scientific data

Check for updates

OPEN

DATA DESCRIPTOR

MSLesSeg: baseline and benchmarking of a new Multiple Sclerosis Lesion Segmentation dataset

Francesco Guarnera[1]✉, Alessia Rondinella[1]✉, Elena Crispino[2], Giulia Russo[3], Clara Di Lorenzo[4], Davide Maimone[5], Francesco Pappalardo[3] & Sebastiano Battiato[1]

This paper presents MSLesSeg, a new, publicly accessible MRI dataset designed to advance research in Multiple Sclerosis (MS) lesion segmentation. The dataset comprises 115 scans of 75 patients including T1, T2 and FLAIR sequences, along with supplementary clinical data collected across different sources. Expert-validated annotations provide high-quality lesion segmentation labels, establishing a reliable human-labeled dataset for benchmarking. Part of the dataset was shared with expert scientists with the aim to compare the last automatic AI-based image segmentation solutions with an expert-biased handmade segmentation. In addition, an AI-based lesion segmentation of MSLesSeg was developed and technically validated against the last state-of-the-art methods. The dataset, the detailed analysis of researcher contributions, and the baseline results presented here mark a significant milestone for advancing automated MS lesion segmentation research.

Background & Summary

Multiple Sclerosis (MS) is a chronic inflammatory disorder that leads to demyelination of the central nervous system. Main neuropathological characteristics of the disease include focal areas of inflammation, accompanied by the loss of both myelin and axons. Magnetic Resonance Imaging (MRI) is widely employed to detect MS lesions in various brain regions and the spinal cord, with lesions progressively accumulating over time[1]. Discriminative circumscription of lesions, including periventricular, cortical/juxtacortical, brainstem/cerebellar, and spinal cord areas, is crucial for MS diagnosis. Monitoring lesion development, such as identifying new or expanding lesions during follow-up, also plays an essential role in assessing therapeutic outcomes and tracking disease progression[2–5]. Handmade annotation of MS lesions on MRI images is labor-intensive and demands significant expertise. Furthermore, bias due to the operator is inevitable, which for sure characterize lesion segmentation modeling[6,7]. Consequently, there is a growing demand for automated MRI analysis solutions to reduce the influence of human error and make these assessments more readily accessible in routine clinical settings[8–10].

In order to identify detailed visualizations of the lesions, axial brain MRI protocols use different imaging sequences. Each sequence provides different contrasts between brain tissues, and the multimodal analysis through the use of all of them has been preferred by scientists in recent years[11,12]. Commonly used sequences to identify MS lesions include Fluid Attenuated Inversion Recovery (FLAIR), T1-weighted (T1-w) and T2-weighted (T2-w) sequences (Fig. 1). T1-w and T2-w capture the response of hydrogen nuclei after being disturbed, with T1-w representing the relaxation time that relates to how quickly the nuclei of hydrogen atoms return to their equilibrium state[13] and T2-w indicating the time for hydrogen nuclei to lose phase coherence among neighboring spins after a radiofrequency disturbance. In T1-w images, white matter appears brighter than gray matter, while CSF appears dark; Conversely, in T2-w images, white matter is darker than gray matter, and CSF appears bright. FLAIR sequence is designed to suppress CSF signals, enhancing the visibility of brain

[1]Department of Mathematics and Computer Science, University of Catania, Catania, Italy. [2]Department of Biomedical and Biotechnological Sciences, University of Catania, Catania, Italy. [3]Department of Drug and Health Sciences, University of Catania, Catania, Italy. [4]UOC Radiologia, ARNAS Garibaldi, Catania, Italy. [5]Centro Sclerosi Multipla, UOC Neurologia, Azienda Ospedaliera per l'Emergenza Cannizzaro, Catania, Italy. ✉e-mail: francesco.guarnera@unict.it; alessia.rondinella@unict.it

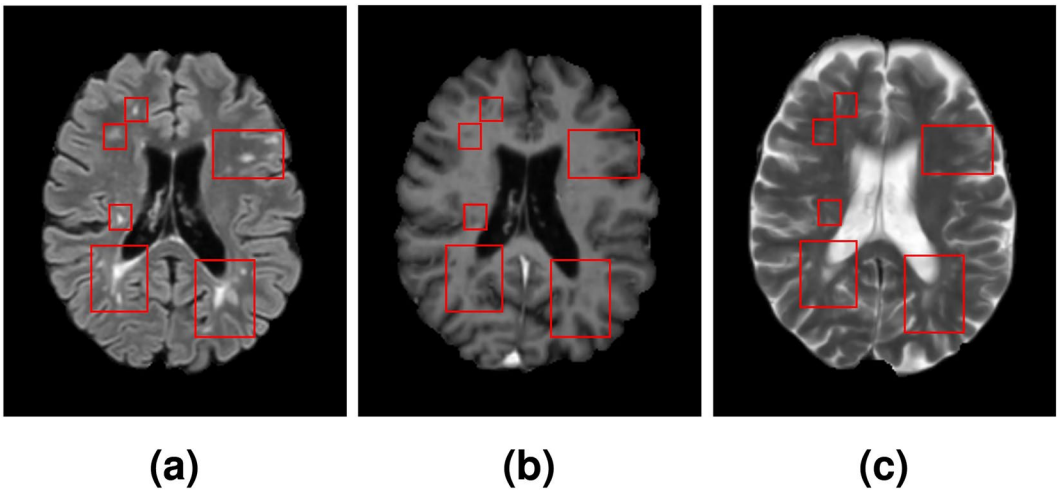


Fig. 1 MRI images MS disease patient in the 3 described sequences FLAIR (**a**), T1-weighted (**b**), and T2-weighted (**c**). Red boxes highlight lesions in the different modality of acquisition.

and spinal cord lesions. This process utilizes an inversion recovery pulse to eliminate the CSF signal, followed by a delay before acquiring a T2-w image; FLAIR images are particularly effective at highlighting MS lesions, where they appear hyperintense and are generally well-differentiated from adjacent tissue. Usually FLAIR sequence is the most important to detect lesions, while T1-w and T2-w are employed to complete the analysis started from the FLAIR.

