Computers in Biology and Medicine 197 (2025) 110919

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

ConfLUNet: Multiple sclerosis lesion instance segmentation in presence of confluent lesions

Maxence Wynen [a,b,*], Pedro M. Gordaliza [c,d], Maxime Istasse [a], Anna Stölting [b], Pietro Maggi [b,e], Benoît Macq [a,1], Meritxell Bach Cuadra [c,d,1]

[a] ICTEAM, Université Catholique de Louvain, Louvain-la-Neuve, Belgium
[b] Louvain Neuroinflammation Imaging Lab (NIL), Université Catholique de Louvain, Brussels, Belgium
[c] CIBM Center for Biomedical Imaging, Switzerland
[d] Department of Radiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland
[e] Cliniques Universitaires Saint-Luc, Université Catholique de Louvain, Belgium

ARTICLE INFO

Keywords:
Multiple sclerosis
Instance segmentation
Confluent lesions
White matter lesion segmentation

ABSTRACT

Accurate lesion-level segmentation on MRI is critical for multiple sclerosis (MS) diagnosis, prognosis, and disease monitoring. However, current evaluation practices largely rely on semantic segmentation post-processed with connected components (CC), which cannot separate confluent lesions (aggregates of confluent lesion units, CLUs) due to reliance on spatial connectivity. To address this misalignment with clinical needs, we introduce formal definitions of CLUs and associated CLU-aware detection metrics, and include them in an exhaustive instance segmentation evaluation framework. Within this framework, we systematically evaluate CC and post-processing-based Automated Confluent Splitting (ACLS), the only existing methods for lesion instance segmentation in MS. Our analysis reveals that CC consistently underestimates CLU counts, while ACLS tends to oversplit lesions, leading to overestimated lesion counts and reduced precision. To overcome these limitations, we propose ConfLUNet, the first end-to-end instance segmentation framework for MS lesions. ConfLUNet jointly optimizes lesion detection and delineation from a single FLAIR image. Trained on 50 patients, ConfLUNet significantly outperforms CC and ACLS on the held-out test set (n = 13) in instance segmentation (Panoptic Quality: 42.0% vs. 37.5%/36.8%; p = 0.017/0.005) and lesion detection (F1: 67.3% vs. 61.6%/59.9%; p = 0.028/0.013). For CLU detection, ConfLUNet achieves the highest $textF1^{CLU}$ (81.5%), improving recall over CC (+12.5%, p = 0.015) and precision over ACLS (+31.2%, p = 0.003). By combining rigorous definitions, new CLU-aware metrics, a reproducible evaluation framework, and the first dedicated end-to-end model, this work lays the foundation for lesion instance segmentation in MS.

1. Introduction

Multiple sclerosis (MS) is a chronic immune-mediated disease and a leading cause of non-traumatic neurological disability in young adults [1]. MS is characterized by the presence of demyelinated lesions in the central nervous system, mainly located in the white matter (WM) and visible on conventional magnetic resonance imaging (MRI). Both the total lesion count and the lesion load (volume) play crucial roles in the diagnosis [2], prognosis [3], and monitoring [4–6] of MS. Recently, advanced lesion-level MRI biomarkers, such as the presence of paramagnetic rim or the central vein sign, are being incorporated in clinical guidelines as they enhance diagnostic accuracy [7,8], improve prognostic assessments [9], and/or uncover underlying pathological mechanisms [10]. Automated analyses of these biomarkers necessitate precise localization and delineation of individual lesions [11–16], which are generally achieved through lesion instance segmentation masks. Compared to semantic masks (Fig. 1b) which classify voxels as lesion/non-lesion, lesion instance segmentation masks (Fig. 1c) also associate each lesion voxel with a unique lesion identifier (id).

Naturally, manually annotating 3D lesion instance segmentation masks is even more laborious than creating their semantic counterparts, particularly in cases with many confluent lesions—aggregates of pathologically distinct focal areas of inflammation, arising from the confluence of spatially and usually temporally separated foci of blood–brain barrier disruption, demyelination and axonal damage. [17,18] (Fig. 1f). To facilitate understanding, we further define the distinct lesions constituting each confluent lesion as *confluent lesion units (CLU)*.

* Correspondence to: ELEN - Maxwell L5.03.02 Place du Levant 3 1348 Louvain-la-Neuve.
E-mail address: maxence.wynen@uclouvain.be (M. Wynen).
[1] The last two authors contributed equally.

https://doi.org/10.1016/j.compbiomed.2025.110919
Received 28 May 2025; Received in revised form 28 July 2025; Accepted 7 August 2025
Available online 23 August 2025

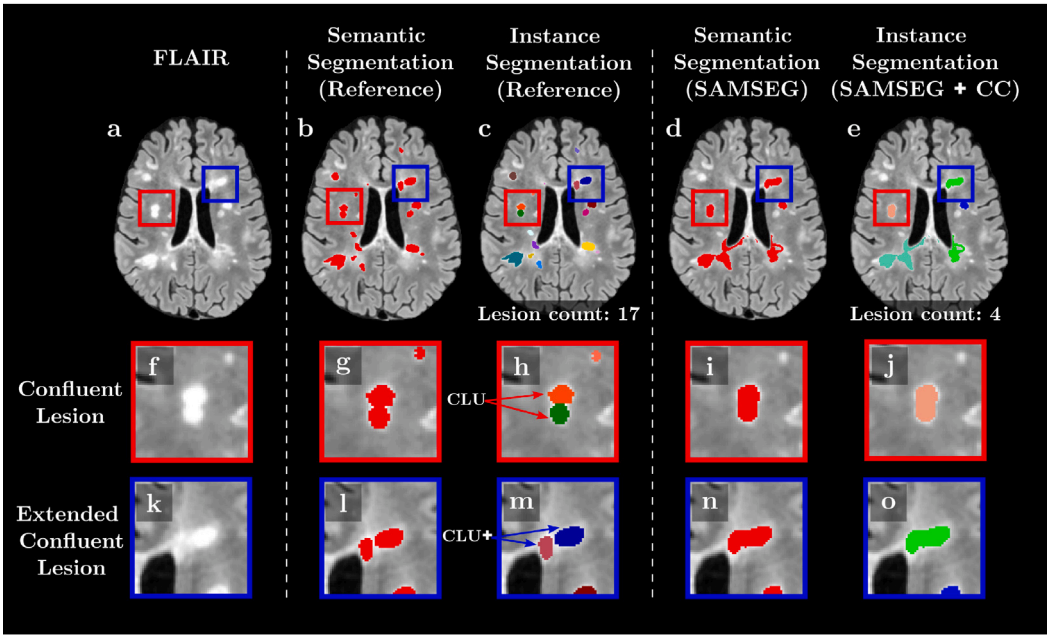

Fig. 1. Illustration of key concepts formalized in this study. (a) Axial FLAIR image; (b–c) Reference semantic and instance segmentations; (d–e) SAMSEG's predicted semantic segmentation and corresponding instance segmentation obtained by applying connected components (CC); (f) Close-up of a confluent lesion; (g–h) Reference semantic and instance segmentations of (f); red arrows indicate distinct confluent lesion units (CLUs); (i–j) SAMSEG predictions for (f); CC merges multiple CLUs into a single instance; (k) Close-up of an extended confluent lesion; (l–m) Reference semantic and instance segmentations of (k); blue arrows indicate extended CLUs (CLU+), defined as lesions whose connectivity becomes apparent only after semantic mask dilation; (n–o) SAMSEG predictions for (k); CC again merges multiple CLU+ into a single instance.

