 10 ml. Each dot represents one subject from the test set.

data that lacks in variability, mainly due to the fact that it comes from a single scanner (Parry et al., 2002; Vrenken et al., 2010; Alfano et al., 2000). Multi-center data is crucial to demonstrate robustness and perform proper validation of the method, as MRI acquisition protocols differ greatly, and the probability distribution of voxel intensities varies based on scanner-dependent parameters that are set during every acquisition. In a review paper published by García-Lorenzo et al. (2013), out of 48 reviewed papers, only 13 had used data coming from more than one scanner, 11 out of which used data from 2 different scanners, and only 2 had used multi-scanner data, which demonstrates the lack of variability that is often present and the need for a well-designed validation framework that would allow for inter-method result comparisons. To our knowledge, the dataset introduced at 2008 MICCAI MS Lesion Segmentation Challenge (Styner et al., 2008) has been one of the best attempts to provide a solid ground for validation and comparison of the developed methods. Even though it is considered to be a large dataset, taking into consideration the average number of subjects considered per study, it consists of 20 training cases that come from only two different scanners. In contrast, just the test set used to conduct the study presented in this paper already consists of 51 subjects coming from 22 different scanner models from 4 major scanner manufacturers. This alone showcases the great difference in dataset variabilities that prevent any fair qualitative or quantitative comparisons with previously developed approaches. The 2016 MICCAI MS Lesion Segmentation Challenge dataset (Commowick et al., 2018) is another publicly available dataset including 15 training scans acquired using 3 different scanner vendors from 3 different imaging centers, and while it does contain manual segmentations created from the fusion of 7 different raters to reduce inter- and intra-rater variability, all the lesions are treated as a single class, as opposed to 4 different ones in our approach.

The obtained voxel-wise Dice scores when evaluating our methods on the training set of 2016 MICCAI MS Lesion Segmentation Challenge were 0.58 for ico**brain ms** 5.0 and 5.1 and 0.62 for the attention-gate U-net. Compared to the reported range of values [0.41–0.62] (McKinley et al., 2021), where several methods were evaluated on the 2016 MICCAI training dataset (in a similar fashion to our approach), these seem to be in line. The same paper by McKinley et al. (2021) showed how misleading it is to train and test on non-independent data, since they got much better performance in their cross-validation experiment that used the 2016 MICCAI training dataset only. Another notable result is the Dice coefficient of 0.70 reported in Hashemi et al. (2018), where the model was trained using the 2016 MICCAI MS Lesion Segmentation Challenge training dataset and evaluated on the corresponding test set, which includes data coming from the same scanners and delineated by the same set of experts. Therefore, our model is at a disadvantage in a sense that it is being evaluated on an unseen type of data. Cerri et al. (2021) reported the Dice value of 0.65, but there is still a significant difference in the problem statement that distinguishes our work from most of the others, which is the fact that the multi-class approach adds another dimension in terms of complexity.

There is also a wide range of metrics different studies have used to quantify the obtained results. Many of those are voxel-based similarity metrics, whereas others focused on lesion-level metrics (Goldberg-Zimring et al., 1998; Sajja et al., 2006; Styner et al., 2008; Yamamoto et al., 2010), such as lesion-wise Dice score. The latter are less sensitive to the manual delineation variabilities, which was one of the reasons we opted for these metrics in our experiments.

5. Conclusions

In this paper, we proposed a method for segmentation and classification of MS lesions, which improves the detection of small and underrepresented lesions (e.g. infratentorial and juxtacortical). The method, based on a combined approach which entails traditional unsupervised machine learning techniques on one side, and deep-learning attention-gate U-net on the other, was validated on a large multi-scanner, multi-center dataset. We trained the U-net in order to tackle