Recent progress in Machine Learning (ML) and Artificial Intelligence (AI) has resulted in the development of advanced algorithms for lesion segmentation in MS, primarily relying on data-driven approaches that depend on well-curated, high-quality training datasets. The creation of segmentation algorithms for MRI, especially for MS lesion detection, has highlighted the necessity for large public datasets containing MRI images and corresponding segmentation masks. However, such large-scale, high-quality datasets for MS lesion segmentation on MRI are still lacking. This study aims to advance MS lesion segmentation research by pursuing the following objectives:

- Introduce MSLesSeg, a new large-scale, publicly available MRI dataset labeled for MS lesion segmentation;
- Provide a detailed analysis of segmentation methods on the field, distinguishing between human expert and automated segmentations;
- Establish a baseline method for MS lesion segmentation on MSLesSeg, validated through comparisons with state-of-the-art techniques.

A broader goal of this work is to create a benchmark dataset for the field, enabling researchers to test algorithms in realistic scenarios and reducing reliance on context-specific solutions.

Methods

This section describes the methodologies employed in constructing the MSLesSeg dataset and the subsequent benchmarking procedures. First, the *Data Collection* sub-section outlines the patients cohort, acquisition timepoints, division into training and test sets, and the accompanying clinical data. *Acquisition* then details the imaging protocols and MRI scanner settings to ensure consistency. In *Preprocessing*, we describe the steps taken to standardize and prepare the images for analysis, including co-registration to achieve cross-sequence compatibility. The *Labelling* sub-section explains the lesion annotation procedure conducted by expert radiologists and the validation process to create reliable segmentation labels. For performance assessment, *Evaluation Metrics* are defined to quantify segmentation accuracy and reliability, both for the baseline model and comparative analyses. Finally, *Consensus Analysis* investigates the consensus delineation, comparing the results of challenge participant segmentations and human labels.

Data collection. MSLesSeg includes data of 75 patients, aged 18 to 59 years, with a mean age of 37 years ($\pm$ 10.3) at baseline. The cohort includes 48 female and 27 male participants. MRI scans were retrospectively collected at varying timepoints per patient, ranging from 1 to 4 scans. Specifically, 50 patients had 1 timepoint, 15 had 2 timepoints, 5 had 3 timepoints, and 5 had 4 timepoints, with a total number of 115 series. For each timepoint, three imaging modalities were acquired: T1-w, T2-w, and FLAIR. The dataset is divided into a training set and a test set according with the setting of the challenge described in[14], where only the training set was privately shared with the participants. The training set includes 53 patients, with 1 to 4 timepoints, including 50 with Relapsing Remitting Multiple Sclerosis (RRMS) and 3 with Secondary Progressive Multiple Sclerosis (SPMS), while the test set includes 22 patients, each with only 1 timepoint, comprising 21 with RRMS and 1 with Primary Progressive Multiple Sclerosis (PPMS). Table 1 describes patient details for the training and test datasets with information regarding MS type, age, EDSS, Lesion Volume and number of lesions.

Set	# Patients	# M/F	MS Type	Age	EDSS	Lesion Volume	# Lesions
				Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Training	53	21/32	50 RRMS	38,46 (±10, 59)	1,9 (±1, 73)	12.167 (±13.233)	28,9 (±19, 2)
			3 SPMS				
Testisng	22	6/16	21 RRMS	36,39 (±10, 16)	1,79 (±1, 70)	10.231 (±10.912)	41,2 (±43, 17)
			1 SPMS				

Table 1. Patient details for the training and test dataset. The columns include the number of patients (# Patients), the number of male/female patients (# M/F), the breakdown of patients by type of multiple sclerosis (MS Type), mean and standard deviation (SD) of age, Expanded Disability Status Scale (EDSS), lesion volume and number of lesions.

Modality	Rows x Columsns	Pixel Spacing	Slice Thickness	Repetition Time (RT)	Echo Time	Inversion Time (IT)	Flip Angle
FLAIR	455,82 × 455,82	0,75 × 0,75	2,61	6859,95	276,31	2072,62	90,55
	(±314,71 × 314,71)	(±0,32 × 0,32)	(±0, 62)	(±1750, 69)	(±89, 49)	(±340, 85)	(±6, 23)
T1-w	288,65 × 288,65	0,91 × 0,91	1,14	102,98	5,96	NA	18,18
	(±74,46 × 74,46)	(±0,15 × 0,15)	(±0, 81)	(±434, 93)	(±21, 04)		(±20, 86)
T2-w	473,76 × 473,76	0,58 × 0,58	4,24	5898,63	102,94	NA	91,05
	(±165,94 × 165,94)	(±0,27 × 0,27)	(±0, 92)	(±1976, 34)	(±10, 58)		(±12, 87)

Table 2. Acquisition details for each sequence used in the MS patient training and testing set.

