OPEN ACCESS
International Journal of Neural Systems, (2025) 2550006 (19 pages)
© The Author(s)
DOI: 10.1142/S0129065725500066

A Context-Dependent CNN-Based Framework for Multiple Sclerosis Segmentation in MRI

Giuseppe Placidi*
A^{2}VI-Lab c/o Department of Life, Health & Environmental Sciences
University of L'Aquila, L'Aquila, Italy
giuseppe.placidi@univaq.it

Luigi Cinque
Department of Computer Science
Sapienza University of Rome, Rome, Italy

Gian Luca Foresti
Department of Mathematics, Computer and Physics Science
University of Udine, Udine, Italy

Francesca Galassi
Univ Rennes, CNRS, Inria, Inserm, IRISA UMR 6074
EMPENN - ERL U 1228, F-35000 Rennes, France

Filippo Mignosi
Department of Information Engineering, Computer Science and Mathematics
University of L'Aquila, L'Aquila, Italy

Michele Nappi
Department of Computer Science
University of Salerno, Fisciano, Italy

Matteo Polsinelli
Department of Management & Innovation Systems
University of Salerno, Fisciano, Italy

Received 22 August 2024
Accepted 5 November 2024
Published Online 13 December 2024

Despite several automated strategies for identification/segmentation of Multiple Sclerosis (MS) lesions in Magnetic Resonance Imaging (MRI) being developed, they consistently fall short when compared to the performance of human experts. This emphasizes the unique skills and expertise of human professionals in dealing with the uncertainty resulting from the vagueness and variability of MS, the lack of specificity of MRI concerning MS, and the inherent instabilities of MRI. Physicians manage this uncertainty in part by relying on their radiological, clinical, and anatomical experience. We have developed an automated framework for identifying and segmenting MS lesions in MRI scans by introducing a novel approach to replicating human diagnosis, a significant advancement in the field. This framework has the potential to revolutionize the way

*Corresponding author.

MS lesions are identified and segmented, being based on three main concepts: (1) Modeling the uncertainty; (2) Use of separately trained Convolutional Neural Networks (CNNs) optimized for detecting lesions, also considering their context in the brain, and to ensure spatial continuity; (3) Implementing an ensemble classifier to combine information from these CNNs. The proposed framework has been trained, validated, and tested on a single MRI modality, the FLuid-Attenuated Inversion Recovery (FLAIR) of the MSSEG benchmark public data set containing annotated data from seven expert radiologists and one ground truth. The comparison with the ground truth and each of the seven human raters demonstrates that it operates similarly to human raters. At the same time, the proposed model demonstrates more stability, effectiveness and robustness to biases than any other state-of-the-art model though using just the FLAIR modality.

Keywords: Multiple sclerosis; MRI; convolutional neural network; U-Net; FLAIR; segmentation; classification; uncertainty.

1. Introduction

Multiple Sclerosis (MS) is a degenerative disease mainly affecting the white matter (WM) and the spinal cord. It has a very heterogeneous clinical presentation across patients in terms of both severity and symptoms.[1] The origins of the disease are not well understood but the characteristic signs of tissue degeneration are the presence of lesions and brain atrophy. Magnetic Resonance Imaging (MRI) enables the observation of most signs of MS and has become the preferred minimally invasive tool for monitoring lesions.[2] Focal lesions are primarily visible in the WM on structural MRI. They are observable as hyper-intensities in T_2-weighted (T_2w) images, proton-density (PD) images, and FLuid-Attenuated Inversion Recovery (FLAIR) images, and as hypointensities in T_1-weighted (T_1w) images. Radiologists often use FLAIR for detecting WM lesions and other modalities to refine their borders, ascertain the presence of cortical lesions, or confirm the choice made with FLAIR.

In an examination, thousands of images are collected before and after contrast administration. However, MRI is not specific to MS and not well correlated with impairment progression, neuroplasticity, and the effects of demyelination of nerves.[2] Lesions and healthy tissue often share the same volume, resulting in the partial volume effect (PVE). Additionally, healthy anatomical structures similar to lesions and close to lesions could contribute to creating further ambiguity in MRI. The wide range of variations in images caused by differences in scanners, magnetic field strength/homogeneity, and parameter settings adds complexity to the framework.[3] While attempts have been made to standardize amplitudes in MRI,[4] the results remain unsatisfactory due to the number of involved variables. As a result, when defining the borders of lesions and identifying entire lesions, disagreement can arise among radiologists as well as uncertainty within each radiologist. This could make manual segmentation not only time-consuming and tedious but also inaccurate, despite the radiologists' extensive experience.

In recent times, several automated frameworks have been proposed and reviewed for this problem.[5–11] However, their results still lag behind those of human experts, leading to an increase in model complexity without the expected improvement materializing[12] and, in some cases, introducing framework-specific biases. Additionally, even though automated strategies have advanced, they still struggle to fully capture the complex medical expertise, human judgment, and adaptability needed for MRI analysis. Finally, automatic strategies do not completely replicate the reasoning process used by radiologists for volume analysis, which involves segmenting 2D axial slices and continuously examining coronal and sagittal images.[2,13] In a recent paper,[14] the significance of 3D Convolutional Neural Networks (CNNs) in MS lesion segmentation is emphasized. However, other recent frameworks,[15,16] favor ensembles of 2D U-Net models, which align with human methodology, due to the anisotropic spatial resolution across the three axes, with a preference for axial planes. Besides, the uncertainty that impacts radiologists during classification is not well-documented in public, binary-labeled datasets. This lack of representation of an expert's "fuzzy" evaluation through a binary choice is often insufficient. Modeling uncertainty could significantly improve segmentation[17] and its impact on knowledge transfer to an automated strategy[18] deserves further study.

We propose a framework to narrow the gap with human experts by implementing the following: (1) Categorizing uncertainty as an intermediary class between background and lesions; (2) Optimizing two CNNs (2D U-Net models), one for the lesion class and one for the lesions contextualized in the brain; (3) Repeating the process for all three spatial directions (axial, coronal, and sagittal) to maintain volume continuity; (4) Defining an ensemble classifier to consolidate the information gathered by all CNNs.

To achieve this goal, we did the following: We used a publicly available large-scale benchmark MRI database of brain images and the corresponding ground truth, as suggested in the MICCAI MS Lesion Segmentation Challenges (MSSEG),[19] ensuring the robustness of our approach; We identified the uncertain regions using the binary classifications of seven human raters in MSSEG. We specifically designed and applied our framework to FLAIR images, underscoring the importance of this imaging modality in our research.

These options enabled us to compare the proposed framework to other competitive automated strategies and seven human experts. The MSSEG dataset's uniqueness for our purposes stems from the presence of data annotated by seven raters, leading us not to consider other important benchmark datasets such as MSSEG2.[20] Moreover, we demonstrated how uncertainty modeling could help in reducing the ambiguity and complexity of the problem and that a single imaging sequence, FLAIR, is sufficient for segmenting MS lesions affecting WM.

This paper is structured as follows. Section 2 presents a review of automatic approaches to MS lesion identification/segmentation, particularly those using CNNs. Section 3 describes the proposed framework in the context of the used data set, the defined three-class consensus used for training, the proposed CNN architecture, and the ensemble system. Section 4 details the indicators used for the comparison. Section 5 reports and discusses experimental results. Section 6 concludes the paper and presents some constructive hints for future investigations.

2. Related Work

Automated strategies are widely used for medical image analysis, mainly of brain images.[21–24] One area of ongoing research is the automated segmentation of MS lesions, with numerous methods being developed and extensively reviewed over time.[5,6,8,25,26]

Automated strategies can be categorized into three main groups: methods using pre-selected features (PSFs), methods using prior information (API), and deep learning (DL).

Some PSFs utilize a wide range of features and then select the ones that are most distinctive through labeled training.[27] Other approaches use topological and statistical atlases[28,29] or Decision Random Forests.[30] Likewise, a framework for segmenting lesions enhanced by contrast agents using conditional random fields was outlined in Ref. 31. The work in Ref. 32 introduces a set of features, including contextual features, registered atlas probability maps, and an outlier map, to automatically segment MS lesions through a voxel-wise approach. Additionally, a rotation-invariant multi-contrast nonlocal means segmentation method was proposed in Ref. 33 for the identification and segmentation of lesions from 3D MRI. Supervised learning through PSF has been widely employed in tasks where the training database and the PSF set cover all possible cases.[34] However, when dealing with MS and MRI, exhibiting significant heterogeneity and huge variability, the size of the training data set and, most importantly, the choice of PSFs, become critical.

API does not require labeled data for training but usually exploits some prior information, such as intensity clustering, to model tissue distribution.[35] In Ref. 36, the distribution of intensities in MRI of healthy brains was modeled by a likelihood estimator. Other methods use threshold with post-processing refinement[37] or probabilistic approaches.[38] A big challenge for API is represented by the outliers that could be due to artifacts, intensity inhomogeneity, and small anatomical structures like blood vessels.[39] Moreover, API heavily relies on the information extracted and managed by the knowledge of specific experts.

Compared to the other categories, DL models extract features directly from data,[40] that makes them particularly suitable for medical imaging,[22,41–47] mainly for studying neurodegenerative diseases.[23,48–52] The recent popularity of DL is due to the U-Nets and their variants.[53–57] While the size and quality of the training dataset are important for DL, the pre-selection of features, such as in PSF, or prior modeling, as in API, are not critical. Comparatively, CNNs have shown significant success in biomedical image analysis, outperforming traditional machine learning approaches.[16,39,41,42] DL has been particularly applied in MS lesion identification and segmentation, as confirmed by recent studies.[5,6,8,10,11,25,58–60]

CNNs applied to MS utilize 2D or 3D convolutional layers[40,57] to integrate spatial and temporal information within the model and a minimum lesion volume threshold to exclude small false positives.