Confluent lesions may comprise anywhere from two to dozens of CLUs. These particularities of CLUs render their identification extremely difficult — if not impossible, particularly in late-stage patients (Fig. A.10) —, especially when using only cross-sectional images. Indeed, regular longitudinal scans with adequate resolution could theoretically help separate spatially confluent lesions based on their temporal appearance, but the frequent acquisition of such scans is very expensive, particularly given that those lesions can form rapidly within the timeframe of weeks, especially in patients with high disease activity [19]. This requirement conflicts with the typical clinical practice of scanning patients only once or twice a year. Although advanced MRI sequences (e.g. EPI, QSM) used to detect paramagnetic rims or perivenular lesions can support lesion instance identification, their high acquisition cost and limited clinical adoption constrain their practical utility.

Despite the increasing interest in lesion-level biomarkers, the vast majority of existing automated approaches for MS lesion instance segmentation do not directly tackle the problem of confluent lesions. Instead, most rely on semantic segmentation followed by connected component (CC) analysis, a simple post-processing step that identifies each connected voxel region in the semantic mask as a separate lesion (Fig. 1d–e). While effective for disconnected lesions, this approach struggles to distinguish CLUs, as their typical spatial connectivity causes CC to consider them as a single instance (Fig. 1e, j, and o), leading to inaccurate lesion counts. To our knowledge, only one alternative, we term *Automated Confluent Lesion Splitting (ACLS)*, attempts to directly address the challenge of lesion splitting [14,20]. However, ACLS, which also operates as a post-processing step on top of semantic segmentation, has not been rigorously validated in the context of instance segmentation. End-to-end instance segmentation could better address CLUs by jointly optimizing detection and segmentation but, to our knowledge, no such methods have yet been proposed.

In parallel, current evaluation strategies often prioritize semantic segmentation metrics, such as the Dice Score Coefficient, which do not fully capture the complexity of instance-level delineation, particularly in the presence of confluent lesions. Our review of recent literature (Appendix E) shows that although 53.5% of studies since 2014 (61 out of 114) also report lesion detection metrics, nearly all rely on CC to generate both predicted *and* reference lesion instances, thus implicitly ignoring CLUs. This results in a misalignment between current evaluation frameworks and clinical needs for accurate counting and lesion-level delineation. Notably, no prior work has rigorously evaluated methods in a true instance segmentation framework, nor have metrics been developed to specifically assess CLU detection accuracy.

This study makes three primary contributions: (i) we introduce the first end-to-end framework for MS lesion instance segmentation, extending our previous ConfLUNet framework [21], to jointly optimize lesion detection and delineation in the presence of CLUs; (ii) we propose an evaluation framework tailored to instance segmentation in MS, including new metrics grounded in formal definitions of confluent lesions and CLUs; and (iii) we conduct the first systematic comparison of existing approaches (CC and ACLS) within this framework, and further demonstrate that ConfLUNet achieves superior instance segmentation performance and promising results regarding CLU detection.

2. Problem formulation and related works

Building on earlier clinical descriptions, we now provide formal definitions to support the technical formulation of the problem.

Let $I : \Omega \to \mathbb{R}$ represent a 3D image, where $\Omega = \{(x, y, z) \in \mathbb{N}^3 | x < w, y < h, z < d\}$ is the voxel image domain. **Semantic segmentation masks** $\boldsymbol{S}$ classify each voxel in an image (Fig. 1b), in our case, delineating lesion tissue, but do not distinguish between individual lesions (i.e., $\boldsymbol{S} : \Omega \to \{0, 1\}$). **Instance segmentation masks** account for this by extending the semantic mask and assigning a unique identifier to each lesion instance (Figs. 1b and d). Formally, an instance segmentation mask $\boldsymbol{M} : \Omega \to \{0, 1, \dots, K\}$ assigns to each voxel a label, where positive values indicate lesion instances. The set of lesion instances is defined as $\mathcal{L} = \{L_1, L_2, \dots, L_K\}$, where each $L_k = \{(x, y, z) \in \Omega | \boldsymbol{M}(x, y, z) = k\}$, is the set of voxel coordinates belonging to the kth lesion. In medical image analysis, the most common method to create an instance segmentation mask from a semantic segmentation mask is via CC analysis (Fig. 1e, Section 2.1). Given a $\boldsymbol{S}$, CC analysis identifies disjoint connected regions. Let $\mathcal{B} = \{B_1, B_2, \dots, B_J\}$ represent the set of connected components ("blobs") in $\boldsymbol{S}$, where each $B_j \subset \Omega$

Table 1
Summary of mathematical formulations. IoU stands for Intersection over Union.

Concept	Mathematical definition	
CC-based Instance Segmentation	$M_{\mathrm{CC}}(\mathrm{x},\mathrm{y},\mathrm{z}) = \begin{cases} j & \text{if} \exists B_j \in \mathcal{B} \text{ such that } (x,y,z) \in B_j \\ 0 & \text{otherwise.} \end{cases}$	(1)
Confluent Lesions	$\mathcal{B}^C = \{B_j \in \mathcal{B} \mid \exists L_i, L_k \in \mathcal{L}, i \neq k, \text{ such that } \mathrm{IoU}(L_i, B_j) > 0 \text{ and } \mathrm{IoU}(L_k, B_j) > 0\}.$	(2)
Confluent Lesion Units (CLUs)	$\mathcal{L}^{CLU} = \{k \in \mathcal{L} \mid \exists B_j \in \mathcal{B}^C \text{ such that } \mathrm{IoU}(L_k, B_j) > 0\} = \{k \in \mathcal{L} \mid \exists B_j \in \mathcal{B} \text{ such that } 0 < \mathrm{IoU}(L_k, B_j) < 1\}.$	(3)
Extended CLUs (CLU+)	$\mathcal{L}^{CLU+} = \{L_i \in \mathcal{L} \mid \exists B_j \in \mathcal{B}^{C+} \text{ such that } \mathrm{IoU}(L_i, B_j) \neq 0\} = \{L_i \in \mathcal{L} \mid \exists B_j \in \mathcal{B}^{+}, \exists L_k \in \mathcal{L}, L_k \neq L_i \text{ such that } \mathrm{IoU}(L_i, B_j) \neq 0 \text{ and } \mathrm{IoU}(L_k, B_j) \neq 0\}.$	(4)

is the set of voxel coordinates belonging to the jth component. The instance segmentation mask $\boldsymbol{M}_{\mathbf{CC}}$ obtained by applying CC on $\boldsymbol{S}$ is defined in Table 1 (Eq. 1). Throughout this work, $\boldsymbol{M}_{\mathbf{CC}}$ is always computed with the most conservative structure (6-connectivity).

Confluent lesions and confluent lesion units. Using M_{CC} and $\mathcal{B}$, we define **confluent lesions** as connected components from M_{CC} that overlap with at least two reference lesion instances from the set $\mathcal{L}$ (Fig. 1g). Formally, the set of confluent lesions $\mathcal{B}^C \subset \mathcal{B}$ is defined by Eq. 1 (Table 1). Thus, a connected component B_j is confluent if at least two lesions L_i, L_k from $\mathcal{L}$ partially overlap with B_j. The individual lesions within a confluent lesion are referred to as **confluent lesion units (CLUs)** (Fig. 1h). Formally, the set of CLUs $\mathcal{L}^{CLU} \subset \mathcal{L}$ is defined by Eq. 1 (Table 1). Thus, each CLU L_i is a distinct lesion instance that overlaps at most partially with a connected component B_j.

Extended definition of confluent lesions. The standard definition of confluent lesions (Eq. 1) may prove overly conservative in practice due to two key factors. First, MS lesions on T2-weighted (T2w) images are often surrounded by non-specific hyper-intensities that are difficult to distinguish — even for trained radiologists — and are frequently included in segmentation masks. As a result, automated methods tend to produce larger connected regions than those seen in reference annotations, especially when trained on masks that do not separate these effects (e.g., SAMSEG in Fig. 1d). Second, manual delineation variability and partial volume effects can fragment visually confluent lesions in the reference masks.