Acquisition. The acquisitions were obtained as part of a study conducted by an Italian hospital. Many of these were performed individually in private clinics and later added to the study by the patients themselves. Consequently, the data were acquired from different scanners with field strength of 3 and 1.5 Tesla which make the dataset proper for challenging research. All the acquisitions are in DICOM (Digital Imaging and Communications in Medicine) format, but did not follow a unique protocol, for this reason the properties of RAW scans are not equally for all items. Table 2 shows the acquisition details for each MRI modality. Specifically, FLAIR have a mean of 455,82 × 455,82 (±314,71 × 314,71) spatial resolution, a mean of 0,75 × 0,75 (±0,32 × 0,32) pixel spacing, a mean of 2,61 (±0, 62) Slice Thickness, a mean of 6859,95 (±1750, 69) Repetition Time (RT), a mean of 276,31 (±89, 49) Echo Time (ET), a mean of 2072,62 (±340, 85) Inversion Time (IT), and a mean of 90,55 (±6, 23) Flip angle, T1-w have a mean of 288,65 × 288,65 (±74,46 × 74,46) spatial resolution, a mean of 0,91 × 0,91 (±0,32 × 0,32) pixel spacing, a mean of 1,14 (±0, 81) Slice Thickness, a mean of 102,98 (±434, 93) Repetition Time (TR), a mean of 5,96 (±21, 04) Echo Time (TE), and a mean of 18,18 (±20, 86) Flip angle, while T2-w have a mean of 473,76 × 473,76 (±165,94 × 165,94) spatial resolution, a mean of 0,58 × 0,58 (±0,27 × 0,27) pixel spacing, a mean of 4,24 (±0, 92) Slice Thickness, a mean of 5898,63 (±1976, 34) Repetition Time (TR), a mean of 102,94 (±10, 58) Echo Time (TE), and a mean of 91,05 (±12, 87) Flip angle. Individual patient data have been collected by Garibaldi Hospital (Catania) upon authorization of ethic committee COMITATO ETICO CATANIA 2 (prot. 810/C.E. del 20/12/2020, favorable opinion) and are irreversibly anonymised. All patients gave their informed consent to the publication of the anonymized data. The hospital authorised the release of the data under a Creative Commons Attribution 4.0 International license (CC-BY-4.0).

Preprocessing. All MRI scans in the dataset underwent comprehensive preprocessing to ensure consistency and standardization across the imaging data. Initially, each scan was fully anonymized to safeguard patient confidentiality and adhere to ethical guidelines on data privacy. Following anonymization, the scans were converted from the original DICOM format to the NIFTI (Neuroimaging Informatics Technology Initiative) format. The NIFTI format was chosen due to its widespread use and compatibility in neuroimaging research, offering significant advantages in terms of storage efficiency and ease of manipulation for further analysis. These preprocessing steps were critical to ensure the dataset's uniformity and its suitability for subsequent analyses. After, each MRI modality underwent co-registration to the MNI152 isotropic template, which has a spatial resolution of $1mm^3$. This process was conducted using FMRIB's Linear Image Registration Tool (FLIRT), a fully automated and widely recognized tool for precise brain image registration. FLIRT[15] was employed to align the scans from all modalities to a standardized reference space, specifically the Montreal Neurological Institute (MNI) 152 template, which serves as a common anatomical reference in neuroimaging research. Specifically, we performed an affine registration with 12 degrees of freedom. By using this technique, spatial discrepancies between different scans were minimized, ensuring that all images were consistently aligned to the same anatomical framework, which is essential for accurate comparison and further analysis across subjects and timepoints. This co-registration process enhances the dataset's compatibility for subsequent computational analysis and cross-modality integration. Following co-registration, brain extraction was performed using the Brain Extraction Tool (BET)[16], a well-established method in neuroimaging for isolating brain tissue. BET efficiently removes non-brain structures, including the skull, scalp and other surrounding tissues, leaving only the brain itself for further analysis. This step is critical in eliminating extraneous anatomical features that could interfere with downstream processing and analysis. By standardising the MRI data through brain extraction, the pre-processing pipeline ensures that all