Some methods improve the MS segmentation by retraining the CNN from the first layers in a kind of self-supervision by hypothesizing that those layers are more informative.[61] Hybrid models, such as Decision Tree (DT) and Support Vector Machines (SVM), combine textural features with machine learning approaches to detect MS lesions.[62] Finally, Federated Learning strategies have successfully improved the segmentation of MS lesions.[56] However, all the above strategies still lag behind human experts and perform poorly on inhomogeneous data sets.[12,56]

3. The Proposed Framework

The framework we propose, outlined in Fig. 1, consists of the following steps:

(1) Automatically classify the cross-sectional images (2D) that make up the MRI volume on a voxel-by-voxel basis, separately for axial, coronal, and sagittal sections (step 1 in Fig. 1).
(2) Fuse the classifications (step 2a in Fig. 1) and then use a majority vote to confirm the classification (step 2b in Fig. 1).
(3) Produce the final output (step 3 in Fig. 1).

The following points are important:

(1) Three classes are considered: Background, Uncertainty, and Lesion (from now on, the capital letter indicate the name of the class). The definition of Uncertainty is described in Sec. 3.2.
(2) We are optimizing two CNNs for each imaging section: one for Lesion (lesion tuned) and the other for Lesion in the context of the whole brain (brain tuned).
(3) For the class confirmation process, the three classes are expected to be arranged in the following order: Lesion > Uncertainty > Background. According to this order, the confirmation starts from Lesion, followed by the upgraded Uncertainty. If the class is not confirmed, it is downgraded by one class.
(4) Just one MRI modality, FLAIR, is used.

In Fig. 1, step 2a is used to incorporate additional information from the specific characteristics of each of the two axial CNNs. It also aims to simulate the decision-making process of radiologists who use axial orientations to form initial hypotheses. Step 2b involves voting on whether the classes resulting from 2a (Lesion or Uncertainty) should remain in the original assignment or be downgraded (from Lesion to Uncertainty and from Uncertainty to Background). In this process, objects are considered confirmed only if at least two of the other four classifications (two coronal and two sagittal) agree with the axial classification. This approach helps reduce false positives, strengthens the consideration of the 3D environment, and mimics the procedure followed by radiologists.

Each classifier is trained, validated, and tested separately. In the following sections, we describe the dataset used, the ground truth including Uncertainty, the CNN-based architecture, the loss function, the hyperparameter optimization, and the final ensemble classification in Fig. 1.

3.1. *MSSEG dataset*

We have chosen the MSSEG dataset[19] for our study as it allows for direct comparison and benchmarking of our proposed framework with state-of-the-art segmentation methodologies and with the assessments of seven independent expert radiologists who annotated data in MSSEG. Indeed, the MSSEG dataset consists of MRI brain images from 53 cases collected in different centers using scanners operating at different magnetic fields: 1.5T Siemens Aera, 3T Siemens Verio, 3T Philips Ingenia, and 3T General Electric Discovery. Each examination in the dataset includes the following imaging sequences: T_1-w, gadolinium-enhanced T_1-w (T_1-w Gd), T_2-w, T_2-FLAIR, and PD-weighted images. Besides the annotated data, MSSEG contains one consensus, a statistical "average" among the assessments of the seven radiologists, used as the ground truth.[19]

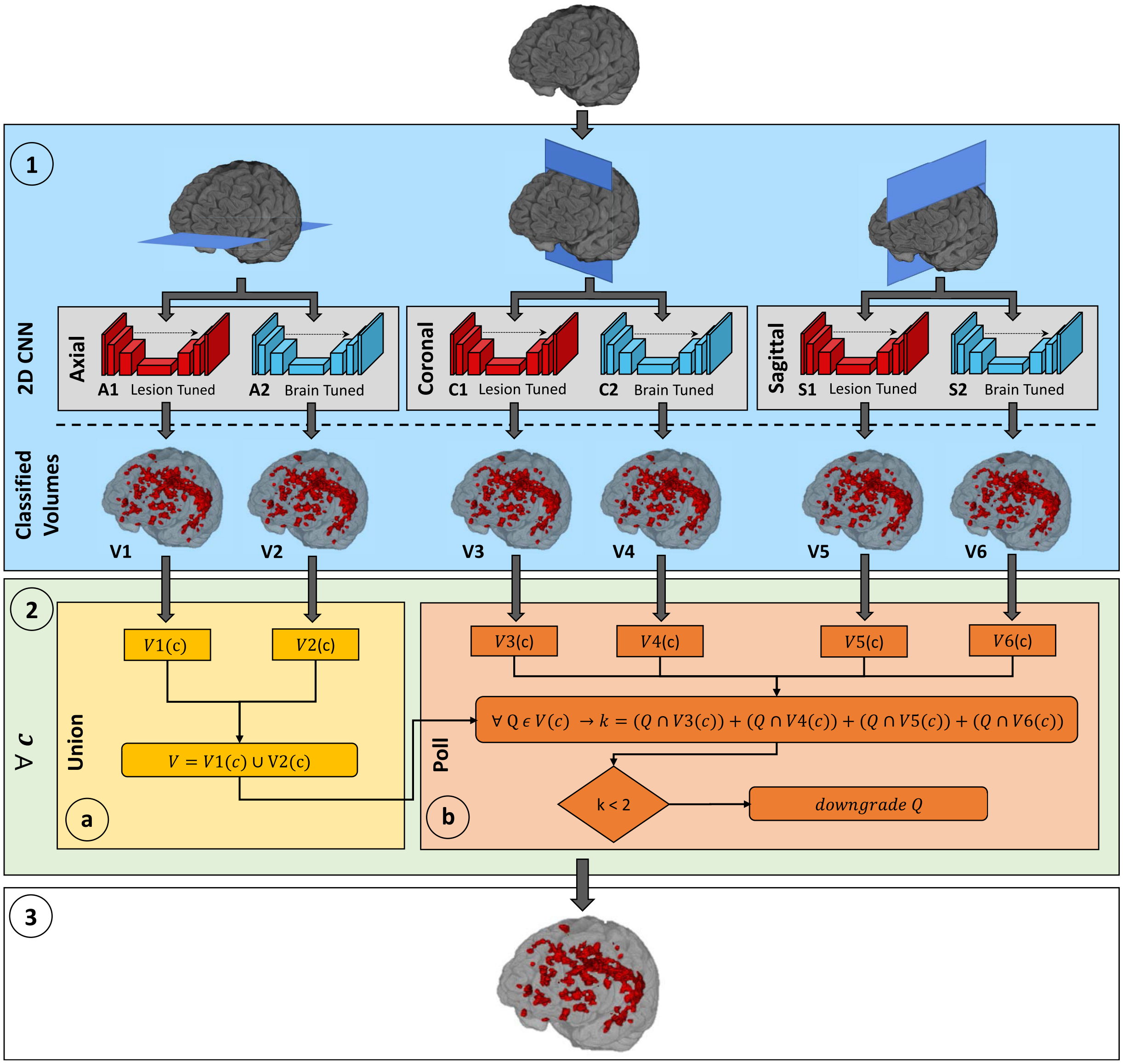


Fig. 1. The model works with three classes: Background (not lesions), Uncertainty (tissues of uncertain nature), and Lesion (lesions). The approach independently processes axial, coronal, and sagittal sections (1), each handled by two separately trained U-Nets. One U-Net is optimized for focusing directly on lesions, while the other is optimized for understanding lesions in the context of the brain. The results are combined by using the Union of axial volumes V1 and V2 (2a), followed by a majority vote strategy on the coronal (V3 and V6) and sagittal (V5 and V6) volumes for confirmation (2b). Voxel classifications that are not confirmed are downgraded: Lesion becomes Uncertainty and Uncertainty becomes Background. The framework operates independently for Lesion and Uncertainty, starting with Lesion. In the confirmation stage, the procedure is voxel-wise, for every single voxel Q belonging to the class $c \in \{\text{Lesion}, \text{Uncertainty}\}$. The segmented volume is the output (3).

The radiologists were tasked with performing binary segmentation, where each voxel was identified as either a lesion or background. The annotated dataset consists of 15 training cases and 38 testing cases.

The images were anonymized and provided in both their original form and pre-processed form to reduce noise and imaging artifacts, equalize space, eliminate outliers, stabilize contrast, and skull stripping.[63] For our purposes, we used the pre-processed data in MSSEG. Additionally, as the MSSEG data from different scanners have varying slice thicknesses,[6] we integrated the pre-processing pipeline with a final interpolation module to standardize our model with a single slice thickness value, which we set at 1 mm. Of the MSSEG, we utilized data solely from the FLAIR modality to detect and segment MS lesions in the WM. We demonstrated that using FLAIR alone is sufficient for MS lesion segmentation in the WM. We divided the dataset of 15 subjects into two subsets for training and validation. The training set comprises data from 12 subjects, with four subjects from each center, while the validation set includes data from three subjects, with one subject from each center. To ensure comprehensive cross-validation, the proportions between centers were maintained. Upon establishing the data set, each image was augmented by adding two random rotations (between -13° and 13° with a resolution of 1°), 1 random scaling (between 1.1 and 1.3 with a resolution of 0.01), and 1 Gaussian random noise addition with a mean of 0 and a variance of 0.001A, where A is the maximum amplitude value in the examined volume. Data augmentation enhanced the model's ability to handle rotation, scaling, and noise. The augmented data set for each orientation (axial, coronal, and sagittal) consisted of 5634 images for training and 435 images for validation. For testing, the data set included 38 subjects. Notably, one of the tested subjects had no lesions.[6] For this specific case, a separate evaluation was conducted based on the number of detected lesions, with the ideal value being 0, in line with the recommendations for objective evaluation,[6] also described below.