To address these limitations and better reflect clinical confluence, we extend the definition of confluent lesions via morphological dilation. We first obtain S^+ by applying a single-iteration binary dilation to the semantic segmentation mask S, then extract the set of dilated connected components $\mathcal{B}^+$ via connected component analysis. Substituting B with B^+ in Eq. 1 yields $\mathcal{B}^{C+}$, the set of extended confluent lesions—dilated components containing multiple lesion instances (Fig. 1l). From $\mathcal{B}^{C+}$, we define the set of extended CLUs $\mathcal{L}^{CLU+}$ as all lesion instances that share a dilated component with at least one other lesion (Eq. 1, Table 1). For readability, we refer to $\mathcal{L}^{CLU+}$ as CLU+ throughout the manuscript.

2.1. Related works on MS lesion instance segmentation

Instance segmentation methods have been widely investigated in computer vision applied to natural images, with extensive reviews highlighting the strengths and limitations of different approaches [22,23]. These methods are commonly categorized as either bottom-up (e.g. Panoptic DeepLab [24]) or top-down (e.g. Mask-RCNN [25]). While end-to-end instance segmentation is common in biomedical imaging (e.g., for nuclei and glands [26]), radiological imaging typically relies on CC analysis for instance identification and downstream analysis [27,28]. Here, we review CC and ACLS, the only two methods proposed for this task in MS.

Connected components (CC) analysis (Fig. 1e) identifies lesion instances by grouping adjacent voxels based on a predefined connectivity structure (e.g., 6-, 18-, or 26-connectivity in 3D). While effective for separating spatially disconnected lesions, CC cannot distinguish multiple CLUs within confluent lesions—typically detecting only the instance with the largest overlap and missing the rest. Our previous work [29] showed that even with ideal segmentation, using CC reduces lesion detection sensitivity by 9.3

Automated confluent lesion splitting (ACLS), the only method specifically designed to address confluent lesions in MS, was introduced as part of a pipeline for detecting paramagnetic rim lesions [14]. ACLS operates in two steps and requires a lesion probability map S_p, typically generated by a semantic segmentation tool. First, lesion centers are detected; then, each lesion voxel in the binary mask S (obtained by thresholding S_p) is assigned to the nearest center, producing an instance map M. The key innovation lies in the center detection step, originally proposed by Dworkin et al. [20] to handle lesion counting in presence of confluent regions. It hypothesizes that voxel probabilities peak at the lesion center — where blood–brain barrier breakdown begins — and decrease outward, reflecting radial inflammatory spread. Center voxels are identified as local maxima with negative Hessian eigenvalues in all three spatial directions, then clustered via CC. In the second step, each voxel in S is assigned to its nearest detected center using a nearest-neighbor approach. Despite its innovative approach, ACLS has some limitations. First, its core assumption — that lesion voxel probabilities peak at the center and decrease peripherally — remains unproven. Second, the method does not consider lesion size during clustering. As illustrated in Fig. C.11, this can lead to incorrect assignments when a small lesion is confluent with a larger one, causing voxels from the larger lesion to be incorrectly attributed to the smaller due to proximity. Finally, ACLS has not been formally validated for lesion instance segmentation. This component was not the primary focus of the original study and has not been evaluated independently. Still, in our previous work [29], we compared ACLS to CC on a private dataset and found that it achieved higher lesion detection sensitivity (+19.1% on average across segmentation tools).

3. Proposed framework

ConfLUNet is a 3D instance segmentation framework for MS lesions in MRI, integrating strengths from two established architectures: nnUNet [27] and Panoptic DeepLab [24]. It builds upon nnUNet's robust, self-configuring 3D U-Net backbone — including components like data fingerprinting, automatic parameter tuning, and data augmentation — while incorporating key ideas from Panoptic DeepLab for instance-aware representation learning. ConfLUNet's pipeline is described in Fig. 2. Additionally, its source code is open, available on GitHub,[2] and containerized using Docker.[3]

[2] https://github.com/maxencewynen/ConfLUNet/
[3] https://hub.docker.com/repository/docker/petermcgor/conflunet

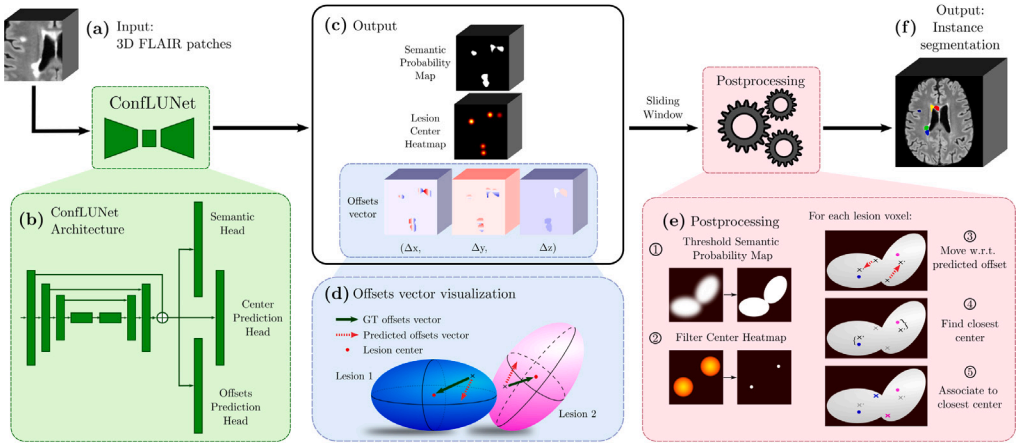

Fig. 2. Illustration of ConfLUNet pipeline. 3D FLAIR patches (a) are provided as input to the ConfLUNet network (b), which simultaneously predicts a semantic probability map, a lesion center heatmap, and an offset vector map (c). For each voxel, the offset vector indicates the direction and distance to the center of the lesion to which it belongs (d). After applying a sliding window inference strategy, the outputs are post-processed (e) to generate the final instance segmentation map (f). This post-processing involves binarizing the semantic probability map using a fixed threshold, filtering the center heatmap to identify discrete lesion centers, and assigning each predicted lesion voxel to the center that lies the closest to their initial position shifted by their predicted offset vector (e).

3.1. Architecture

The resulting architecture (Fig. 2b) follows a standard encoder–decoder U-Net design with skip connections and three output heads: one for semantic segmentation, one for lesion center detection, and one for predicting offset vectors from voxels positions to their corresponding centers. These additions enable ConfLUNet to capture both the extent and the instance structure of confluent MS lesions. ConfLUNet's three heads are as follows:

1. **Semantic Segmentation:** This head, inherited from nnUNet, outputs a two-channel semantic map (background/lesion). The loss L_{seg} combines focal loss and Dice loss: $L_{\text{seg}} = L_{\text{focal}} + \gamma L_{\text{dice}}$, where L_{focal} mitigates class imbalance by focusing on hard voxels, and L_{dice} measures overlap with reference. We use $\gamma = 0.5$, following prior work [30].
2. **Center Prediction:** This head outputs a heatmap where each voxel represents the probability of being a lesion center. The reference is a binary mask of lesion centers of mass $C = C_k = (C_k^x, C_k^y, C_k^z); |; k \in \mathcal{L}$, convolved with a 3D Gaussian ($\sigma = 2$ voxels). The loss L_{center} is the mean squared error between predicted and reference heatmaps.
3. **Offset Prediction:** This head predicts 3D offset vectors from each voxel's position to the center of its lesion instance (Fig. 2c,d). During training, reference offsets are computed as the difference between each lesion voxel's coordinates (x, y, z) and its corresponding lesion center:

$$\mathcal{O}(x,y,z) = \begin{cases} (C_k^x - x, C_k^y - y, C_k^z - z) & \text{if } (x,y,z) \in \mathcal{L}_k, \\ (0,0,0) & \text{otherwise,} \end{cases}$$

The offset loss L_{offsets} is computed using the L1 norm.