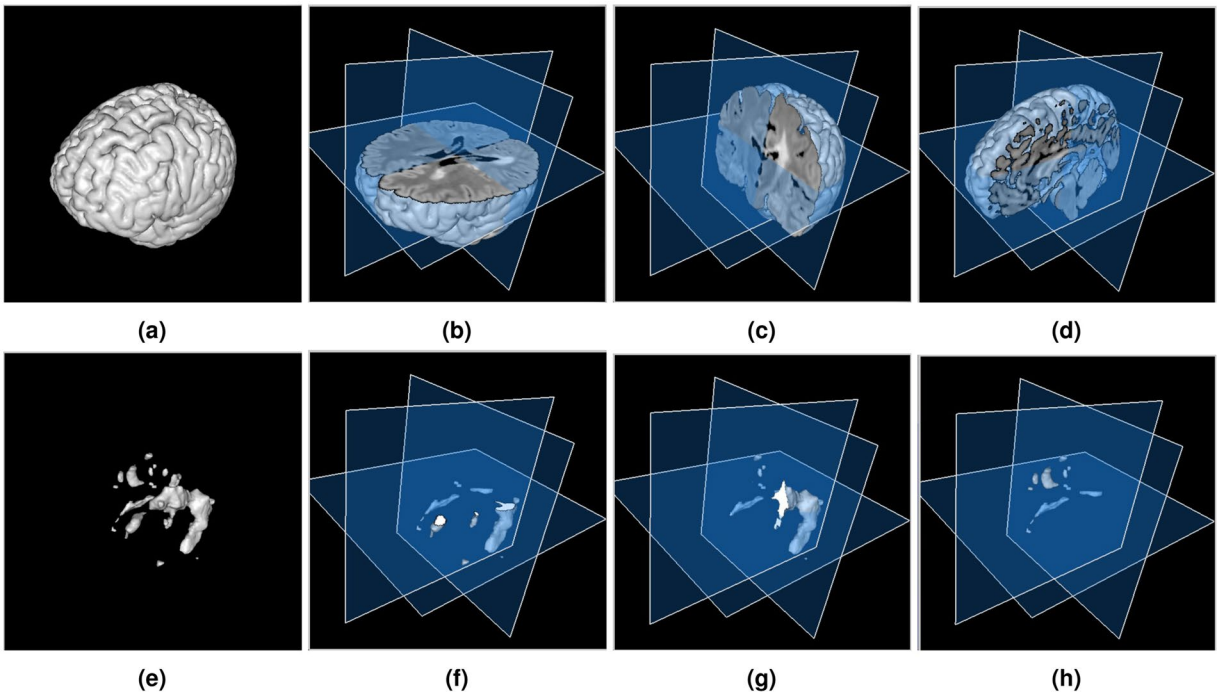


Fig. 2 3D images of acquisition (**a**) and full annotation mask (**e**). The other three columns have the same meaning after a cut carried out in the central slice of each one of the 3 plane and for both acquisition and mask to make visible brain shape and lesion segmentation.

scans are consistent and optimised for subsequent tasks, such as the development and evaluation of multiple sclerosis (MS) lesion segmentation algorithms. This meticulous pre-processing improves the quality and reliability of the dataset, enabling more accurate and reproducible computational analysis in MS research. Anonymization, conversion, registration and brain-extraction were done with proper python scripts; in particular the last two employed FSL library, a collection of analysis tools for FMRI, MRI and diffusion brain imaging data.

Labelling. The dataset underwent meticulous manual annotation to generate ground-truth masks of hyperintense lesions on FLAIR images for each timepoint across all patients. Manual segmentation of MS lesions on MRI presents significant challenges due to the inherent complexity of the task, coupled with the natural variability in interpretations among different expert annotators. The primary modality for lesion identification was FLAIR, but to increase accuracy and confidence, cross-checking was conducted using T2-weighted (T2-w) and T1-weighted (T1-w) sequences.

The segmentation process was carried out by one junior rater and two senior experts, a highly experienced neuroradiologist and a senior neurologist, both specializing in MS. Multiple training sessions were held between the junior and senior experts to develop a standardized and consistent lesion segmentation approach. These sessions also involved the introduction and training of the junior rater on the use of JIM 9.0 (Xinapse Systems Ltd., UK), a sophisticated software tool known for its advanced image registration, segmentation, and analysis capabilities. Following the training, the junior rater began annotating the MRI data. Throughout this process, frequent meetings were held between the junior rater and the senior experts to review and validate the annotations. This iterative approach, with regular feedback, helped maintain accuracy and consistency across all segmentations. The lesions were segmented on registered FLAIR images, while T2-w and T1-w scans were used to resolve any ambiguities or challenging cases encountered on the FLAIR images. Once the junior rater completed the segmentation, the annotations were thoroughly reviewed and validated by the two senior experts. After this final validation process, the annotated lesion masks were accepted as the ground-truth for the dataset. An example of annotation mask generated as described is visible in Fig. 2: it describes acquisition and mask in first column, and the relative central sections w.r.t. the 3 planes in the other columns.

To note that our group chose to implement this labelling protocol voluntarily: many state-of-the-art datasets proposed multiple annotations of the same element as ISBI-2015[17] in the MS field. On the one hand, this choice avoids bias, but on the other hand, it does not impose a specific way of using the labels. As a consequence, the labels are used in different ways, some methods use only one or a few evaluator labels, others combine evaluators with a consensus mask: often this possibility is an advantage, and does not allow comparison. In our case, all analyses performed on MsLesSeg will use the same labels, generated from more experts and then unbiased.

Labelling protocol. The junior rater, given a scan, follow a specific protocol during the labelling phase as follow:

1. Open a visualization tool for MRI (JIM9 or others) and load T2-w and T1-w sequences on the tool;
2. Open Jim9 and import the FLAIR sequence;
3. Adjust the display settings for optimal visualization (contrast, brightness, view, zoom);
4. Enable Region of Interest (ROIs) Identification and Lesion Marking;
5. Navigating through Slices;
6. Identify a Lesion through boundaries;
7. Create ROI (Region of Interest) around the lesion;
8. Avoid Non-Lesion Areas (for challenging scenarios the rater use the comparison in T1-w and T2-w previously opened);
9. Propagate ROIs Across Slices (propagation allows to apply a defined ROI across multiple slices in a 3D MRI scan, useful when the ROI shape does not change drastically across neighboring slices):
    a. Select ROI to Propagate;
    b. Enable Propagation;
    c. Adjust Propagation Settings (combination mode like Union, Intersection, Exclusive OR and Pixel Area Threshold to define the intensity or size threshold for the inclusion of pixels in the ROI);
    d. Manual Adjustments (review the propagated ROIs in the subsequent slices making manual adjustments).
10. Create the Mask (after multiple ROIs creation):
    a. Open the Image Masking Tool;
    b. Load the Input FLAIR scan (load the scan on which the ROIs were created; this will serve as the base image where the mask will be applied);
    c. Set Pixel Intensity for the Mask;
    d. Save mask.