3.2. *The ground-truth including uncertainty*

In medical imaging regarding MS, it is often assumed that there exists a single, unknown, true segmentation map of the underlying anatomy and that each radiologist produces an approximation with variations reflecting individual experience.[64] On the other hand, it can also be assumed that variable annotations from experts are all realistic and acceptable instances of true segmentation. It's important to consider that the truth often lies somewhere in between: an ideal and unique true segmentation is impossible, due to the unpredictability of MS and the nonspecificity of MRI, and, at the same time, not all the variability in expert annotations are acceptable instances of an ideal true segmentation, due to human mistakes or oversights. However, human subjectivity is a significant factor that cannot be overlooked. Its effects are due to a combination of prior assumptions, such as experience in the field, utilization of additional meta-information (e.g. anatomical/radiological/clinical knowledge), mistakes, or oversights, particularly regarding small and/or low-intensity lesions or their borders. Recognizing the importance of understanding and addressing human subjectivity is a crucial step toward improving training and standardizing practices in medical imaging.

When radiologists are required to provide a simple "yes" or "no" answer, as in the case of MSSEG, they are unable to convey any uncertainty caused by the ambiguities mentioned above. This can lead to decisions that may not truly reflect the rater's beliefs and could also be confusing for an automated system. In similar situations, a rater may have to make ambiguous decisions (like deeming an uncertain area as healthy in one instance and as a lesion in another), which could further confuse the automated system.[18]

To train a model to recognize problem-specific uncertainty, we needed to combine binary ground truth with human uncertainty. To maintain the original ground truth, we identified voxels as uncertain if at least three out of seven human raters considered them as lesions while the binary consensus did not. This approach created room for Uncertainty while keeping the binary consensus for Lesion unchanged, allowing comparison with current strategies used for lesion detection. It's important to note that in this context, introducing Uncertainty helps to reduce ambiguity in identifying Lesion, rather than aiming to provide an optimal definition of Uncertainty, which is beyond the scope of the paper.

Our definition of Uncertainty is different from others in the literature.[11,65] First, it aims to maintain the original structure of Lesion. Second, it accounts for the uncertainty affecting both the problem and the raters. Third, it avoids the possibility of the new class Uncertainty capturing part of the Lesion from the binary ground truth, making a direct comparison with other methods impossible. Fourth, it quantifies the improvement obtained when Uncertainty is introduced compared to when Uncertainty is not used. Lastly, it allows the learning strategy to consider not only lesion borders as uncertain, as other authors do,[65]

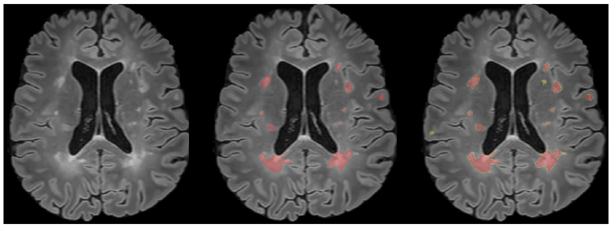


Fig. 2. (Color online) A FLAIR image from the MSSEG data set (left), the binary consensus (middle), and the proposed ternary consensus (right). The image is superimposed on both consensuses. Lesion (identical in both consensuses) is in red, Uncertainty in yellow.

but also entire regions. Figure 2 shows an example of a FLAIR image with the corresponding binary and ternary consensuses, the latter used to train the proposed framework.

3.3. *CNN architecture*

In this work, we utilized the U-Net "D" architecture,[53] where "D" refers to a variant incorporating a Dropout layer for regularization, as the foundational element of the suggested image-based segmentation framework (Fig. 3).

Compared to traditional architectures, we added a batch normalization layer in each block to mitigate the effects of gradient amplification around the lesions,[66] despite the 30% increase in computational overhead. We performed initial training to determine the optimal number of blocks, denoted as n, where $n \in \{3, 4, 5\}$. We focused on this range because with $n = 5$, the U-Net model started to overfit, even with high L_2-Regularization values. However, with $n = 2$, there was a noticeable decrease in performance. We observed that with $n = 4$, overfitting was mitigated, and the performance was acceptable, although some redundancy remained. In contrast, with $n = 3$, redundancy was significantly decreased, and training converged more quickly compared to $n = 4$. Consequently, we opted to proceed with $n = 3$ for subsequent procedures.

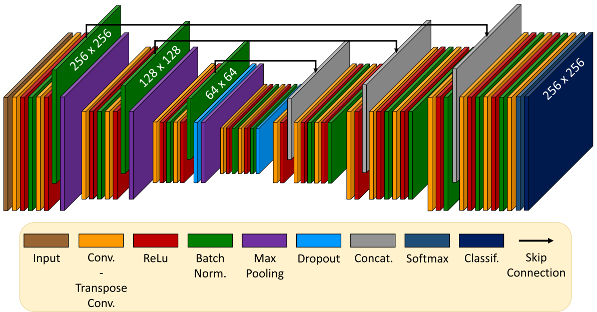


Fig. 3. The U-Net "D" Architecture used for the six classifiers in Fig. 1: A1 and A2 (axial), C1 and C2 (coronal), and S1 and S2 (sagittal).

3.4. *Loss function and hyperparameter optimization*

The proposed architecture had to solve a three-class annotation, for which the following Multi-label Cross Entropy Loss Function was used:

$$\text{loss} = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{K} (T_{n,i} \log(Y_{n,i}) + (1 - T_{n,i}) \times \log(1 - Y_{n,i})), \qquad (1)$$

where N is the number of observations, K is the number of classes, T is the true label, and Y is the predicted label.

The use of three classes enabled better confidence in defining both Lesion and Background through Uncertainty. This approach allowed for the optimization of two CNNs sharing the same loss function (Eq. (1)), with a different learning focus: Lesion and Lesion in the whole brain context. The class Uncertainty acts as a "*buffer class*". In the case of binary classification, this optimization would not have been possible, as what is not Lesion is Background and vice-versa. The use of Uncertainty provided both CNNs with a way to overcome this limitation. To achieve faster training and improved performance, we controlled the training process of each CNN using the following hyperparameters[67,68]:

(1) **Starting Learning Rate (LR):** determined by the dataset and the type of neural network.
(2) ***L*$_2$-Regularization (*L*$_2$-Reg): used to prevent overfitting.**
(3) **Class balancing:** using weights to balance the different cardinality of the classes. With three classes, two weights were needed (Lesion weight (LW) and Background weight (BW)), with the third being the complement to one of the others.

Of the above, the first two are standard for CNNs, while Class balancing is problem-specific as far as it helps to differentiate the paths of optimization.

The selected hyperparameters were optimized through a Bayesian approach[67] applied to the following optimization problem:

$$x^* = \text{argmin}_{x \in X} f(x), \tag{2}$$

where X is the space of solutions, x^* is the optimal hyperparameter setting we were searching for, and $f(x)$ the objective function:

$$f(x) = 1 - \text{IoU}(x), \tag{3}$$

where IoU is the Intersection over Union score[5] defined in Sec. 4. In our analysis, we calculated the IoU for the first CNN regarding the Lesion and for the second CNN regarding the Background. We optimized the hyperparameter settings differently for each CNN. Multiple short training sessions were conducted for each CNN with different hyperparameter configurations. The maximum number of epochs of each attempt was set to 15. In addition, an early stopping criterion was implemented: if the validation loss did not improve within the last five epochs, training was terminated early. The total number of short training sessions was constrained by the experimental duration, which was limited to 48 h. This time limit was imposed to ensure that all training experiments could be completed within a practical timeframe, striking a balance between computational efficiency and model performance exploration. For the training, the Adam optimizer was applied with a batch size of 4. The search space for the Bayesian optimization algorithm was limited to the following ranges of hyperparameter values: LR in [1E−4, 1E−2], L_2-Reg in [1E−10, 1E−2], LW in [0.02, 0.10], and BW in [0.78, 0.90]. These boundaries guided the optimization process, facilitating efficient model tuning. Table 1 reports the best hyperparameter setting obtained for the 6 CNN in our model. As can be observed, the overall difference in settings justifies different training paths for the CNN and different points of convergence for each of them. In Fig. 4, we illustrated the different behavior of the two CNNs in the segmentation grad-cams of a sample axial slice. The CNN optimized for Lesion tends to enlarge lesions and uncertainty compared to the CNN optimized for the lesion in the context of the whole brain.

Table 1. Optimal hyperparameter configurations calculated through the Bayesian approach for the 6 CNNs used in the model.

CNN	LR	L_2-Reg	LW	BW
Axial In	5.32E−04	3.66E−10	7.98E−02	8.91E−01
Axial Out	4.54E−04	1.14E−10	2.99E−02	8.99E−01
Cor. In	6.50E−04	3.43E−10	7.98E−02	8.71E−01
Cor. Out	3.18E−04	7.92E−09	6.59E−02	8.58E−01
Sag. In	1.01E−04	5.53E−09	7.01E−02	8.74E−01
Sag. Out	1.08E−04	3.41E−09	6.02E−02	8.50E−01

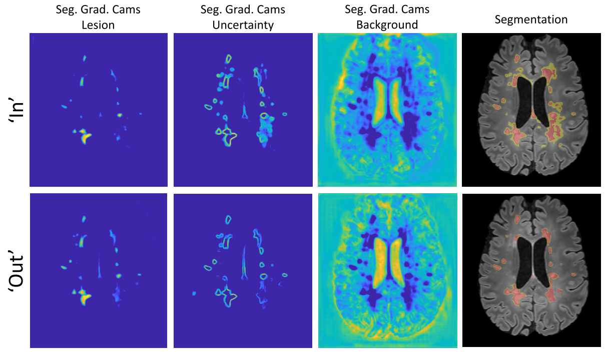


Fig. 4. Grad-cams showing the action of the two CNNs in a sample axial image: Lesions (In) and Lesions from the whole brain (Out): Lesion (first column), Uncertainty (second column), and Background (third column). The resulting segmentation is in the last column.