The final instance segmentation mask is obtained by combining semantic, center, and offset outputs during post-processing (Section 3.3).

3.2. Training

ConfLUNet is trained using 5-fold cross-validation, with ensemble predictions applied at inference. The total loss combines semantic segmentation (L_{seg}), center prediction (L_{center}), and offset prediction (L_{offsets}) losses:

$$L_{\text{total}} = L_{\text{seg}} + \alpha \cdot L_{\text{center}} + \beta \cdot L_{\text{offsets}}.$$

To balance magnitudes, we set $\alpha = 100$ and $\beta = 1$. Optimization is performed with Adam [31], using a constant learning rate of 0.002 and weight decay of 3×10^{-5}. The final model is trained for 15,000 epochs, while hyperparameter tuning is conducted over 2500 epochs.

We adapt nnUNet's patch sampling by disabling subject replacement and enforcing a 50% probability of sampling patches containing lesion voxels. Validation is performed every 25 epochs via full inference on the validation set, using panoptic quality (Section 4.2) as the epoch selection metric.

3.3. Inference

The full inference process is shown in Fig. 2e. A sliding window with 50% overlap and Gaussian importance weighting aggregates the three output patches (segmentation, center and offset maps). The semantic segmentation mask $\hat{S}$ is obtained by thresholding the softmax output $\hat{S}_p$ at 0.5. Lesion voxel clustering then starts by detecting predicted lesion centers $\hat{C} = \{\hat{C}_k : (x_k, y_k, z_k)\}$ via max pooling and filtering of the center heatmap.

An embedding space is created by adding each voxel's predicted offset vector $\hat{\mathcal{O}}(x,y,z)$ to its spatial coordinates. Each lesion voxel is then assigned to the nearest predicted center in this space. Specifically, the instance *id* $\hat{c}(x,y,z)$ of a voxel is the ID of the closest predicted center after applying the offset, expressed as:

$$\hat{C}(x,y,z) = \underset{k}{\operatorname{argmin}} \left\| \hat{C}_k - \left((x,y,z) + \hat{\mathcal{O}}(x,y,z)\right) \right\|^2,$$

where $\hat{C}_k$ denotes the coordinates of the kth predicted center. Finally, following the clinical definition of an MS lesion [32], predicted instances smaller than 3 mm along any axis or under 14 mm^3 in volume are removed.

4. Evaluation framework

4.1. Materials

4.1.1. Dataset and image acquisition details

This study used a private retrospective cohort of 65 patients recruited between January 2021 and April 2023 at the Saint-Luc University Hospital, Brussels Belgium. Following qualitative evaluation by clinical experts, two patients were excluded due to extreme lesion confluency, rendering their confluent lesions indiscernible. The remaining 63 patients (40 female) were diagnosed with MS per the 2017 McDonald criteria [2], including 45 with relapsing-remitting

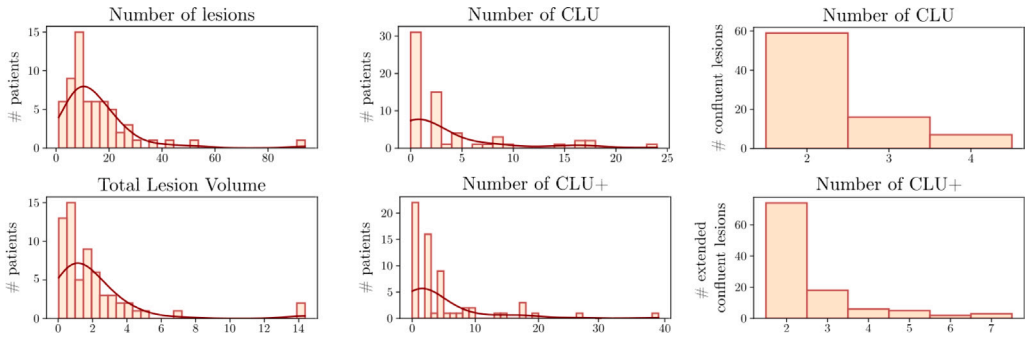

Fig. 3. Histograms of lesion and CLU(+) distributions in the dataset.

(RRMS), 11 secondary progressive (SPMS), and 7 primary progressive (PPMS) forms. The cohort's *mean* (*median*, *[range]*) age is 42.4 ± 12.1 (42.0, [21.5–67.3]) years old, EDSS 2.7 ± 2.0 (2.0, [0–7]), and disease duration 7.4 ± 8.0 years (5.9, [0–33.3]).

The study explores cross-sectional MRI acquisitions conducted at 3T whole-body MR scanner. The protocol (detailed in Appendix B) included 3D Fluid Attenuated Inversion Recovery (FLAIR), 3D Magnetization-Prepared Rapid Gradient Echo (MPRAGE) and 3D Echo Planar Imaging (EPI). FLAIR images were registered to the EPI magnitude images ($0.66 \times 0.66 \times 0.7$ mm) for another study, and MPRAGE was aligned to FLAIR. All registrations used ANTS rigid registration (v2.5.0) [33]. Brain masks were generated with Synthstrip [34] on FLAIR images and applied to all registered scans. Data management and registration were performed using the BIDS Managing and Analysis Tool (BMAT) [35].

4.1.2. Manual segmentation

Manual instance segmentation was performed by a trained neurobiologist (A.S., 1.5 years' experience) and reviewed by an experienced neurologist (P.M., > 10 years). Annotations were based on FLAIR images, with MPRAGE and EPI phase images consulted for challenging confluent lesions. Only clearly inflammatory WM lesions ≥3 mm in diameter [32] were included, excluding infratentorial and cortical lesions. All annotations and visualizations were done using ITK-Snap (v3.8.0) [36].

The dataset included 973 lesions, averaging 15.4 per subject (median: 11; range: 1–94; standard deviation: 13.8). Total lesion volume averaged 2048.8 mm^3 per subject (median: 1431.7 mm^3; range: 33.0–14368.0 mm^3; standard deviation: 2603.7 mm^3). According to Section 2 definitions, 32 patients had at least two CLUs and 41 had at least two CLU+. On average, each subject had 3.2 CLUs and 4.5 CLU+, totaling 199 and 285 across the cohort. These distributions are shown in Fig. 3.

4.2. Evaluation metrics

To better reflect clinical relevance, we evaluate MS lesion instance segmentation using established instance segmentation metrics. For completeness, we also report semantic segmentation and detection metrics, following the *Metrics Reloaded* framework [37]. In addition, we introduce three custom metrics specifically designed to assess CLU detection. All metrics are computed per subject and averaged across the cohort.

Lesion matching strategy. To compute instance-level metrics, we establish a one-to-one correspondence between predicted and reference lesions (Fig. 4). Our strategy ensures: (i) each reference lesion is matched to the predicted lesion with the highest overlap (if IoU ≥ $\lambda \in]0,1]$); (ii) each predicted lesion is matched similarly to its best-overlapping reference; and (iii) only mutually consistent (bidirectional) matches are retained. Formally, we define a matching function $\phi^\lambda : L_1 \to L_2 \cup \emptyset$ that assigns to each instance in L_1 its best match in L_2 if the maximum IoU exceeds λ, or $\emptyset$ otherwise:

$$\phi^\lambda(l_1) = \begin{cases} \underset{l_2 \in L_2}{\operatorname{argmax}} \operatorname{IoU}(l_1, l_2) & \text{if } \operatorname{IoU}(l_1, l_2) \geq \lambda \\ \emptyset & \text{otherwise} \end{cases} \tag{5}$$

While a 0.5 IoU threshold is common in 2D image tasks [24], we use $\lambda = 0.1$ to accommodate the variability in 3D lesion size and shape [37].