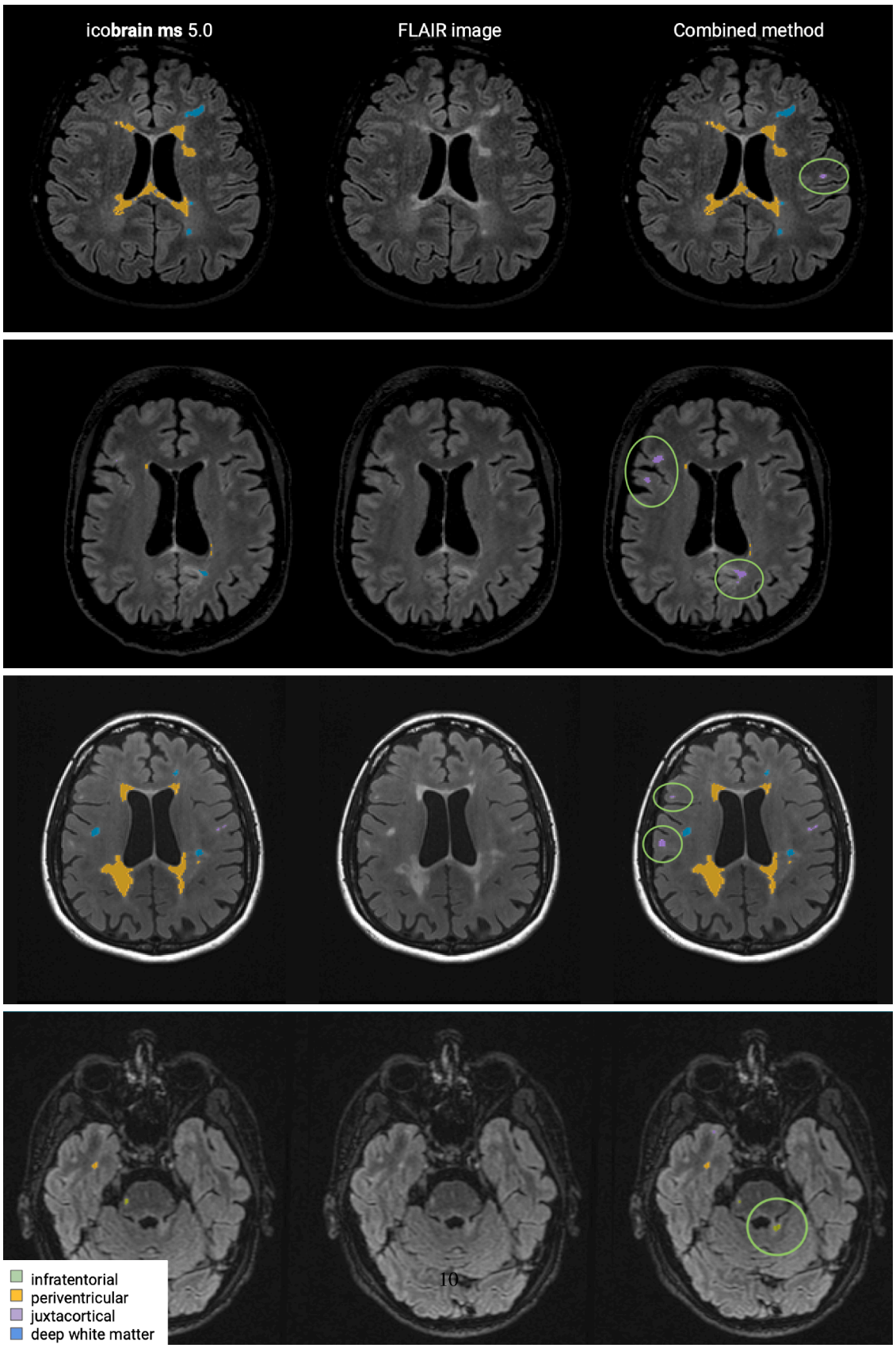


Fig. 7. Combined method (right column) that combines unsupervised segmentation (left column) with a deep learning network equipped with attention gates allows better detection/classification/segmentation performance of MS lesions on FLAIR images (middle column), especially of small juxtacortical and infratentorial lesions. The examples have been selected to reflect some of the most extreme cases in terms of differences in performance between the two methods; the difference was quantified using 90th percentile of the absolute error in ml.