Once the optimal hyperparameter configurations were identified, an extended training of 100 epochs was performed for each of the 6 CNN. Under this configuration, each U-Net comprises 7.7 million parameters, resulting in a total of 46.2 million parameters across the entire framework. We used MATLAB 2024a to develop the CNNs and run relevant scripts in our experimental setup. The hardware we used for training the CNNs and running the scripts included a system with the following specifications: Windows 11 OS, an AMD Ryzen 5 3600X CPU, 32 GB of RAM, dual Nvidia RTX 2080 Super GPUs, each with 16 GB of VRAM, and two 1 TB SSDs.

3.5. *Ensemble classification*

It is a well-known fact that ensemble classifiers often outperform individual components, and they tend to be more stable as well.[15,69,70] For classification, we utilized 2D slices of the complete volume, with specific CNNs trained separately for Lesion and Lesion in the context of the whole brain, each for axial, radial, and sagittal orientations. This approach ensures that no particular orientation is biased towards lesions or the classifier. Additionally, it helps to provide context for lesions in the brain environment and to maintain 3D continuity.

We used six classifiers to mimic the decision-making process of radiologists. Although 3D FLAIR data were collected along sagittal planes, radiologists usually analyze axial slices for data interpretation and the other orientations are used for verification.[2,13] In the same way, we prioritized axial classifications and used coronal and sagittal outputs for confirmation.

In terms of axial classification, both CNNs gathered valuable contributions and, for this reason, a Union operation was necessary to combine the two classifications. This aligns with the double reading procedure followed by radiologists.[71] The resulting classified volume was a three-value data set, so the Union operation did follow the traditional binary union rules. In our case, "Lesion" taken priority, followed by "Uncertainty," and then "Background." A voxel was classified as a Lesion if at least one of the two classifications identified it as such. Elsewhere, if at least one of the two classifications identified it as an Uncertainty, it was classified that way. If neither of these applied, it was set as Background. After combining the above sets, we noticed an increase in false positives compared to each individual classifier. To address this, we used a majority vote among the other four classifications (two coronal and two sagittal). For each voxel, a class was maintained if at least two other classifiers confirmed it. Otherwise, it was downgraded by one (a potential Lesion became Uncertainty, and a potential Uncertainty became Background). First, a decision is made on the Lesion, and then on the Uncertainty.

The use of multiple classifiers is justified because it allowed us to combine both common and specific information from axial classifiers. This helps to ensure that any potentially positive voxel is confirmed by the coronal and sagittal classifiers, maintaining 3D continuity. We decided to follow the typical procedure used by radiologists and take advantage of the benefits of using multiple classifiers.

In the proposed automatic pipeline, we initially focused on using axial sections over other orientations, but we also conducted trials to test the preference of other orientations in the fusion process. The results, not reported, confirm that the axial preference yields the best results, followed closely by the coronal preference, with the sagittal preference being the last, despite it being the direction used for FLAIR data collection. This could be partly explained by the fact that axial and coronal slices display highly symmetrical shapes in both brain anatomy and lesions, making the learning process easier compared to sagittal slices, which lack symmetry and can result in significant variations also for minor head rotations.

4. Performance Parameters

In binary classification, voxels are categorized as positive (P or Lesion) or negative (N or Background). In a ternary classification, P represents the voxels of the current class, while N represents the negative voxels (those of the other two classes). Each rater follows the same rules for the current class: True Positive (TP) are the correctly identified positive voxels, True Negative (TN) are the correctly identified negative voxels, False Positive (FP) are the incorrectly identified positive voxels, and False Negative (FN) are the incorrectly identified negative voxels. In our specific case, we only checked Lesion and the other two classes are fused and considered as Background. To thoroughly compare all raters (artificial, human, and ground truth) and to account for the lack of a single performance parameter, we calculated various well-known scores and metrics. Despite some redundancy, we prioritized over-describing the results to expose any potential errors and prevent them from benefiting the model. We defined and calculated scores, typically ranging from 0 to 1 with 1 indicating the ideal value, and metrics, typically ranging from 0 to infinity with 0 indicating the best value. Detailed definitions can be found elsewhere.[5,6,72] Table 2 provides the list of scores and metrics used for our comparisons, along with their respective formulas.

5. Results and Discussion

The proposed framework has undergone training, validation, and testing using the ternary ground truth provided above. By keeping the original binary ground truth unchanged, we ensured a direct

Table 2. Scores (left) and metrics (right) used for the comparisons.

Scores (optimum is 1)	Metrics (optimum is 0)
$\text{SENS} = \frac{\text{TP}}{\text{TP+FN}}$	$\text{EF} = \frac{\text{FP}}{\text{TP+FN}}$
$\text{OSENS} = \frac{\text{TP}_o}{\text{TP}_o+\text{FN}_o}$	$\text{DER} = \frac{\text{DE}}{\text{MTA}}$
$\text{SPEC} = \frac{\text{TN}}{\text{TN+FP}}$	$\text{OER} = \frac{\text{OE}}{\text{MTA}}$
$\text{ACC} = \left(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}\right)/2$	$\text{FDE} = \frac{\text{FP}}{P}$
$\text{PPV} = \frac{\text{TP}}{\text{TP+FP}}$	$\text{RAE} = \frac{\text{TP+FP}-P}{P}$
$\text{OPPV} = \frac{\text{TP}_o}{\text{TP}_o+\text{FP}_o}$	$\text{HD}(A,B) = \max(h(A,B), h(B,A))$
$\text{Dice} = \frac{2*\text{TP}}{2*\text{TP+FP+FN}}$	$\text{ED}(A,B) = \max(d(A,B), d(B,A))$
$\text{IoU} = \frac{\text{TP}}{\text{TP+FP+FN}}$	$\text{SD} = \frac{\sum_{i\in A_S} d(x_i,G_S)+\sum_{j\in G_S} d(x_j,A_S)}{N_A+N_G}$
$F1 = 2 * \frac{\text{OSENS}*\text{OPPV}}{\text{OSENS+OPPV}}$	
$\text{PCC}(A,B) = \frac{\text{cov}(A,B)}{\sigma_A * \sigma_B}$	

Notes: Left: sensitivity (SENS); object-wise SENS (OSENS), where the subscript o indicates the whole object; specificity (SPEC); accuracy (ACC); positive predictive value (PPV); object-wise PPV (OPPV); Dice score (Dice); intersection over union (IoU); $F1$; Pearson correlated coefficient (PCC), ranging in $[-1, 1]$ and calculated between two data sets A and B of which it uses the covariance between them and their standard deviation. Right: extra fraction (EF); detection error rate (DER), using the detection error (DE), calculated by summing the voxels of connected regions incorrectly labeled as positives, and the mean total area (MTA), where MTA is the average of the number of positive voxels from both the rater and the ground truth; the outline error rate (OER), where OE is the outline error calculated as the difference between the number of voxels of the union and that of the intersection between the positively connected regions; false detection error (FDE); relative area error (RAE); Hausdorff distance (HD) among two sets A and B, using $h(A,B) = \max_{a\in A} \min_{b\in B}\|a-b\|$; pseudo-Euclidean distance (ED), using $d(A,B) = \frac{1}{N}\sum_{a\in A} \min_{b\in B}\|a-b\|$; surface distance (SD), using two segmentations, the rater segmentation (A_S) and the ground truth segmentation (G_S), and their corresponding number of points, N_A and N_G, respectively.

comparison with both human raters and state-of-the-art automated methods.

We evaluated the proposed framework and the human radiologists by applying cross-validation, as defined in Sec. 3.1. We calculated the average and standard deviation for the indicators defined in Sec. 4, divided into scores and metrics. The initial results reported in Fig. 5 involve a comparison between the raters and the proposed framework with the ground truth concerning Lesion. This comparison also enables an indirect assessment of the proposed framework against the human raters, using the ground truth as a reference.

To provide a clearer overview, the average values are displayed in Fig. 6 using a radar visualization. These values confirm that the proposed framework aligns with the variability among different raters. This can be explained by the fact that our framework has been trained using a consensus approach, which tends to average the assessments of different raters involved in its development. A statistical test (Wilcoxon signed-rank test) of the metric values shows, with a significance level of 0.01, that there is no significant difference between the performance of our method and that of the seven human raters. In other words, if data are presented without labels, distinguishing the behavior of our framework from that of humans would be nearly impossible.

The positive results mentioned above are not enough on their own for us to conclude that our framework behaves like human raters. This is because the comparison was made against the ground truth. Put differently, our framework could be at the same distance from the ground truth as the human raters, but from opposite sides. A direct comparison was necessary. To do this, we experimented by

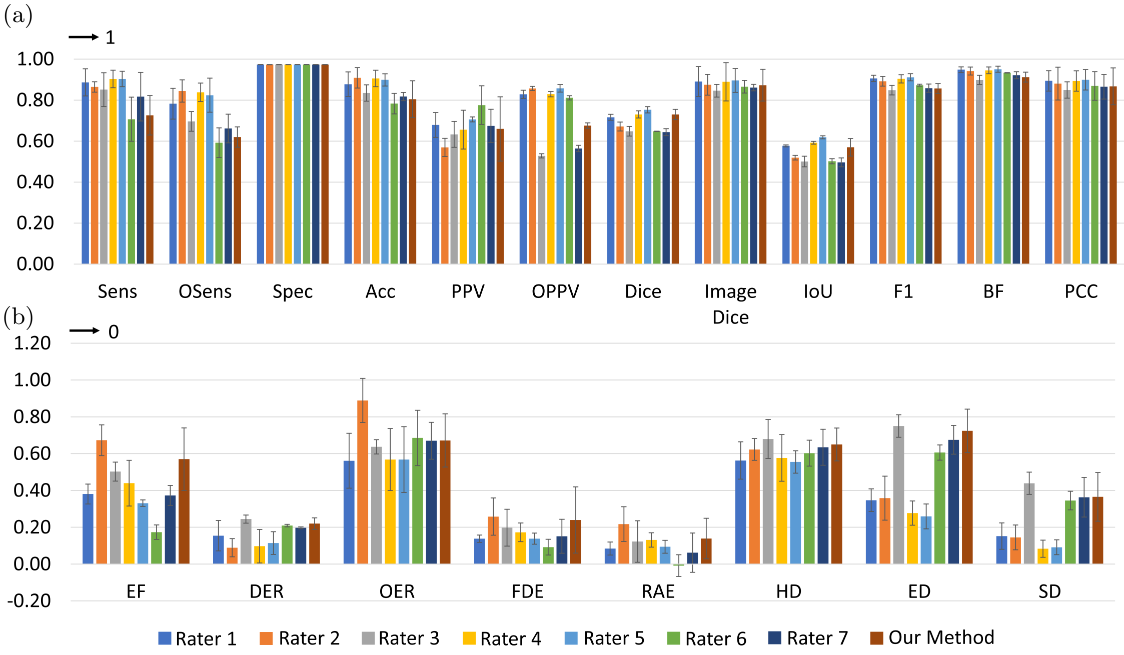


Fig. 5. The raters and the proposed framework are compared to the ground truth for the class Lesion both in scores (a) and in metrics (b). The average and the standard deviation define the interval of confidence for the reported results.