We define $\mathcal{P}$ as the set of mutually matched pairs between predicted instances $\hat{\mathcal{L}}$ and reference instances $\mathcal{L}$:

$$\mathcal{P} = \left\{(i,j) | j = \phi^\lambda(i), i = \phi^\lambda(j), \quad i \in \hat{\mathcal{L}}, j \in \mathcal{L}\right\}$$

A pair (i,j) is included in $\mathcal{P}$ only if i is the best match for j and vice versa. This strict 1:1 matching avoids cases where a single prediction covers multiple reference lesions, or a reference lesion is split across multiple predictions (Fig. 4a).

Detection and semantic segmentation accuracy. From $\mathcal{P}$, we derive the true positive (TP), false negative (FN), and false positives (FP) sets. Then, we compute overall lesion-wise Precision, Recall, and F1 score.

Additionally, we assess the semantic segmentation performance using classical Dice score coefficient (DSC) and normalized DSC (nDSC) [38] which reduces the DSC bias due to patients with different lesion loads.

Instance segmentation assessment. Panoptic Quality (PQ), our primary instance segmentation metric, combines Segmentation Quality (SQ) and Recognition Quality (RQ) [24] as PQ = SQ × RQ, where:

$$\text{SQ} = \frac{\sum_{(i,j)\in\mathcal{P}} \operatorname{IoU}(i,j)}{|\text{TP}|}; \quad \text{RQ} = \frac{|\text{TP}|}{|\text{TP}| + \frac{1}{2}|\text{FP}| + \frac{1}{2}|\text{FN}|}. \tag{6}$$

To evaluate CLU detection, we define the set of matched pairs $\mathcal{P}^{\text{CLU}}$ between predicted lesions and reference CLU instances:

$$\mathcal{P}^{\text{CLU}} = \left\{(i,j) | j = \phi^\lambda(i), i = \phi^\lambda(j), \quad i \in \hat{\mathcal{L}}, j \in \mathcal{L}^{\text{CLU}}\right\}$$

where $\mathcal{L}^{\text{CLU}}$ is the set of CLU reference instances (Table 1). Then (Fig. 4b):

- $\text{TP}^{\text{CLU}} = \mathcal{P}^{\text{CLU}}$: matched reference CLU considered correctly detected;
- $\text{FN}^{\text{CLU}} = \{j \in \mathcal{L}^{\text{CLU}} | \nexists i \in \hat{\mathcal{L}}, (i,j) \in \mathcal{P}^{\text{CLU}}\}$: reference CLU with no matching prediction;
- $\text{FP}^{\text{CLU}} = \{i \in \hat{\mathcal{L}} | \exists j \in \mathcal{L}, j = \phi^\lambda(i) \wedge i \neq \phi^\lambda(j)\}$: predicted lesions that match to a reference lesion ($j = \phi^\lambda(i)$) but are *not* the best match for that reference lesion ($i \neq \phi^\lambda(j)$)

These sets are used to compute CLU-wise precision ($\text{Precision}^{\text{CLU}}$ sensitive to over-splitting), recall ($\text{Recall}^{\text{CLU}}$, sensitive to under-splitting) and the corresponding F1 score (F1^{CLU}). Boundary conditions are handled as follows: if no confluent lesions are present in the reference annotations, $\text{Recall}^{\text{CLU}} = 1$ by definition. Similarly, $\text{Precision}^{\text{CLU}} =$

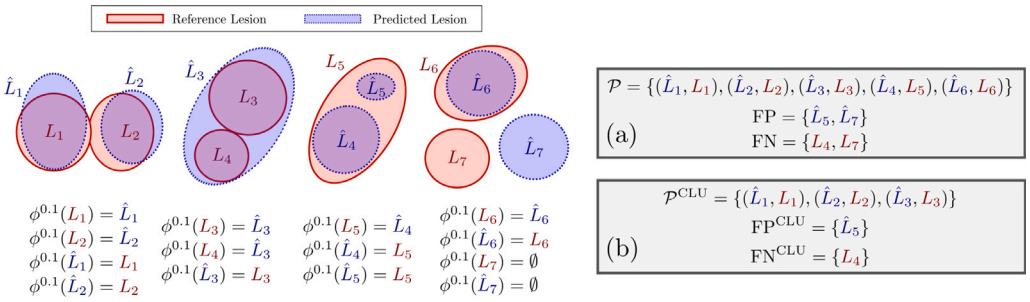

Fig. 4. Examples of the mutually consistent lesion matching strategy and of $\mathcal{P}^{\text{CLU}}$ (= TP^{CLU}), FP^{CLU}, and FN^{CLU}.

1 when both true and false positive CLU detections are absent ($\text{TP}^{\text{CLU}} = \text{FP}^{\text{CLU}} = 0$), reflecting that no incorrect predictions were made.

The same logic and evaluation procedure apply to CLU+ lesions, as defined in Eq. 1 (Table 1), with the corresponding metrics denoted as $\text{Precision}^{\text{CLU+}}$, $\text{Recall}^{\text{CLU+}}$, and $\text{F1}^{\text{CLU+}}$.

4.3. Experimental setting

Experiment 1: Instance segmentation with CC and ACLS. This experiment formally evaluated the instance segmentation performance of CC and ACLS using 3D FLAIR images from all 63 patients (Section 4.1). Probabilistic semantic segmentation masks (S_p) were generated using three established tools: (i) LST-LPA, a logistic regression model in SPM [39]; (ii) FLAMeS, a nnUNet model pretrained for MS lesion segmentation [40,41]; and (iii) SAMSEG [42], a probabilistic atlas-based tool.

All outputs were binarized at 0.5 and post-processed into instance masks using either CC or ACLS. CC used a 6-connectivity structure to avoid over-merging, while ACLS employed 26-connectivity for lesion center detection, which yielded better results. Finally, small predicted lesions were discarded based on earlier described strategy. This yielded six instance segmentation maps per patient, which were evaluated using the metrics described in Section 4.2.

Experiments 2 and 3: Validating ConfLUNet. To evaluate ConfLUNet, the dataset was split into training (n = 50) and test (n = 13) sets, ensuring balanced distribution of patients with confluent lesions. A baseline model, matching ConfLUNet's architecture but excluding the center and offset heads, was trained with only the segmentation loss. Training followed nnUNet's pipeline [27], including a "Poly" learning rate scheduler (initial rate 0.01), patch sampling strategy, and 15,000 training epochs. Validation was performed every 25 epochs, computing PQ on the validation set. To compare ConfLUNet with CC and ACLS fairly, the baseline model's outputs were post-processed using both CC and ACLS under the same conditions as in Experiment 1. The best-performing epoch was selected based on PQ, ensuring consistency across methods. Experiment 2 reports results from 5-fold cross-validation on the training set for ConfLUNet, U-Net+CC, and U-Net+ACLS, while Experiment 3 evaluates the final models on the held-out test set.

4.4. Statistical analyses

Statistical analyses were conducted at the patient level using Python's SciPy library [43]. We used the Wilcoxon signed-rank test for paired method comparisons (with Holm–Bonferroni correction for multiple comparisons provided in supplementary materials). Bland–Altman plots were used to assess the agreement between predicted and reference lesion, CLU, and CLU+ counts. Spearman's rank correlation coefficient quantified relationships between means and differences. Statistical significance was set at $p < 0.05$.