the main weakness of the former method, which is the detection of infratentorial and juxtacortical plaques in particular. Our results on independent test data illustrate the improvement in detection of these two classes when using the U-net, both qualitatively and quantitatively, through lesion-wise Dice scores (U-net shows improvement of 14% for infratentorial and 28% for juxtacortical plaques, with respect to ico**brain ms** 5.0). The combined method uses a simple label fusion of both methods, which outperforms either method individually and increases the improvement on juxtacortical lesion detection to 31% with respect to ico**brain ms** 5.0. To make the method as robust as possible, we applied data stratification strategies to make training and testing sets variable enough, but without introducing a potential bias during the network training stage. Comparison of the achieved results with the other works is mainly challenged by the lack of a standardized and large enough data set that includes multi-center data and is thus variable enough. Still, the validation of our method on the 2016 MICCAI dataset demonstrates a good level of consistency when it comes to lesion detection on an independent, publicly available dataset. Since the new deep-learning pipeline is now a part of a clinical product ico**brain ms** 5.1 that is FDA and CE approved and is constantly being validated on real-world clinical data, this method will be continually refined based on direct feedback from clinical users worldwide.

6. Ethical disclaimer about the usage of data

After bringing ico**brain ms** on the market, ico**metrix** has analysed thousands of MR scans from MS patients as part of clinical practice in many hospitals worldwide. As of February 1, 2020, 6239 potential MS subjects, having both T1 and FLAIR images and comprising 44 projects (with one or more centers), had agreed to allow ico**metrix** to use an anonymised version of the already analysed MR images for post-market research purposes. Data is only considered anonymous if it cannot be used in any manner to identify the subject. ico**metrix** processes personal data received from the hospitals in conformity with the applicable data protection and privacy legislation. ico**metrix** has sufficient security measures to guarantee this, including procedures to contact the hospitals in the event of information security incidents. Authors Mladen Rakić, Sophie Vercruyssen, Simon Van Eyndhoven, Ezequiel de la Rosa, Saurabh Jain, Dirk Smeets and Diana M. Sima are all (or have been at the time of writing) employees of ico**metrix** (Leuven, Belgium).

CRediT authorship contribution statement

Mladen Rakić: Investigation, Methodology, Software, Writing - original draft. **Sophie Vercruyssen:** Data curation, Software, Validation. **Simon Van Eyndhoven:** Formal analysis, Writing - review & editing. **Ezequiel de la Rosa:** Methodology, Software. **Saurabh Jain:** Conceptualization, Methodology, Software. **Sabine Van Huffel:** Supervision, Writing - review & editing. **Frederik Maes:** Supervision, Writing - review & editing. **Dirk Smeets:** Supervision, Conceptualization, Funding acquisition. **Diana M. Sima:** Supervision, Conceptualization, Writing - review & editing, Funding acquisition.

Acknowledgments

The authors would like to acknowledge Prof. Guy Nagels for valuable feedback on manual delineation protocol and corrections, Delphine Segaert for contributions to the manual delineation task, as well as Joep Stevens for his attention-gate block code. This work was supported by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreements No 813120 and No 765148. Sabine Van Huffel and Frederik Maes received funding from the Flemish Government (AI Research Program). Sabine Van Huffel and Frederik Maes are affiliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium.

References

Alfano, B., Brunetti, A., Larobina, M., Quarantelli, M., Tedeschi, E., Ciarmiello, A., Covelli, E.M., Salvatore, M., 2000. Automated segmentation and measurement of global white matter lesion volume in patients with multiple sclerosis. Journal of Magnetic Resonance Imaging 12, 799–807.

Ashton, E.A., Takahashi, C., Berg, M.J., Goodman, A., Totterman, S., Ekholm, S., 2003. Accuracy and reproducibility of manual and semiautomated quantification of MS lesions by MRI. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 17, 300–308.