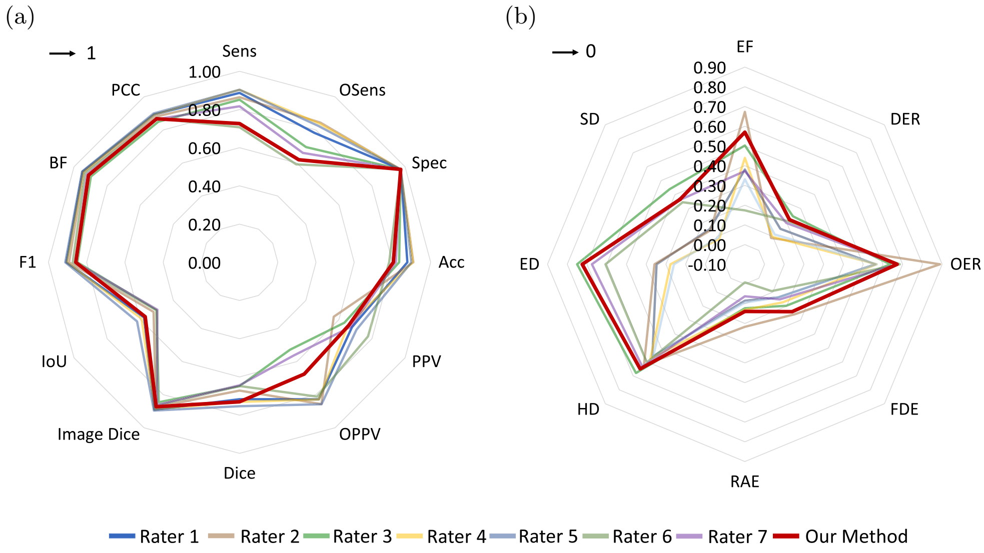


Fig. 6. (Color online) Radar plots of the average values of the indicators depicted in Fig. 5. Only the average values are presented for clarity, where the line of the proposed framework (in red) is highlighted for emphasis.

comparing all raters to each other, in rotation. The results for the most commonly used indicators are presented in Fig. 7, confirming that the proposed framework's behavior is similar to that of other human raters and that it is not biased toward a specific rater or the ground truth. Additionally, as noted by other authors,[11] the results reveal similarities between some human raters (R4 with R5 and R6 with R7). Thankfully, the results also confirm that the ground truth is not influenced by these similarities among raters and that it maintains a "human" behavior, closely resembling raters R1 and

R2. This is crucial as it indicates that all people striving to train automated systems to the ground truth, including ourselves, are not chasing an unattainable goal, where the closer we get, the farther we move from the true objective. A visual representation of the behavior of the proposed framework throughout the process of identification/segmentation, for both Lesion and Uncertainty, is depicted in Fig. 8 as an example of a typical examination (the more frequent situation). For both Lesion and Uncertainty, the proposed framework tends to produce FP. Interestingly, FN is almost absent from the segmented volume. Regarding FP, a question arises: if, as shown earlier, the proposed framework is just as good as human raters, then why is it affected by FP? The answer is simple: both the human raters and the consensus are also influenced by FP. If this weren't the case, the proposed framework would have shown different performance, mainly affecting those indicators relying on FP, compared to human raters. This is also confirmed by analyzing the unique no-lesion subject of the test dataset in MSSEG. The proposed framework detected three lesions with a lesion load of $0.096\,\mathrm{cm}^3$. However, according to Commowick *et al.,*[6] human raters assigned the number of lesions for this case in the range of 0–8, with just two raters indicating 0 lesions. The reported lesion load ranged from 0 to $10.88\,\mathrm{cm}^3$. In this no-lesion subject, as in the others, the proposed framework falls in the middle. It's important to note that none of the state-of-the-art models listed in Table 3 considered this special case.

Finally, we have summarized the comparison between the proposed framework and recently introduced automated strategies in Table 3. The necessary condition for a method to be considered is that it must have been trained, validated, and tested on the MSSEG dataset. This ensures that the comparison is consistent and conducted under the same conditions as those for the seven human raters. The considered indicators are those calculated in at least one strategy, apart our framework, and their values are collected from the referenced papers.

Despite the difficulty of creating a global ranking, the data in Table 3 clearly show that the proposed framework is the most stable across indicators and generally performs better than other methods, even those using multiple imaging modalities. In fact, while the proposed framework excels in only two of the reported indicators, its stability ensures that biases are not introduced, as could happen with other models that optimize a limited number of potentially favorable indicators. For instance, the model developed by Ghosal *et al.* demonstrates a high accuracy value of 0.97. In contrast, the accuracy levels of the seven human experts in Fig. 5 range from 0.78 to 0.91. A similar situation is observed in the model developed by Alijamaat *et al.*, which shows a high Dice score level of 0.82, compared to that of the human experts, which ranges from 0.64 to 0.75. A particularly strong performance by one indicator could generate a "distinctive" sign (a bias) in the model. However, this doesn't occur in the proposed framework, as it maintains a good balance with input from the seven human experts. This is supported by the huge number of calculated indicators, which prevents the introduction of new biases beyond those originally present in the problem, by which even human experts are influenced, as any other supervised automatic strategies trained on their labeled data.

The use of a single imaging modality in the proposed framework could have interesting implications:

- FLAIR not only contains the necessary information but also provides sufficient details to identify and segment all MS lesions in the WM.
- The performance could be enhanced more than using multiple modalities due to the extensive variability of MRI.
- Acquisition time, patient stress, and time for diagnosis could be reduced.
- Valuable information for radiologists could be obtained.

An important factor contributing to the valuable performance of the proposed framework is the utilization of the ternary ground truth. Figure 9 displays the results when the proposed framework is trained on the binary consensus (without Uncertainty) compared to those obtained when trained on the ternary consensus (with Uncertainty). The ensemble method trained without Uncertainty outperforms similar automated strategies (TF[6]); however, it still falls far behind human methodologies. What put the proposed framework on par with human performance was the inclusion of Uncertainty.