5. Results

5.1. Evaluating CC and ACLS for lesion instance segmentation

Our first experiment, which compared the performance of CC and ACLS when applied to outputs from LST-LPA, SAMSEG, and FLAMeS, revealed distinct behavioral patterns between the two methods (Fig. 5). ACLS achieved slightly higher instance segmentation performance than CC (PQ: +1.6% on average), with significant differences for SAMSEG and FLAMeS, but not for LST-LPA. It also outperformed CC in semantic segmentation metrics, with modest yet significant improvements in DSC (+1.0%) and nDSC (+0.2%), and a notably higher recall (+17.7% on average). In contrast, CC yielded significantly higher precision (+11.5%), but no significant differences were observed in overall lesion detection (F1). For CLU detection, CC consistently achieved significantly higher precision (+22.5%), while ACLS showed superior recall (+20.7%). These trends persisted when using the CLU+ definition. Although no significant difference was found in F1^{CLU}, the use of ACLS instead of CC led to significantly higher $\text{F1}^{\text{CLU+}}$ scores (+8.2% on average). Nearly all findings were consistent across all three segmentation tools. Interestingly, limiting the analysis to patients with confluent lesions resulted in a substantial drop in CC's CLU recall (LST-LPA: 70.0%→41.0%; SAMSEG: 64.4%→29.8%; FLAMeS: 66.8%→34.7%), whereas ACLS experienced a more moderate decrease (89.8%→80.0%; 84.1%→68.7%; 89.2%→78.8%).

Bland–Altman plots (Fig. 6) for overall lesion count exhibit no clear trend across segmentation tools when using CC. In contrast, ACLS systematically overestimates lesion count, with this overestimation increasing alongside the average of predicted and reference counts. However, ACLS provides more accurate estimates for CLU and CLU+ counts, as indicated by lower mean absolute differences and narrower 95% limits of agreement. Conversely, CC consistently underestimates these counts, with the difference between predicted and reference CLU count decreasing as their average increases.

5.2. Cross-validation

The results of the 5-fold CV analysis (Table 2) revealed that ConfLUNet outperformed both CC and ACLS in instance segmentation (+2.3% and +2.8%, resp.) and overall lesion detection (F1 +1.4% and +2.1%). On the other hand, semantic segmentation-based approaches showed higher DSC (+2.4%/2.3% for CC/ACLS) and nDSC (+2.7%/+2.1% for CC/ACLS). CC and ACLS demonstrated comparable performance in terms of PQ, DSC, and F1, with their main difference lying in the balance between recall and precision: ACLS favors recall, while CC prioritizes precision, following the trend found in Experiment 1 (Section 5.1). ConfLUNet achieves a higher F1 score by striking a compromise between the two, attaining greater recall than CC (though lower than ACLS) and higher precision than ACLS (though lower than CC).

Regarding CLU detection (Table 3), ConfLUNet achieved the highest F1^{CLU} score, with improvements of +4.1% and +15.2% compared to

Table 2
Results of the 5-fold cross-validation comparison between ConfLUNet and 3D U-Net semantic segmentation baseline model + Connected Components (CC) or Automated Confluent Lesion Splitting (ACLS) as post-processing. Variations across folds are indicated. Column groups separated by dashed lines represent the three task evaluations: instance segmentation, semantic segmentation and detection. PQ: Panoptic Quality, DSC: Dice Score, nDSC: normalized dice score.

Method	PQ (%)	DSC (%)	nDSC (%)	F1 (%)	Recall (%)	Precision (%)
3D U-Net + CC	47.9 ± 4.6	**70.7 ± 2.9**	**77.5 ± 1.5**	74.9 ± 4.4	72.3 ± 7.5	**81.9 ± 3.7**
3D U-Net + ACLS	47.4 ± 3.2	70.6 ± 2.8	76.9 ± 1.7	74.2 ± 3.1	**79.0 ± 4.2**	72.6 ± 3.9
ConfLUNet (ours)	**50.2 ± 3.0**	68.3 ± 3.2	74.8 ± 3.9	**76.3 ± 3.7**	75.6 ± 9.1	80.2 ± 3.3

Table 3
Confluent lesion unit (CLU) detection performance resulting from the 5-fold cross-validation comparison between ConfLUNet and 3D U-Net semantic segmentation baseline model + Connected Components (CC) or Automated Confluent Lesion Splitting (ACLS) as post-processing. Variations across folds are indicated.

Method	$F1^{CLU}$ (%)	$Recall^{CLU}$ (%)	$Precision^{CLU}$ (%)	$F1^{CLU+}$ (%)	$Recall^{CLU+}$ (%)	$Precision^{CLU+}$ (%)
3D U-Net + CC	79.9 ± 5.3	72.0 ± 6.5	**97.7 ± 3.9**	78.4 ± 5.8	69.3 ± 7.4	**99.8 ± 0.4**
3D U-Net + ACLS	68.8 ± 9.3	**88.1 ± 2.9**	70.1 ± 9.9	74.6 ± 8.3	**86.8 ± 3.9**	75.9 ± 8.1
ConfLUNet (ours)	**84.0 ± 5.3**	80.0 ± 8.2	92.8 ± 4.8	**82.9 ± 2.8**	77.1 ± 6.2	95.5 ± 4.5

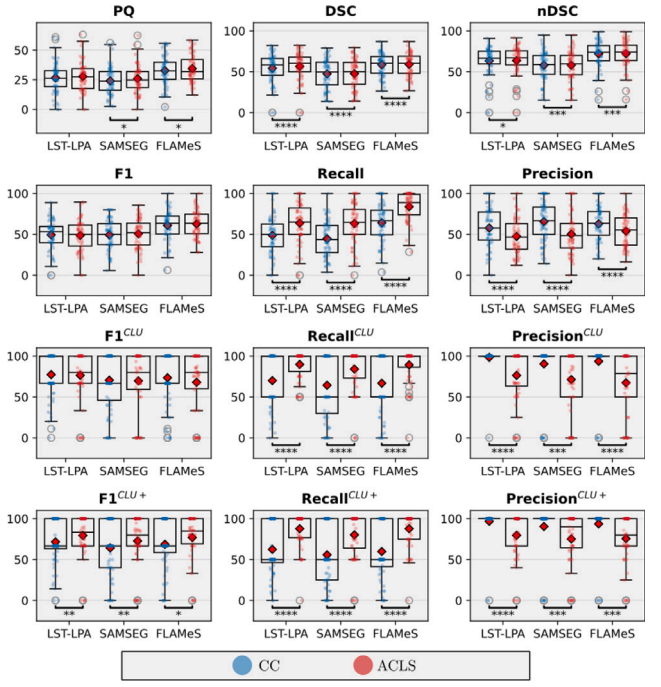

Fig. 5. Box plots of the patient-wise metrics obtained for the three lesion segmentation tools, to which we applied either CC or ACLS. P-values are computed with the Wilcoxon signed rank test. ACLS: Automated Confluent Lesion Splitting; CC: Connected Components; DSC: Dice Score Coefficient; nDSC: normalized Dice Score Coefficient; CLU: Confluent Lesion Unit; CLU+: Extended definition of Confluent Lesion Units (Eq. 1).

CC and ACLS, respectively, reflecting a strong balance between precision and recall. CC exhibited the highest Precision but the lowest Recall, while ACLS showed the highest Recall and the lowest Precision. These trends persisted upon further examination of CLU+ (eq. 1). All three methods showed moderate gains in precision and slight declines in recall, with corresponding changes in $F1^{CLU+}$: ConfLUNet and CC experienced slight decreases (−1.1% and −1.5%, respectively), while ACLS improved (+6.2%). Notably, CC achieved near-perfect precision (99.8%) under the CLU+ definition.