Evaluation Metrics. Segmentation performance was evaluated through a set of standard metrics that capture different aspects of the result's quality. Using multiple metrics is crucial to providing a comprehensive performance assessment, as each metric measures different aspects of the segmentation, such as region overlap, the balance between false positives and false negatives, and the similarity between the surfaces of the segmented lesions.

- Dice Similarity Coefficient (DSC): The DSC is a measure of overlap between the predicted and reference sets, accounting for both sensitivity and precision. It is defined as:

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

  where X is the set of segmented lesions and Y is the set of reference lesions. Higher DSC values indicate greater similarity between the two segmentations.
- True Positive Rate (TPR): Also known as sensitivity or recall, TPR measures the proportion of correctly identified lesions relative to the total number of actual lesions. It is defined as:

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

  where TP represents true positives and FN false negatives.
- Positive Predictive Value (PPV): PPV, or precision, measures the proportion of correctly segmented lesions relative to all lesions identified by the model. The formula is:

$$PPV = \frac{TP}{TP + FP} \tag{3}$$

  where FP represents false positives.
- Lesion-wise True Positive Rate (LTPR): measures the proportion of correctly identified lesions at the individual level. This metric is particularly useful for assessing the model's ability to segment small lesions correctly.
- Lesion-wise False Positive Rate (LFPR): LFPR measures the proportion of incorrectly segmented lesions relative to the total number of lesions detected by the model, indicating the model's tendency to over-segment.
- Absolute Volume Difference (AVD): AVD measures the volume difference between the segmented and actual lesions. A lower AVD indicates that the total volume of the segmented lesions is closer to the actual volume. The formula is:

$$AVD = \frac{|V_{pred} - V_{true}|}{V_{true}} \tag{4}$$

  where V_{pred} and V_{true} are the volumes of the segmented and true lesions, respectively.

# Team	GT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Challenge result	1.00	0.71	0.71	0.71	0.70	0.69	0.68	0.68	0.65	0.65	0.64	0.64	0.61	0.57	0.54	0.50
STAPLE*C1*	0.69	0.83	0.77	0.75	0.75	0.69	0.75	0.89	0.68	0.78	0.78	0.79	0.60	0.64	0.73	0.62
STAPLE*C2*	0.75	0.95	0.90	0.90	0.88	0.83	-	-	-	-	-	-	-	-	-	

Table 3. Consensus results in terms of Dice Score employing all 15 participants masks (*C1*) and only the best 5 (*C2*).

- Average Symmetric Surface Distance (ASSD): ASSD measures the average distance between the surfaces of the segmented and true lesions, providing an evaluation of the segmentation's boundary accuracy. It is calculated as:

$$ASSD = \frac{1}{|S_X| + |S_Y|}\left(\sum_{x \in S_X} \min_{y \in S_Y} d(x, y) + \sum_{y \in S_Y} \min_{x \in S_X} d(y, x)\right) \quad (5)$$

where S_x and S_Y are the surfaces of the segmented and reference lesions, and $d(x, y)$ is the Euclidean distance between points x and y.

These metrics provide a comprehensive view of the model's performance, from volumetric similarity to surface segmentation quality, enabling a thorough comparison between different segmentation methods.

Consensus analysis. Recently, our research group organized a challenge[14] that involved sharing part of MSLesSeg exclusively with registered teams. The goal was to promote the development of AI-based algorithms for the automatic segmentation of MS lesions and to compare these results against human annotations. Participants were provided with a portion of the dataset (the training set) that included human-labeled ground-truth, while the remaining data (the test set) was shared without ground-truth labels. Teams were challenged to generate segmentation masks for the test set using their proposed models. Following submission, an in-depth consensus delineation analysis was conducted to evaluate and compare the segmentations predicted by each participating team. Although the MSLesSeg segmentation masks were created by a single rater, the dataset underwent a validation process involving three experts with varying levels of experience. This approach reduces potential bias, as it ensures that the dataset is not solely influenced by one rater's perspective. To further investigate the distinctions between AI-generated masks and human annotations, a comprehensive consensus analysis was performed. This analysis enabled a direct comparison of the quality of our manually labeled ground-truths with the segmentation results generated by the teams' automatic algorithms, providing insight into the relative accuracy and consistency of human and AI-based annotations. In the first experiment consensus masks were generated employing all participants masks and the ground-truth masks with the Simultaneous Truth And Performance Level Estimation (STAPLE) method[18]: STAPLE is an advanced algorithm based on expectation-maximization, used to statistically merge multiple binary segmentations into a cohesive result. Its function is to analyze multiple input masks and generate a probabilistic estimate of what the true segmentation might be, along with additional relevant data. For each scan of the test set, the created consensus mask was employed as a ground-truth reference to calculate the performance of the same masks used to create the consensus ones.

The Dice scores of this first experiment is shown in Table 3 (STAPLE *C1* row) where the better results of all teams with the new consensus mask seem to highlight a common error. To confirm that, a visual analysis on consensus mask was carried out, to understand what is happening: some representative elements are shown in Fig. 3. The presence of isolated lesion (Fig. 3 (d)) minimises the likelihood of error (fewer false positives and negatives, as shown in Figures 3(e) and 3(f)), whereas in presence of lesions located in uncommon areas or of larger dimensions (Fig. 3(a)), AI algorithms tend to outline a wider area, often producing the same error (a kind of lesion contour as shown in Fig. 3(b) and 3(c)). To further confirm the aforementioned insight the same consensus mask was generated employing only the best 5 results of the challenge: the results described in Table 3 (STAPLE *C2* row) and the images on last column of Fig. 3 highlight how the errors are similar in shape and that a flattening of the lesions contours on the consensus mask corresponds to an improvement in the results (better results of GT w.r.t. STAPLE consensus mask). A more detailed analysis of the single participant methods shows how the described "lesion contour" error is more accentuated when the method uses only the FLAIR sequence to predict the mask.