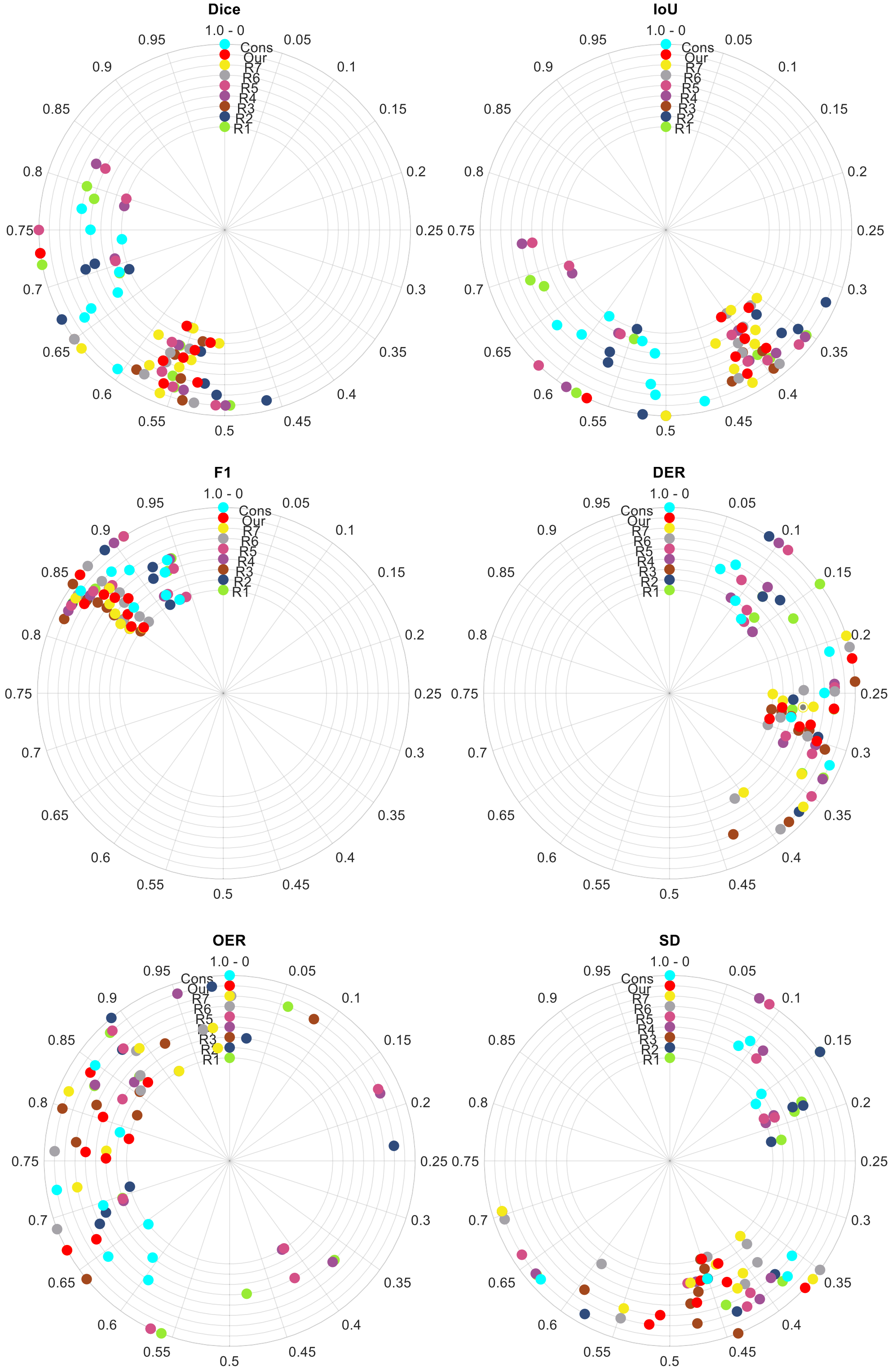
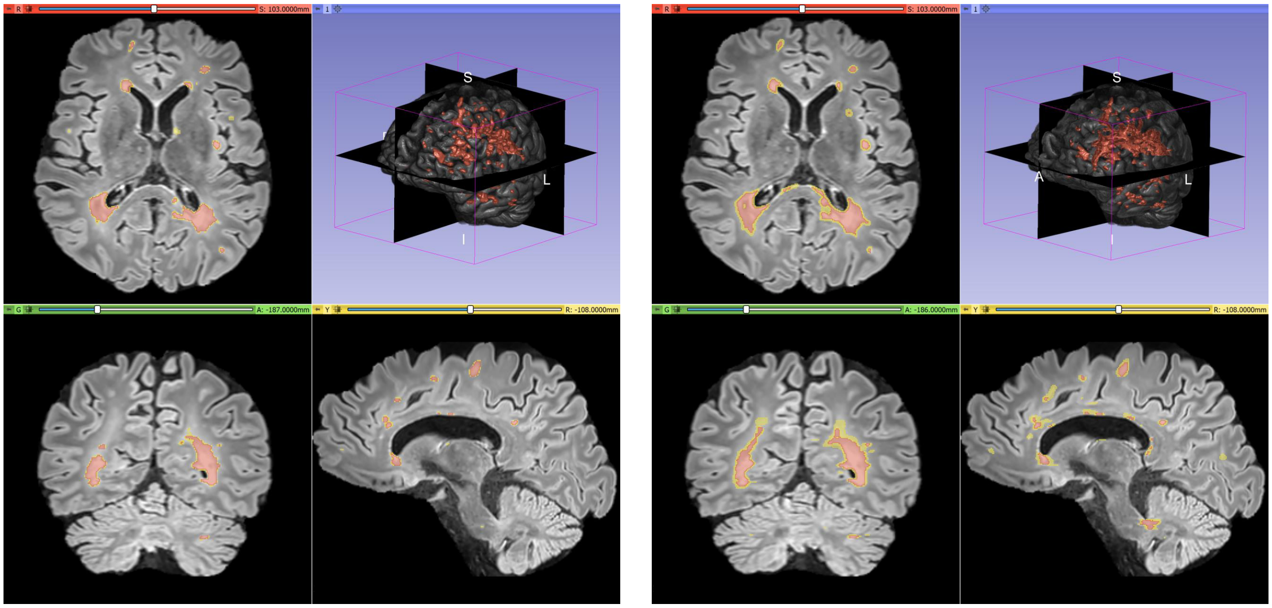


Fig. 7. All raters are compared to each other, including our framework and the consensus, with each one alternatively considered as the ground truth. To make it easier to read, we only reported on some metrics. The angular position indicates the metric value: clockwise for scores converging to 1, and counterclockwise for metrics converging to 0. Different raters are indicated radially.



(a)

(b)

Fig. 8. (Color online) Comparison between the ground truth (left) and the proposed automated framework (right). Lesions are marked in red and Uncertainty in yellow. The upper right panel on each side shows the 3D localization of the reported slices for readability.

Table 3. Comparison of our framework with the state-of-the-art methods, whose indicators are collected from the original papers. Unavailable data are represented by "—". Team Fusion (TF)[6] is reported for reference.

Method	MRI mod.	Sens	OSens	Acc	PPV	OPPV	Dice	F1	SD
TF[6]	FLAIR, PD, T2, T1, G-E T1	0.71	0,60	—	0.65	0.53	0.64	0.50	0.91
Ref. 9	FLAIR, PD, T2, T1, G-E T1	0.65	—	**0.97**	—	—	0.76	—	—
Ref. 73	FLAIR, T1	0.55	—	—	—	**0.79**	0.63	—	—
Ref. 10	FLAIR, PD, T1, G-E T1	**0.76**	—	—	—	—	**0.82**	—	—
Ref. 12	FLAIR, T1, T2	—	—	—	—	—	0.76	0.59	—
Ref. 74	FLAIR	—	—	—	—	—	0.74	—	—
Ref. 75	FLAIR, T2, T1	0.74	—	—	0.65	—	0.67	0.59	—
Ref. 56	T2-FLAIR, T1	0.68	0.62	—	0.64	—	0.72	—	—
Ref. 57	FLAIR, T1	0.68	**0.83**	—	**0.68**	0.55	0.65	—	—
OUR	FLAIR	0.73	0.62	0.80	0.66	0.68	0.73	**0.86**	**0.37**

This is consistent with the findings in Ref. 11: the framework improves its understanding of what is definitely a Lesion and what is definitely Background, and uses Uncertainty as a buffer class for ambiguous cases. The polarized and ambiguous classification of uncertain voxels, sometimes as Lesion and other times as Background, confuses any automated strategy and leads it away from human-like reasoning. In the proposed study, it would have been important to evaluate the nature of the Uncertainty class segmented by the proposed model. However, we could not do it for different reasons: (1) the Uncertainty class, which we defined from the differences from the seven Raters with the ground truth, was not originally present in the MSSEG dataset (the raters were not originally asked to segment, besides Lesion and Background, the Uncertainty class — in a ternary way — and this could have been led to a completely different output from the Raters. (2) Our definition of the Uncertainty

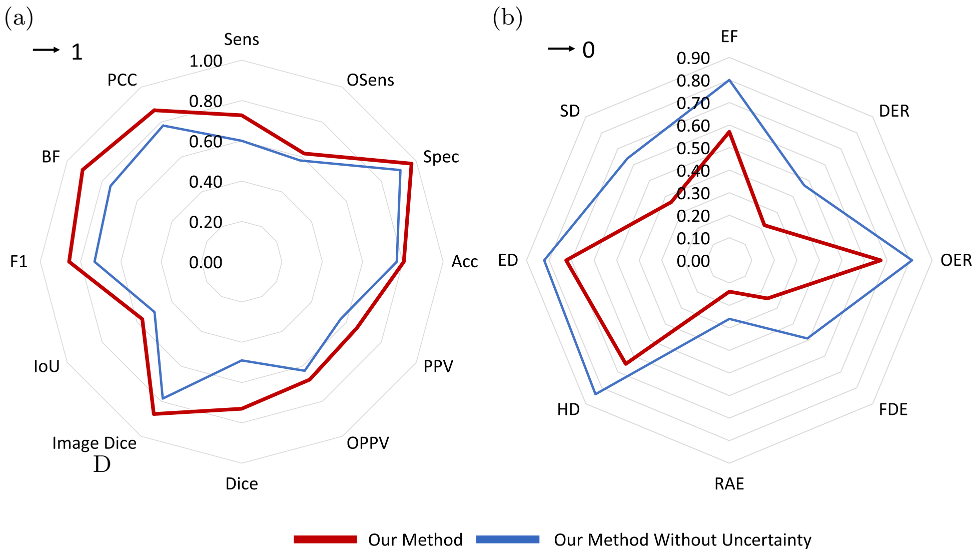


Fig. 9. (Color online) The proposed framework indicators when trained with (red) and without (blue) Uncertainty, for scores converging to 1 (a) and for metrics converging to 0 (b).

class was arbitrary: we used it just to demonstrate that if an AI model could adopt an intermediate class between Background and Lesions, the segmentation of the Lesion class was more effective. (3) We lacked a ground truth for the Uncertainty class, so any comparison would have been arbitrary.

6. Conclusions

An automated framework has been presented for identifying and segmenting MS lesions in WM from FLAIR MRI images. The framework utilizes CNN-based architectures that are adapted to behave like human specialists. The strengths of the proposed framework include training the system to recognize lesions in their environment, emulating the procedures of human radiologists, using ensemble classification, operating solely on FLAIR images, and incorporating an artificially generated Uncertainty class to improve performance. This Uncertainty class represents the system's confidence level in its predictions, which is a key factor in improving the system's performance.

Results have shown that the framework closely resembles human raters in both behavior and performance. It outperforms state-of-the-art strategies and has exhibited behavior equivalent to human raters. The use of Uncertainty during training has significantly improved the framework's performance.

A recent report by the JASON Advisory Group[76] emphasizes the importance of new technologies addressing significant clinical needs and reducing medical costs. Demonstrating improved performance by incorporating key concepts such as Uncertainty and Ensemble aligns with these recommendations.

Future directions of exploration include defining "Uncertainty" and studying its role in decision-making, implementing specific pre-processing strategies for FLAIR images to improve method robustness concerning MRI variability,[77] addressing unbalancing problems via different loss functions, exploring consensus and loss functions based on probability values, testing the framework for identifying cortical lesions and in longitudinal studies,[20] implementing pre-processing strategies for MRI harmonization,[78] and studying the role of additional meta-information in improving lesion identification.