5.3. Performance on held-out test set

Similar to nnUNet, we performed ensemble inference on the independent test set — including 13 patients — using the models trained on each fold. Among these patients, seven presented at least two CLUs, yielding a total of 41 CLUs (range: 0–24 per patient), and ten had at least two CLU+, totaling 66 (range: 0–26 per patient). Table 4 and Fig. 7 highlight ConfLUNet's improved performance across instance segmentation and general lesion detection metrics on the held-out test set. ConfLUNet significantly outperformed both CC and ACLS in PQ (resp. +4.5%, $p = 0.017$ and +5.2%, $p = 0.005$) and F1 score (resp. +4.5%, $p = 0.028$ and +5.2%, $p = 0.013$). We found no significant difference in semantic segmentation metrics across methods. In terms of detection, ConfLUNet achieved the highest Recall, reaching a significant difference only against CC ($p = 0.015$), and the highest Precision, reaching a significant difference only against ACLS ($p = 0.003$). Notably, all methods exhibited a general decrease in performance on the test set compared to the 5-fold CV results (Section 5.2).

Regarding confluent lesions (Table 5, Fig. 7), ConfLUNet achieved the highest $F1^{CLU}$ score, surpassing CC by 1.7% and ACLS by 18.1%, though without statistical significance. Compared to CC, it showed significantly higher recall (+12.5%, $p = 0.015$) and slightly lower precision (−8.1%, not significant). Against ACLS, ConfLUNet achieved significantly higher precision (+31.2%, $p = 0.003$), with a non-significant drop in recall (−10.0%). Under the extended CLU+ definition (Eq. 1), ConfLUNet improved its own $F1^{CLU}$ by 7.2%, while CC declined (−1.6%) and ACLS gained 6.2%. The improvement over CC was significant ($p = 0.046$), but not over ACLS ($p = 0.08$). ConfLUNet also showed significantly higher recall than CC (+16.1%, $p = 0.015$), and significantly higher precision than ACLS (+29.7%, $p = 0.011$), with no significant difference in precision compared to CC. Overall, ConfLUNet achieves a more robust balance between precision and recall, particularly under the clinically motivated CLU+ definition.

When computing metrics across all lesions in the test set — rather than averaging patient-wise — to obtain a global view of CLU detection performance (Table 6), the gap between ConfLUNet and CC widens for both $F1^{CLU}$ (+25.6%) and $F1^{CLU+}$ (+22.0%). Interestingly, CC's recall drops strikingly to 41.5% for CLU and 50.0% for CLU+, whereas all metrics are relatively similar to their patient-wise equivalent for ConfLUNet and ACLS. This lesion-wise analysis is valuable, as it gives more weight to patients with a higher number of confluent lesions—typically those with longer disease duration and/or more aggressive progression, for whom accurately separating these confluent lesions is clinically most relevant.

Finally, Bland–Altman plots (Fig. 8) reveal similar trends as described in Section 5.1, with CC underestimating CLU count and ACLS overestimating overall lesion count. ConfLUNet, however, exhibited a larger bias compared to CC (resp. 4.5 and 3.3), but smaller compared to ACLS (10.8), and weaker trends compared to both CC and ACLS, suggesting greater consistency across lesion load levels. ConfLUNet's errors in estimating CLU counts were less systematically affected by the reference CLU count, as indicated by the lower coefficient of determination. Illustrative examples of predictions from ConfLUNet, CC and ACLS on the test set are shown in Fig. 9.

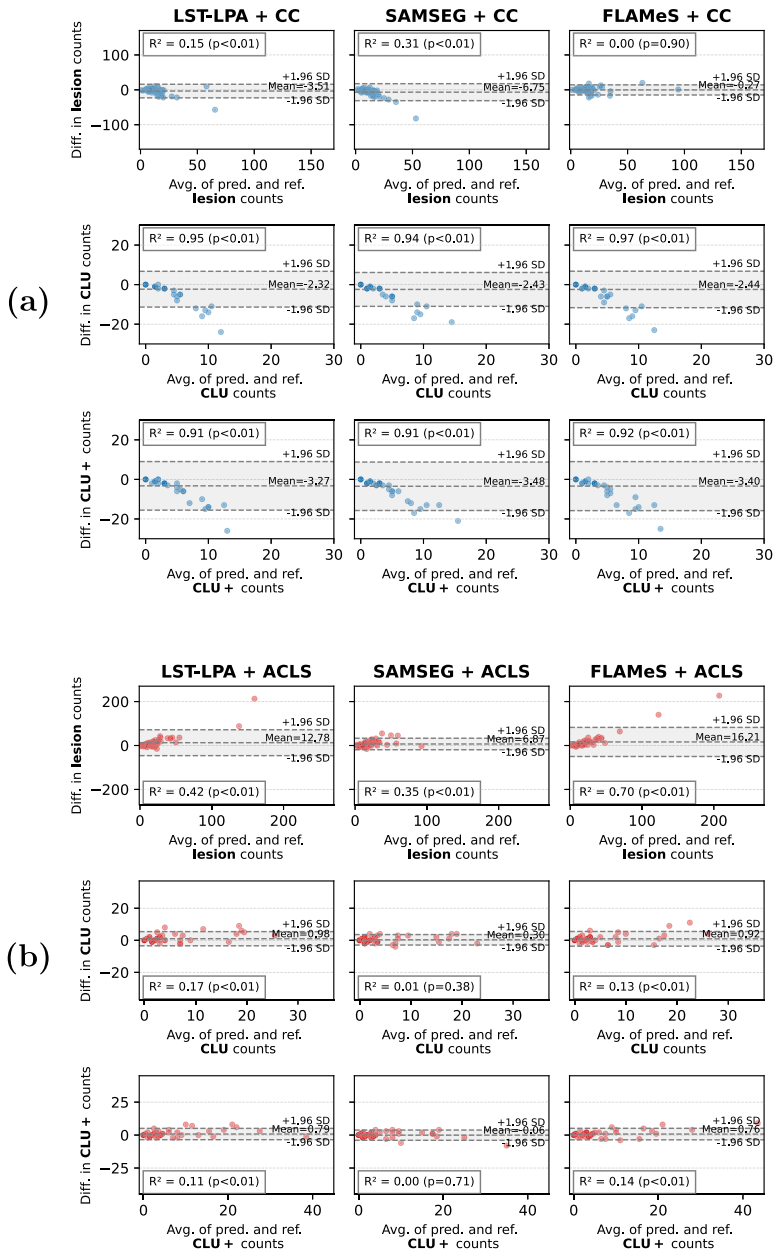

Fig. 6. Bland–Altman plots for predicted and reference lesion/CLU/CLU+ counts using masks from the lesion segmentation tool LST, with either (a) CC or (b) ACLS post-processing. CLU: Confluent Lesion Unit; CLU+: Extended definition of CLUs.

Table 4
Performance on the test set (average and standard deviation) of ConfLUNet and 3D U-Net semantic segmentation baseline model + Connected Components (CC) or Automated Confluent Lesion Splitting (ACLS) as post-processing. Column groups separated by dashed lines represent the three task evaluations: instance segmentation, semantic segmentation and detection. PQ: Panoptic Quality, DSC: Dice Score, nDSC: normalized dice score, DiC: Absolute Different in Count.