Bricq, S., Collet, C., Armspach, J.P., 2008. Lesions detection on 3D brain MRI using trimmmed likelihood estimator and probabilistic atlas. In: 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, IEEE, pp. 93–96.

Brust, J.C., 2018. Current Diagnosis & Treatment Neurology. McGraw Hill Professional.

Calabresi, P.A., 2004. Diagnosis and management of multiple sclerosis. American Family Physician 70, 1935–1944.

Cerri, S., Puonti, O., Meier, D.S., Wuerfel, J., Mühlau, M., Siebner, H.R., Van Leemput, K., 2021. A contrast-adaptive method for simultaneous whole-brain and lesion segmentation in multiple sclerosis. NeuroImage 225, 117471.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-net: learning dense volumetric segmentation from sparse annotation. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer 424–432.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S.C., Girard, P., Ameli, R., Ferré, J.C., et al., 2018. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Scientific Reports 8, 1–17.

Crum, W.R., Camara, O., Hill, D.L., 2006. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Transactions on Medical Imaging 25, 1451–1461.

Evans, A.C., Marrett, S., Neelin, P., Collins, L., Worsley, K., Dai, W., Milot, S., Meyer, E., Bub, D., 1992. Anatomical mapping of functional activation in stereotactic coordinate space. Neuroimage 1, 43–53.

Filippi, M., Horsfield, M., Bressi, S., Martinelli, V., Baratti, C., Reganati, P., Campi, A., Miller, D., Comi, G., 1995. Intra-and inter-observer agreement of brain MRI lesion volume measurements in multiple sclerosis: a comparison of techniques. Brain 118, 1593–1600.

Forbes, F., Doyle, S., Garcia-Lorenzo, D., Barillot, C., Dojat, M., 2010. Adaptive weighted fusion of multiple MR sequences for brain lesion segmentation. In: 2010 IEEE international symposium on biomedical imaging: from nano to macro, IEEE, pp. 69–72.

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D.L., Collins, D.L., 2013. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Medical Image Analysis 17, 1–18.

Goldberg-Zimring, D., Achiron, A., Miron, S., Faibel, M., Azhari, H., 1998. Automated detection and characterization of multiple sclerosis lesions in brain MR images. Magnetic Resonance Imaging 16, 311–318.

Goldenberg, M.M., 2012. Multiple Sclerosis Review. Pharmacy and Therapeutics 37, 175.

Grimaud, J., Lai, M., Thorpe, J., Adeleine, P., Wang, L., Barker, G., Plummer, D., Tofts, P., McDonald, W., Miller, D., 1996. Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques. Magnetic Resonance Imaging 14, 495–505.

Hashemi, S.R., Salehi, S.S.M., Erdogmus, D., Prabhu, S.P., Warfield, S.K., Gholipour, A., 2018. 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.

Hauser, S.L., Goodin, D.S., 2005. Multiple sclerosis and other demyelinating diseases. Harrisons Principles of Internal Medicine 16, 2461.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al., 2018. nnU-net: Self-adapting framework for U-net-based medical image segmentation. arXiv preprint arXiv:1809.10486.

Jain, S., Sima, D.M., Ribbens, A., Cambron, M., Maertens, A., Van Hecke, W., De Mey, J., Barkhof, F., Steenwijk, M.D., Daams, M., et al., 2015. Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. NeuroImage: Clinical 8, 367–375.

Johnston, B., Atkins, M.S., Mackiewich, B., Anderson, M., 1996. Segmentation of multiple sclerosis lesions in intensity corrected multispectral MRI. IEEE Transactions on Medical Imaging 15, 154–169.

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., et al., 2017. Ensembles of multiple models and architectures for robust brain tumour segmentation. International MICCAI Brainlesion Workshop, Springer. 450–462.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. Imagenet classification with deep convolutional neural networks. Communications of the ACM 60, 84–90.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models, in: Proc. icml, p. 3.