Acknowledgments

This research was partially supported by The European Union — NextGenerationEU under the Italian Ministry of University and Research (MUR) National Innovation Ecosystem grant ECS00000041 - VITALITY — CUP E13C22001060006. The FVG Project "Supporting the diagnosis of rare diseases (MR) through artificial intelligence" (2023–2026)

(CUP:F53C22001770002). Project PNC 0000001 D3-4 Health - CUP B83C22006120001 through National Plan for Complementary Investments to the NRRP, Funded by the European Union - NextGenerationEU.

ORCID

Giuseppe Placidi https://orcid.org/0000-0002-4790-4029
Luigi Cinque https://orcid.org/0000-0001-9149-2175
Gian Luca Foresti https://orcid.org/0000-0002-8425-6892
Francesca Galassi https://orcid.org/0000-0002-8788-2856
Filippo Mignosi https://orcid.org/0000-0001-9599-5730
Michele Nappi https://orcid.org/0000-0002-2517-2867
Matteo Polsinelli https://orcid.org/0000-0002-4215-2630

References

1. M. L. Steinman, Multiple sclerosis: A coordinated immunological attack against myelin in the central nervous system, *Cell* **85** (1996) 299–302.
2. M. Filippi *et al.*, Assessment of lesions on magnetic resonance imaging in multiple sclerosis: Practical guidelines, *Brain* **142** (2019) 1858–1875.
3. G. Placidi, *MRI: Essentials for Innovative Technologies*, 1st edn. (CRC Press, 2012).
4. A. Carré *et al.*, Standardization of brain MR images across machines and protocols: Bridging the gap for MRI-based radiomics, *Sci. Rep.* **10** (2020) 12340.
5. A. Danelakis, T. Theoharis and D. A. Verganelakis, Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging, *Comput. Med. Imaging Graph.* **70** (2018) 83–100.
6. O. Commowick *et al.*, Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure, *Sci. Rep.* **8** (2018) 13650.
7. G. Placidi, L. Cinque, M. Polsinelli, A. Splendiani and E. Tommasino, Automatic framework for multiple sclerosis follow-up by magnetic resonance imaging for reducing contrast agents, in *Image Analysis and Processing – ICIAP 2019*, eds. E. Ricci, S. Rota Bulò, C. Snoek, O. Lanz, S. Messelodi and N. Sebe (Springer International Publishing, Cham, 2019), pp. 367–378.
8. H. Zhang and I. Oguz, Multiple sclerosis lesion segmentation — A survey of supervised CNN-based methods, in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds. A. Crimi and S. Bakas (Springer International Publishing, Cham, 2021), pp. 11–29.
9. P. Ghosal, P. K. C. Prasad and D. Nandi, A light weighted deep learning framework for multiple sclerosis lesion segmentation, in *2019 Fifth Int. Conf. Image Information Processing (ICIIP)*, Solan, Himachal Pradesh, India (IEEE, 2019), pp. 526–531.
10. A. Alijamaat, A. NikravanShalmani and P. Bayat, Multiple sclerosis lesion segmentation from brain MRI using U-Net based on wavelet pooling, *Int. J. Comput. Assist. Radiol. Surg.* **16**(9) (2021) 1459–1467.
11. C. Gros, A. Lemay and J. Cohen-Adad, SoftSeg: Advantages of soft versus binary training for image segmentation, *Med. Image Anal.* **71** (2021) 102038.
12. R. McKinley *et al.*, Simultaneous lesion and brain segmentation in multiple sclerosis using deep neural networks, *Sci. Rep.* **11**(1) (2021) 1–11.
13. A. Traboulsee *et al.*, Revised recommendations of the consortium of MS centers task force for a standardized MRI protocol and clinical guidelines for the diagnosis and follow-up of multiple sclerosis, *Am. J. Neuroradiol.* **37**(3) (2016) 394–401.
14. U. Macar, E. N. Karthik, C. Gros, A. Lemay and J. Cohen-Adad, Team NeuroPoly: Description of the pipelines for the MICCAI 2021 MS new lesions segmentation challenge, arXiv:abs/2109.05409.
15. V. Sundaresan, G. Zamboni, P. M. Rothwell, M. Jenkinson and L. Griffanti, Triplanar ensemble U-Net model for white matter hyperintensities segmentation on MR images, *Med. Image Anal.* **73** (2021) 102184.
16. Y. Yang, Y. Hu, X. Zhang and S. Wang, Two-stage selective ensemble of CNN via deep tree training for medical image classification, *IIEEE Trans. Cybern.* **52**(9) (2021) 9194–9207.
17. R. Alizadehsani *et al.*, Handling of uncertainty in medical data using machine learning and probability theory techniques: a review of 30 years (1991–2020), *Ann. Oper. Res.* **339** (2021) 1077–1118.
18. O. Vincent, C. Gros and J. Cohen-Adad, Impact of individual rater style on deep learning uncertainty in medical imaging segmentation, arXiv:abs/2105.02197 (2021).
19. C. Olivier, C. Frédéric and A. Roxana, MSSEG challenge proceedings: Multiple sclerosis lesions segmentation challenge using a data management and processing infrastructure, in *MICCAI*, 2016, Athénes, Greece.
20. O. Commowick, F. Cervenansky, F. Cotton and M. Dojat, MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure, in *MICCAI 2021-24th Int. Conf. Medical Image Computing and Computer Assisted Intervention*, 2021, Strasbourg, France, p. 126.
21. D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold and D. L. Collins, Review of automatic

segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging, *Med. Image Anal.* **17** (2013) 1–18.

22. G. Mirzaei and H. Adeli, Segmentation and clustering in brain MRI imaging, *Rev. Neurosci.* **30**(1) (2018) 31–44.
23. A.-M. Tăuţan, B. Ionescu and E. Santarnecchi, Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques, *Artif. Intell. Med.* **117** (2021) 1–22.
24. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken and C. I. Sánchez, A survey on deep learning in medical image analysis, *Med. Image Anal.* **42** (2017) 60–88.
25. A. Kaur, L. Kaur and A. Singh, State-of-the-art segmentation techniques and future directions for multiple sclerosis brain lesions, *Arch. Comput. Methods Eng.* **28** (2021) 951–977.
26. H. M. R. Afzal, S. Luo, S. Ramadan and J. Lechner-Scott, The emerging role of artificial intelligence in multiple sclerosis imaging, *Mult. Scler. Int.* **28**(6) (2022) 849–858.
27. M. Zurita, C. Montalba, T. Labbé, J. P. Cruz, J. D. da Rocha, C. Tejos, E. Ciampi, C. Cárcamo, R. Sitaram and S. Uribe, Characterization of relapsing-remitting multiple sclerosis patients using support vector machine classifications of functional and diffusion MRI data, *NeuroImage: Clin.* **20** (2018) 724–730.
28. N. Shiee, P.-L. Bazin, A. Ozturk, D. S. Reich, P. A. Calabresi and D. L. Pham, A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions, *NeuroImage* **49** (2010) 1524–1535.
29. H. Wu, X. Chen, P. Li and Z. Wen, Automatic symmetry detection from brain MRI based on a 2-channel convolutional neural network, *IEEE Trans. Cybern.* **51**(9) (2021) 4464–4475.
30. E. Geremia, O. Clatz, B. H. Menze, E. Konukoglu, A. Criminisi and N. Ayache, Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images, *NeuroImage* **57** (2011) 378–390.
31. Z. Karimaghaloo, M. Shah, S. J. Francis, D. L. Arnold, D. L. Collins and T. Arbel, Automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain MRI using conditional random fields, *IEEE Trans. Med. Imaging* **31**(6) (2012) 1181–1194.
32. M. Cabezas, A. Oliver, S. Valverde, B. Beltran, J. Freixenet, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira and X. Lladó, BOOST: A supervised approach for multiple sclerosis lesion segmentation, *J. Neurosci. Methods* **237** (2014) 108–117.
33. N. Guizard, P. Coupé, V. S. Fonov, J. V. Manjón, D. L. Arnold and D. L. Collins, Rotation-invariant multi-contrast non-local means for MS lesion segmentation, *NeuroImage: Clin.* **8** (2015) 376–389.
34. A. Carass *et al.*, Longitudinal multiple sclerosis lesion segmentation: Resource and challenge, *NeuroImage* **148** (2017) 77–102.
35. Y. Yoo, T. Brosch, A. Traboulsee, D. K. Li and R. Tam, Deep learning of image features from unlabeled data for multiple sclerosis lesion segmentation, in *Machine Learning in Medical Imaging: 5th Int. Workshop, MLMI 2014, Held in Conjunction with MICCAI 2014*, Boston, MA, USA, September 14, 2014. Proc. 5 (Springer, Cham, 2014), pp. 117–124.
36. L. S. Aït-Ali, S. Prima, P. Hellier, B. Carsin, G. Edan and C. Barillot, STREM: A robust multidimensional parametric method to segment MS lesions in MRI, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005*, eds. J. S. Duncan and G. Gerig (Springer, Berlin, Heidelberg, 2005), pp. 409–416.
37. E. Roura, A. Oliver, M. Cabezas, S. Valverde, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira and X. Lladó, A toolbox for multiple sclerosis lesion segmentation, *Neuroradiology* **57** (2015) 1031–1043.
38. M. Strumia, F. R. Schmidt, C. Anastasopoulos, C. Granziera, G. Krueger and T. Brox, White matter MS-lesion segmentation using a geometric brain model, *IEEE Trans. Med. Imaging* **35**(7) (2016) 1636–1646.
39. T. Brosch, L. Y. W. Tang, Y. Yoo, D. K. B. Li, A. Traboulsee and R. Tam, Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation, *IEEE Trans. Med. Imaging* **35**(5) (2016) 1229–1239.
40. S. Aslani, M. Dayan, L. Storelli, M. Filippi, V. Murino, M. Rocca and D. Sona, Multi-branch convolutional neural network for multiple sclerosis lesion segmentation, *NeuroImage* **196** (2019) 1–15.
41. D. Nie, L. Wang, E. Adeli, C. Lao, W. Lin and D. Shen, 3-D fully convolutional networks for multimodal isointense infant brain image segmentation, *IEEE Trans. Cybern.* **49**(3) (2019) 1123–1136.
42. J. Xue, K. He, D. Nie, E. Adeli, Z. Shi, S.-W. Lee, Y. Zheng, X. Liu, D. Li and D. Shen, Cascaded multitask 3-D fully convolutional networks for pancreas segmentation, *IEEE Trans. Cybern.* **51**(4) (2021) 2153–2165.
43. O. Lucena, R. Souza, L. Rittner, R. Frayne and R. Lotufo, Convolutional neural networks for skull-stripping in brain MR imaging using silver standard masks, *Artif. Intell. Med.* **98** (2019) 48–58.
44. J. Hu, C. Yu, Z. Yi and H. Zhang, Enhancing robustness of medical image segmentation model with neural memory ordinary differential equation, *Int. J. Neural Syst.* **33**(12) (2023) 2350060.
45. H. S. Nogay and H. Adeli, Multiple classification of brain MRI autism spectrum disorder by age and gender using deep learning, *J. Med. Syst.* **48**(1) (2024) 15.