Data Records

MSLesSeg is public available on https://doi.org/10.6084/m9.figshare.27919209[19]. In order to permit the comparison with the proposed baseline, the dataset was split into training set and test set, but users can merge them for deeper analysis. The dataset is accompanied by an overview file, that explains the structure of the directories w.r.t. patients, timepoints, type of sequences and segmentation mask; moreover it furnish additional information about the patients like age, sex, acquisition date and other details related to the scan (these data are referring to the patients and timepoints through a unique ID). All FLAIR, T2-weighted, T1-weighted and segmentation masks are in NIFTI format with a standardized size 182 × 218 × 182 in LAS orientation (standard "Radiological" convention) due to the registration step performed with MNI152 template.

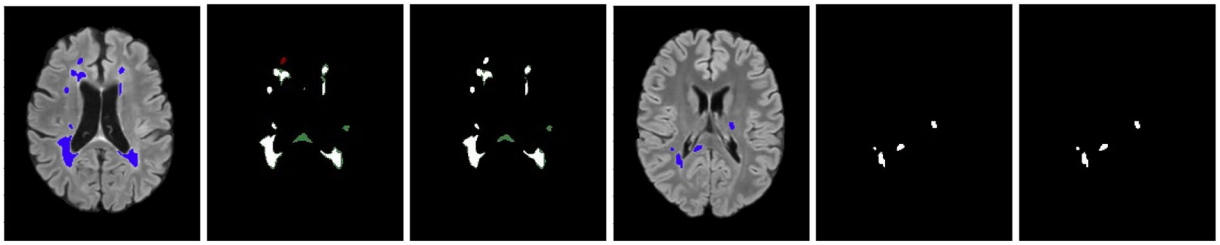


Fig. 3 Difference between human and AI-based segmentation masks on two different types of lesion; (**a**) and (**d**) are the input with the corresponding segmentation masks produced by the raters highlighted in blue, (**b**) and (**e**) are the mean masks extracted from the 15 AI methods (test *C1*) while (**c**) and (**f**) are the mean mask extracted from the best 5 AI methods (test *C2*). In columns (**b**),(**c**),(**e**) and (**f**), the green pixels represent the false positives of the mean mask w.r.t. (**a**) and (**d**), and the red pixels the false negatives; this means that the green pixels are part of the mean mask.

N° Fold	1	2	3	4	5
Training patients	1-43	1-33, 44-53	1-23, 34-53	1-13, 23-53	11-53
Validation patients	44-53	34-43	24-33	14-23	1-10

Table 4. Cross-folding configuration: each fold presents different couple of training/validation sets.

Mean on 5 Fold							
Model	**DSC↑**	**TPR↑**	**PPV↑**	**LTPR↑**	**LFPR↓**	**AVD↓**	**ASSD↓**
MSSegDiff[20]	**0.6851**	0.6719	**0.7241**	0.6250	**0.2226**	**24.45**	**3.474**
UNETR[26]	0.6421	**0.6751**	0.6496	**0.7332**	0.4331	90.16	4.524
SwinUNETR[25]	0.6786	0.6676	0.7157	0.7017	0.2957	25.12	3.622
TransBTS[27]	0.4917	0.5447	0.5027	0.5110	0.5408	67.9509	7.5484

Table 5. Obtained results with the proposed baseline and other comparative models. All the results are averaged by the 5 folds described in Table 4.

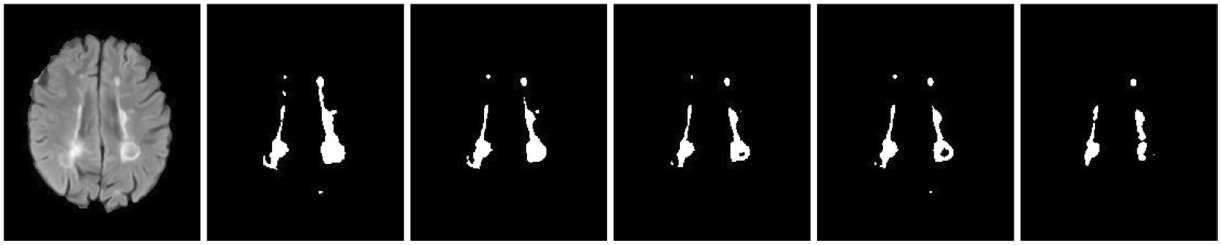
Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
MSSegDiff[20]	**0.6836**	0.6801	**0.6811**	**0.7067**	**0.6742**
UNETR[26]	0.6541	0.6387	0.6374	0.6460	0.6347
SwinUNETR[25]	0.6819	**0.6838**	0.6701	0.6885	0.6690
TransBTS[27]	0.5045	0.4668	0.5231	0.5524	0.4118

Table 6. Dice Score of the proposed baseline and other comparative models over the folds.