McDonald, W.I., Compston, A., Edan, G., Goodkin, D., Hartung, H.P., Lublin, F.D., McFarland, H.F., Paty, D.W., Polman, C.H., Reingold, S.C., et al., 2001. Recommended diagnostic criteria for multiple sclerosis: guidelines from the international panel on the diagnosis of multiple sclerosis. Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society 50, 121–127.

McKinley, R., Wepfer, R., Aschwanden, F., Grunder, L., Muri, R., Rummel, C., Verma, R., Weisstanner, C., Reyes, M., Salmen, A., et al., 2021. Simultaneous lesion and brain

segmentation in multiple sclerosis using deep neural networks. Scientific Reports 11, 1–11.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention U-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Parodi, R.C., Sardanelli, F., Renzetti, P., Rosso, E., Losacco, C., Ferrari, A., Levrero, F., Pilot, A., Inglese, M., Mancardi, G.L., 2002. Growing region segmentation software (GRES) for quantitative magnetic resonance imaging of multiple sclerosis: intra-and inter-observer agreement variability: a comparison with manual contouring method. European Radiology 12, 866–871.

Parry, A., Clare, S., Jenkinson, M., Smith, S., Palace, J., Matthews, P.M., 2002. White matter and lesion T1 relaxation times increase in parallel and correlate with disability in multiple sclerosis. Journal of Neurology 249, 1279–1286.

Rakić, M., Cabezas, M., Kushibar, K., Oliver, A., Lladó, X., 2020. Improving the detection of autism spectrum disorder by combining structural and functional MRI information. NeuroImage: Clinical 25, 102181.

Sajja, B.R., Datta, S., He, R., Mehta, M., Gupta, R.K., Wolinsky, J.S., Narayana, P.A., 2006. Unified approach for multiple sclerosis lesion segmentation on brain MRI. Annals of Biomedical Engineering 34, 142–151.

Shiee, N., Bazin, P.L., Ozturk, A., Reich, D.S., Calabresi, P.A., Pham, D.L., 2010. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. NeuroImage 49, 1524–1535.

Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., Markovic-Plese, S., Jewells, V., Warfield, S., 2008. 3D segmentation in the clinic: A grand challenge II: MS lesion segmentation. Midas Journal 2008, 1–6.

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J., 2017. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 240–248.

Udupa, J.K., Wei, L., Samarasekera, S., Miki, Y., van Buchem, M.A., Grossman, R.I., 1997. Multiple sclerosis lesion quantification using fuzzy-connectedness principles. IEEE Transactions on Medical Imaging 16, 598–609.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2016. Instance normalization: The missing ingredient for fast stylization arXiv preprint arXiv:1607.08022.

Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À., Oliver, A., Lladó, X., 2017. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. NeuroImage 155, 159–168.

Vrenken, H., Seewann, A., Knol, D., Polman, C., Barkhof, F., Geurts, J., 2010. Diffusely abnormal white matter in progressive multiple sclerosis: in vivo quantitative MR imaging characterization and comparison between disease types. American Journal of Neuroradiology 31, 541–548.

Weinshenker, B.G., 1996. Epidemiology of multiple sclerosis. Neurologic Clinics 14, 291–308.

Yamamoto, D., Arimura, H., Kakeda, S., Magome, T., Yamashita, Y., Toyofuku, F., Ohki, M., Higashida, Y., Korogi, Y., 2010. Computer-aided detection of multiple sclerosis lesions in brain magnetic resonance images: False positive reduction scheme consisted of rule-based, level set method, and support vector machine. Computerized Medical Imaging and Graphics 34, 404–413.

Younis, A.A., Soliman, A.T., Kabuka, M.R., John, N.M., 2007. MS lesions detection in MRI using grouping artificial immune networks. In: 2007 IEEE 7th International Symposium on BioInformatics and BioEngineering, IEEE, pp. 1139–1146.

Zijdenbos, A.P., Forghani, R., Evans, A.C., 2002. Automatic pipeline analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. IEEE Transactions on Medical Imaging 21, 1280–1291.