46. J. Á. Díaz-Francés, J. D. Fernández-Rodríguez, K. Thurnhofer-Hemsi and E. López-Rubio, Semi-supervised semantic image segmentation by deep diffusion models and generative adversarial networks, *Int. J. Neural Syst.* **34**(11) (2024) 2450057.
47. C. Zhou, L. Ye, H. Peng, Z. Liu, J. Wang and A. Ramírez-De-Arellano, A parallel convolutional network based on spiking neural systems, *Int. J. Neural Syst.* **34**(5) (2024) 2450022.
48. J. Bernal, K. Kushibar, D. S. Asfaw, S. Valverde, A. Oliver, R. Martí and X. Lladó, Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: A review, *Artif. Intell. Med.* **95** (2019) 64–81.
49. G. Castellano, G. Placidi, M. Polsinelli, G. Tulipani and G. Vessio, Unsupervised brain MRI anomaly detection for multiple sclerosis classification, in *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, eds. J.-J. Rousseau and B. Kapralos (Springer Nature Switzerland, Cham, 2023), pp. 644–652.
50. H. S. Nogay and H. Adeli, Diagnostic of autism spectrum disorder based on structural brain MRI images using, grid search optimization, and convolutional neural networks, *Biomed. Signal Process. Control* **79** (2023) 104234.
51. J. Dong, G. Zhang, Y. Hu, Y. Wu and H. Rong, An optimization numerical spiking neural membrane system with adaptive multi-mutation operators for brain tumor segmentation, *Int. J. Neural Syst.* **34**(8) (2024) 2450036.
52. F. Mercaldo, M. Di Giammarco, F. Ravelli, F. Martinelli, A. Santone and M. Cesarelli, Alzheimer's disease evaluation through visual explainability by means of convolutional neural networks, *Int. J. Neural Syst.* **34**(2) (2024) 2450007.
53. O. Ronneberger, P. Fischer and T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, eds. N. Navab, J. Hornegger, W. M. Wells and A. F. Frangi (Springer International Publishing, Cham, 2015), pp. 234–241.
54. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox and O. Ronneberger, 3D U-Net: Learning dense volumetric segmentation from sparse annotation, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (Springer International Publishing, 2016), pp. 424–432.
55. F. Milletari, N. Navab and S.-A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in *2016 Fourth Int. Conf. 3D Vision (3DV)*, 2016, California, USA, pp. 565–571.
56. L. Bai *et al.*, Improving multiple sclerosis lesion segmentation across clinical sites: A federated learning approach with noise-resilient training, *Artif. Intell. Med.* **152** (2024) 102872.
57. T. Wiltgen *et al.*, LST-AI: A deep learning ensemble for accurate MS lesion segmentation, *NeuroImage: Clin.* **42** (2024) 103611.
58. G. Kang, B. Hou, Y. Ma, F. Labeau, Z. Su *et al.*, Acu-Net: A 3D attention context U-Net for multiple sclerosis lesion segmentation, in *ICASSP 2020-2020 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain (IEEE, 2020), pp. 1384–1388.
59. Y. S. Vang, Y. Cao, P. D. Chang, D. S. Chow, A. U. Brandt, F. Paul, M. Scheel and X. Xie, SynergyNet: A fusion framework for multiple sclerosis brain MRI segmentation with local refinement, in *2020 IEEE 17th Int. Symp. Biomedical Imaging (ISBI)*, Iowa City, IA, USA (IEEE, 2020), pp. 131–135.
60. H. Zhang, J. Zhang, R. Wang, Q. Zhang, S. A. Gauthier, P. Spincemaille, T. D. Nguyen and Y. Wang, Geometric loss for deep multiple sclerosis lesion segmentation, in *2021 IEEE 18th Int. Symp. Biomedical Imaging (ISBI)*, Nice, France (IEEE, 2021), pp. 24–28.
61. A. Fenneteau, P. Bourdon, D. Helbert, C. Fernandez-Maloigne, C. N. Habas and R. Guillevin, Learning a CNN on multiple sclerosis lesion segmentation with self-supervision, in *Int. Symp. Electronic Imaging 2020 3D Measurement and Data Processing*, 2020, San Francisco, USA, pp. 1–8.
62. R. ElSebely, A. H. Yousef, A. A. Salem and B. Abdullah, Automatic segmentation of multiple sclerosis lesions in brain MR images using ensemble machine learning, in *2021 Int. Mobile, Intelligent, and Ubiquitous Computing Conf. (MIUCC)*, Cairo, Egypt (IEEE, 2021), pp. 28–33.
63. Ž. Špiclin, F. Pernuš, B. Likar, T. Jerman and D. Ravnik, Dataset variability leverages white-matter lesion segmentation performance with convolutional neural network, in *Medical Imaging 2018: Image Processing*, eds. E. D. Angelini and B. A. Landman, Vol. 10574 (SPIE, 2018), pp. 388–396.
64. R. Walsh, C. Meurée, A. Kerbrat, A. Masson, B. R. Hussein, M. Gaubert, F. Galassi and B. Combès, Expert variability and deep learning performance in spinal cord lesion segmentation for multiple sclerosis patients, in *IEEE CBMS 2023-36th IEEE Int. Symp. Computer-Based Medical Systems (CBMS)*, 2023, L'Aquila, Italy, pp. 463–470.
65. E. Kats, J. Goldberger and H. Greenspan, Soft labeling by distilling anatomical knowledge for improved MS lesion segmentation, in *2019 IEEE 16th Int. Symp. Biomedical Imaging (ISBI 2019)*, 2019, Venice, Italy, pp. 1563–1566.
66. S. Santurkar, D. Tsipras, A. Ilyas and A. Madry, How does batch normalization help optimization? in *Proc. 32nd Int. Conf. Neural Information Processing*

Systems, NIPS'18 (Curran Associates, Red Hook, NY, USA, 2018), p. 2488–2498.

67. J. Snoek, H. Larochelle and R. P. Adams, Practical Bayesian optimization of machine learning algorithms, in *Proc. 25th Int. Conf, Neural Information Processing Systems — Volume 2, NIPS'12*, (Curran Associates, Red Hook, NY, USA, 2012), p. 2951–2959.
68. X. Zhang, X. Chen, L. Yao, C. Ge and M. Dong, Deep neural network hyperparameter optimization with orthogonal array tuning, in *Neural Information Processing*, eds. T. Gedeon, K. W. Wong and M. Lee (Springer International Publishing, Cham, 2019), pp. 287–295.
69. L. Gu and X.-C. Cai, Fusing 2D and 3D convolutional neural networks for the segmentation of aorta and coronary arteries from CT images, *Artif. Intell. Med.* **121** (2021) 102189.
70. M. H. Abd Latif and I. Faye, Automated tibiofemoral joint segmentation based on deeply supervised 2D-3D ensemble U-Net: Data from the osteoarthritis initiative, *Artif. Intell. Med.* **122** (2021) 102213.
71. H. Geijer and M. Geijer, Added value of double reading in diagnostic radiology, a systematic review, *Insights Imaging* **9** (2018) 287–301.
72. G. Csurka, D. Larlus, F. Perronnin and F. Meylan, What is a good evaluation measure for semantic segmentation? in *British Machine Vision Conf.*, Vol. 27, 2013, Bristol, UK, pp. 10–5244.
73. S. Valverde, M. Salem, M. Cabezas, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, J. Salvi, A. Oliver and X. Lladó, One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks, *NeuroImage: Clin.* **21** (2019) 101638.
74. D. Liu *et al.*, Multiple sclerosis lesion segmentation: revisiting weighting mechanisms for federated learning, *Front. Neurosci.* **17** (2023) 1167612.
75. B. Sarica, D. Z. Seker and B. Bayram, A dense residual U-Net for multiple sclerosis lesions segmentation from multi-sequence 3D MR images, *Int. J. Med. Inform.* **170** (2023) 104965.
76. D. Derrington, Artificial intelligence for health and health care (2017), https://www.healthit.gov/sites/default/files/jsr-17-task-002_aiforhealthandhealthcare12122017.pdf.
77. G. Placidi, L. Cinque, M. Nappi, M. Polsinelli, A. Sciarra and G. Tortora, Star-Net: A multi-branch convolutional network for multiple source image segmentation, in *2022 16th Int. Conf. Signal-Image Technology & Internet-Based Systems (SITIS)*, Dijon, France (IEEE, 2022), pp. 127–134.
78. M. Polsinelli, H. B. Li, F. Mignosi, L. Zhang and G. Placidi, Siamese network to assess scanner-related contrast variability in MRI, *Image Vis. Comput.* **145** (2024) 104997.