Technical Validation

The goal of this section is to demonstrate how MSLesSeg, equally to other technical validate dataset in the field, is applicable for segmentation tasks. The results obtained with the proposed baseline and other state-of-the-art methods demonstrate the dataset's congruence w.r.t. ISBI-2015[17] once, which is the reference dataset for the task. Many methods in the field suffer from a generalisation property, often due to the low cardinality of the data: in order to propose a generalised method and to highlight its goodness, a cross-validation strategy was employed. In particular, starting from the training/test split described in previous sections, 5 different folds were created for training, as described in Table 4. The numbers in columns *Training patients* and *Validation patients* correspond to the range patients (i.e. 1-43 are the scans of all patients from number 1 to number 43); this choice permits to avoid that a patient employed for training will be evaluated in the next phase (methods with same brain-shape in training and validation/test steps perform naturally better). All the tests were done employing the patients from number 54 to 77 and the numbers of the following Tables (5) (6) are referred to them.

With the mentioned configuration, the idea was to propose a baseline and to compare it with some significant state-of-the-art models. The deep learning framework chosen as the baseline has been presented by Rondinella *et al.*[20]: this diffusion-based architecture, based on BasicUNet[21], integrates attention mechanisms to improve MS lesion detection[22]. Noise is added in t steps to ground-truth masks, and the input volume is passed through an Encoder Module to extract key features. These features, combined with noisy labels, flow through the downsampling path of the Denoising Attention U-Net, which then reverses the process to produce a clean segmentation mask. Squeeze-and-Attention blocks[23], customized for 3D images, are added after each module, enabling the



(a) FLAIR **(b)** GT **(c)** MSSegDiff **(d)** SwinUNETR **(e)** UNETR **(f)** TransBTS

Fig. 4 Example of segmented cases showing (**a**) the FLAIR and its relative Ground-truth (**b**) and predictions from MSSegDiff, SwinUNETR, UNETR, and TransBTS models. The results highlight that MSSegDiff produces fewer errors compared to SwinUNETR, UNETR and TransBTS, demonstrating its superior performance in lesion segmentation.

network to focus on relevant pixel groups. Both encoders maintain consistent features and dimensions for seamless merging, with Squeeze-and-Attention blocks placed after each up and downsampling block. During evaluation, segmentation masks are generated at each step using the Denoising Diffusion Implicit Model[24]. These masks are merged using the Step-Uncertainty-based Fusion (SUF) module[22], ensuring more stable results as prediction accuracy improves with more testing steps. This architecture demonstrated its goodness[20] with the reference dataset ISBI-2015[17], for this reason the same implementation has been evaluated employing MSLesSeg: a similar results would show the quality and validity of the dataset proposed in this paper. For technical validation of dataset and baseline, a comparison was made with two of the best architectures in the field, SwinUNETR[25], UNETR[26], and TransBTS[27]. UNETR and SwinUNETR architectures are similar in some parts, but both are actually the state-of-the-art in medical imaging segmentation. UNETR[26] is a transformer-based encoder that learns sequence representations from the input volume. The encoder is connected to a CNN-based decoder through skip connections to produce the final segmentation output. SwinUNETR[25] is a neural network that merges the advantages of the Swin Transformer and UNETR models. The Swin Transformer is employed as the encoder, utilizing hierarchical feature extraction with a shifted windows approach. The architecture is highly effective at capturing both local and global context, making it well-suited for complex segmentation tasks. The encoder's output is linked to a CNN-based decoder at various resolutions via skip connections. TransBTS[27], on the other hand, is designed to balance local and global feature extraction in 3D medical image segmentation tasks. The encoder in TransBTS first applies 3D CNNs to capture local spatial features within volumetric data. These features are then transformed into tokens fed into a Transformer module, which performs global feature modeling through self-attention mechanisms. The decoder progressively upsamples the embedded features to produce a detailed segmentation map.

The comparison was done through a cross-validation strategy using 5 different folds for training/validation steps (Table 4) and other 24 patients for testing (patients from number 54 to number 77 of MSLesSeg). The results described by all the metrics employed for the lesions segmentation task shown as MSSegDiff generally achieve better results, with the mean of the Dice Score over the folds being higher compared to other methods (Table 5); also the results on single folds (Table 6) show better performances of MSSegDiff in almost all cases, with a constant improvement w.r.t. other methods (minimum with SwinUNETR). The other metrics described in Table 5 confirm this analysis, with MSSegDiff and SwinUNETR showing the same behaviour (small difference in all the metrics) while UNETR outperforms in terms of TPR and LTRP, which means a probable larger segmentation in the conflict zone. Fig. 4 further illustrates this by providing an example of the predictions generated by each network, clearly highlighting the superior accuracy of MSSegDiff as compared to the other models. For the task of MS lesions segmentation AI solutions play a pivotal role in enabling automated and accurate analysis of brain MRI scans, assisting neurologists in their clinical practice. Deep Learning and AI algorithms have the potential to produce robust, quantitative results for understanding the progression and impact of multiple sclerosis in both clinical and research settings. The availability of public datasets and benchmarks is crucial for advancing the field. Existing datasets are often full of constraints helping some research, but lead to solutions that are not applicable to the real world. Our dataset provides a collection of real cases with an expert hand label, useful for train and evaluate algorithms, supporting new segmentation techniques. The algorithms presented together with this dataset serve as basic results with which to compare future methods.

In addition, many human-based checks were carried out to validate the data. Each step of the registration phase was checked by experts to verify that the co-registration and skull removal could be considered consistent. The labelling phase also went through many human checks to certify the quality. This was done by visual analysis of the output.

Code availability

The code used for preprocessing the dataset is available at https://github.com/alessiarondinella/MSLesSeg-2024. The MSSegDiff baseline algorithm is publicly available at https://github.com/alessiarondinella/MSSegDiff. The baseline implementations for SwinUNETR, UNETR and TransBTS can be accessed at the following links: https://github.com/Project-MONAI/research-contributions/tree/main/SwinUNETR for SwinUNETR, https://github.com/Project-MONAI/research-contributions/tree/main/UNETR for UNETR, and https://github.com/Rubics-Xuan/TransBTS for TransBTS.

Received: 9 December 2024; Accepted: 21 May 2025;
Published online: 31 May 2025

References

1. Filippi, M. *et al*. Multiple sclerosis. Nat. Rev. Dis. Primers 4 (2018).
2. Tullman, M. J. Overview of the epidemiology, diagnosis, and disease progression associated with multiple sclerosis. *Am J Manag. Care* **19**, S15–20 (2013).
3. Thompson, A. J. *et al*. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurol* **17**, 162–173 (2018).
4. Kuhlmann, T. *et al*. Multiple sclerosis progression: time for a new mechanism-driven framework. *The Lancet Neurol* **22**, 78–88 (2023).
5. Pinto, M. F. *et al*. Prediction of disease progression and outcomes in multiple sclerosis with machine learning. *Sci. reports* **10**, 21038 (2020).
6. Van der, W. & Chris, W. J. *et al*. Myelin quantification with mri: A systematic review of accuracy and reproducibility. *Neuroimage* **226**, 117561 (2021).
7. Molyneux, P. *et al*. Visual analysis of serial T2-weighted MRI in multiple sclerosis: intra-and interobserver reproducibility. *Neuroradiology* **41**, 882–888 (1999).
8. Aslani, S. *et al*. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *NeuroImage* **196**, 1–15 (2019).
9. Rondinella, A. *et al*. Boosting multiple sclerosis lesion segmentation through attention mechanism. *Comput. Biol. Medicine* **161**, 107021 (2023).
10. Shoeibi, A. *et al*. Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review. *Comput. Biol. Medicine* **136**, 104697 (2021).
11. Barquero, G. *et al*. Rimnet: A deep 3d multimodal mri architecture for paramagnetic rim lesion assessment in multiple sclerosis. *NeuroImage: Clin* **28**, 102412 (2020).
12. Fink, J. R., Muzi, M., Peck, M. & Krohn, K. A. Multimodality brain tumor imaging: Mr imaging, pet, and pet/mr imaging. *J. Nucl. Medicine* **56**, 1554–1561 (2015).
13. Janardhan, V., Suri, S. & Bakshi, R. Multiple sclerosis: hyperintense lesions in the brain on nonenhanced t1-weighted mr images evidenced as areas of t1 shortening. *Radiology* **244**, 823–831 (2007).
14. Rondinella, A. *et al*. ICPR 2024 Competition on Multiple Sclerosis Lesion Segmentation—Methods and Results. In *International Conference on Pattern Recognition* (pp. 1–16). Cham: Springer Nature Switzerland (November, 2024).
15. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841 (2002).
16. Smith, S. M. Fast robust automated brain extraction. *Hum. brain mapping* **17**, 143–155 (2002).
17. Carass, A. *et al*. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. NeuroImage 148, 77–102, https://doi.org/10.1016/j.neuroimage.2016.12.064 (2017).
18. Warfield, S. K., Zou, K. H. & Wells, W. M. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging* **23**, 903–921 (2004).
19. Guarnera, F. *et al*. MSLesSeg: baseline and benchmarking of a new Multiple Sclerosis Lesion Segmentation dataset. figshare https://doi.org/10.6084/m9.figshare.27919209 (2025).
20. Rondinella, A. *et al*. Enhancing multiple sclerosis lesion segmentation in multimodal mri scans with diffusion models. In 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 3733-3740 (IEEE, 2023).
21. Falk, T. *et al*. U-net: deep learning for cell counting, detection, and morphometry. *Nat. methods* **16**, 67–70 (2019).
22. Xing, Z., Wan, L., Fu, H., Yang, G. & Zhu, L. Diff-unet: A diffusion embedded network for volumetric segmentation. arXiv preprint arXiv:2303.10326 (2023).
23. Zhong, Z. *et al*. Squeeze-and-attention networks for semantic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 13065-13074 (2020).
24. Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020).
25. Hatamizadeh, A. *et al*. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In International MICCAI brainlesion workshop, 272-284 (Springer, 2021).
26. Hatamizadeh, A. *et al*. Unetr: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 574-584 (2022).
27. Wang, W. *et al*. Transbts: Multimodal brain tumor segmentation using transformer. In International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) (2021).

Acknowledgements

Francesco Guarnera is funded by the PNRR MUR project PE0000013-FAIR. Alessia Rondinella is a PhD candidate enrolled in the National PhD in Artificial Intelligence, XXXVII cycle, course on Health and life sciences, organized by Universitá Campus Bio-Medico di Roma. Francesco Pappalardo is funded through the Programma di ricerca CN00000013 -National Centre for HPC, Big Data and Quantum Computing-, finanziato dal Decreto Direttoriale di concessione del finanziamento n.1031 del 17.06.2022 a valere sulle risorse del PNRR-M4C2-Investimento 1.4—Avviso -Centri Nazionali—D.D. n. 3138 del 16 dicembre 2021.

Author contributions

F.G. and A.R contributed equally to this work, initiated the study, developed the code, and conducted the experiments. F.P. and S.B. supervised the project. D.M. and C.D.L. provided the scans and validated the labels produced. E.C., G.R. and F.P. contributed to the labeling process. F.G., A.R. and S.B. wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.G. or A.R.