Method	PQ (%)	DSC (%)	nDSC (%)	F1 (%)	Recall (%)	Precision (%)	DiC ↘
3D U-Net + CC	37.5 ± 11.9	69.0 ± 10.3	72.7 ± 10.7	61.6 ± 13.2	69.2 ± 17.5	57.2 ± 12.9	**4.7 ± 3.3**
3D U-Net + ACLS	36.8 ± 12.1	**69.5 ± 11.1**	73.2 ± 11.6	59.9 ± 14.5	78.4 ± 21.7	50.4 ± 13.8	11.4 ± 12.7
ConfLUNet (ours)	**42.0 ± 10.2**	69.0 ± 9.7	**73.6 ± 9.7**	**67.3 ± 11.2**	**78.8 ± 13.0**	**60.2 ± 13.6**	5.1 ± 3.3

Table 5
Performance (average and standard deviation across patients) of confluent lesion unit (CLU) and extended confluent lesions (CLU+) (Eq. 1) detection on the test set. Comparison between ConfLUNet and 3D U-Net semantic segmentation baseline model + Connected Components (CC) or Automated Confluent Lesion Splitting (ACLS) as post-processing.

Method	$F1^{CLU}$ (%)	$Recall^{CLU}$ (%)	$Precision^{CLU}$ (%)	$F1^{CLU+}$ (%)	$Recall^{CLU+}$ (%)	$Precision^{CLU+}$ (%)
3D U-Net + CC	79.8 ± 19.3	70.8 ± 27.4	**100.0 ± 0.0**	78.2 ± 17.6	67.9 ± 24.7	**100.0 ± 0.0**
3D U-Net + ACLS	63.4 ± 39.2	**93.3 ± 15.3**	60.7 ± 40.4	71.5 ± 33.4	**87.4 ± 27.0**	68.0 ± 35.0
ConfLUNet (ours)	**81.5 ± 27.0**	83.3 ± 20.7	91.9 ± 26.6	**88.9 ± 11.0**	84.0 ± 17.6	97.7 ± 6.7

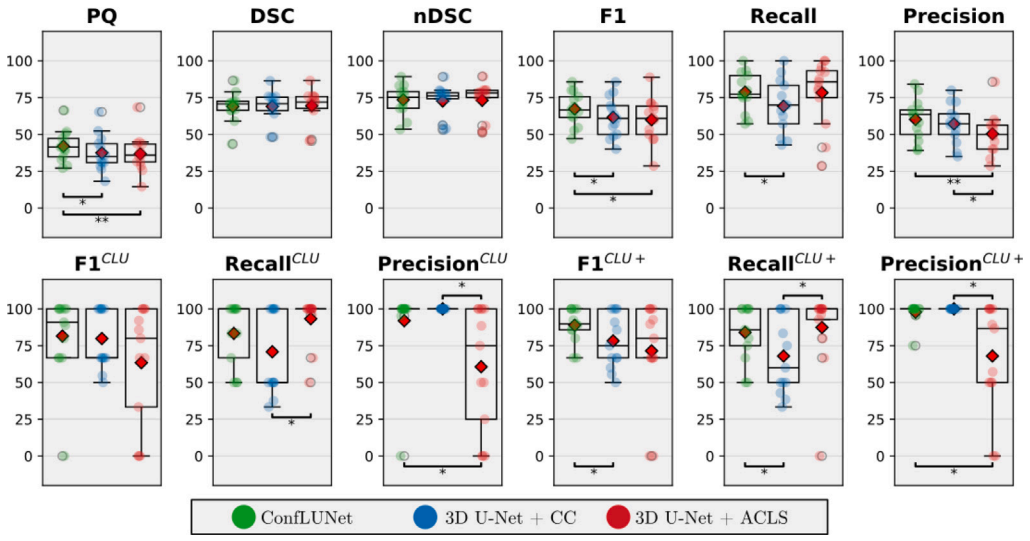

Fig. 7. Box plots of the patient-wise general metrics obtained for the three models. The p-values are computed with the Wilcoxon signed rank test. ACLS: Automated Confluent Lesion Splitting; CC: Connected Components; DSC: Dice Score Coefficient; nDSC: normalized Dice Score Coefficient; CLU: Confluent Lesion Unit; CLU+: Extended definition of Confluent Lesion Units (Eq. 1).

Table 6
Lesion-wise performance of confluent lesion unit (CLU) and extended confluent lesions (CLU+) (Eq. 1)) detection on the test set. Comparison between ConfLUNet and 3D U-Net semantic segmentation baseline model + Connected Components (CC) or Automated Confluent Lesion Splitting (ACLS) as post-processing.

Method	$F1^{CLU}$ (%)	$Recall^{CLU}$ (%)	$Precision^{CLU}$ (%)	$F1^{CLU+}$ (%)	$Recall^{CLU+}$ (%)	$Precision^{CLU+}$ (%)
3D U-Net + CC	58.6	41.5	**100.0**	66.7	50.0	**100.0**
3D U-Net + ACLS	76.8	**92.7**	65.5	82.2	**90.9**	75.0
ConfLUNet (ours)	**84.2**	78.1	91.4	**88.7**	83.3	94.8

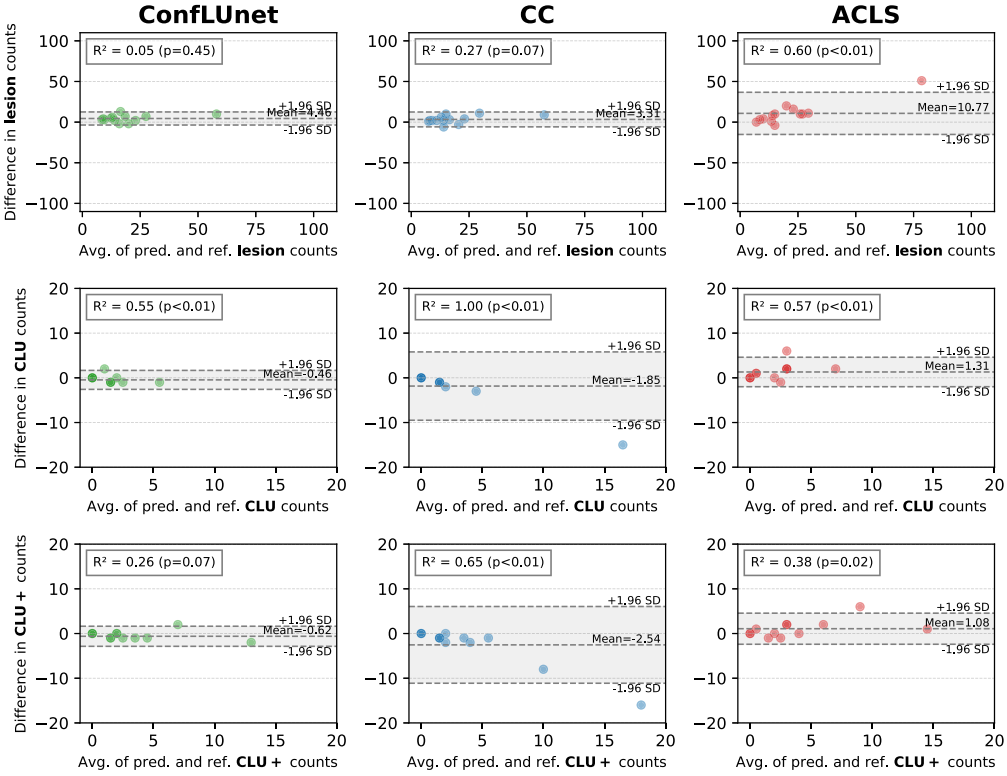

Fig. 8. Bland–Altman plot for predicted and reference lesion/CLU/CLU+ counts. Comparison between ConfLUNet and 3D U-Net semantic segmentation baseline model + Connected Components (CC) or Automated Confluent Lesion Splitting (ACLS) as post-processing. CLU: Confluent Lesion Unit; CLU+: Extended definition of CLUs.

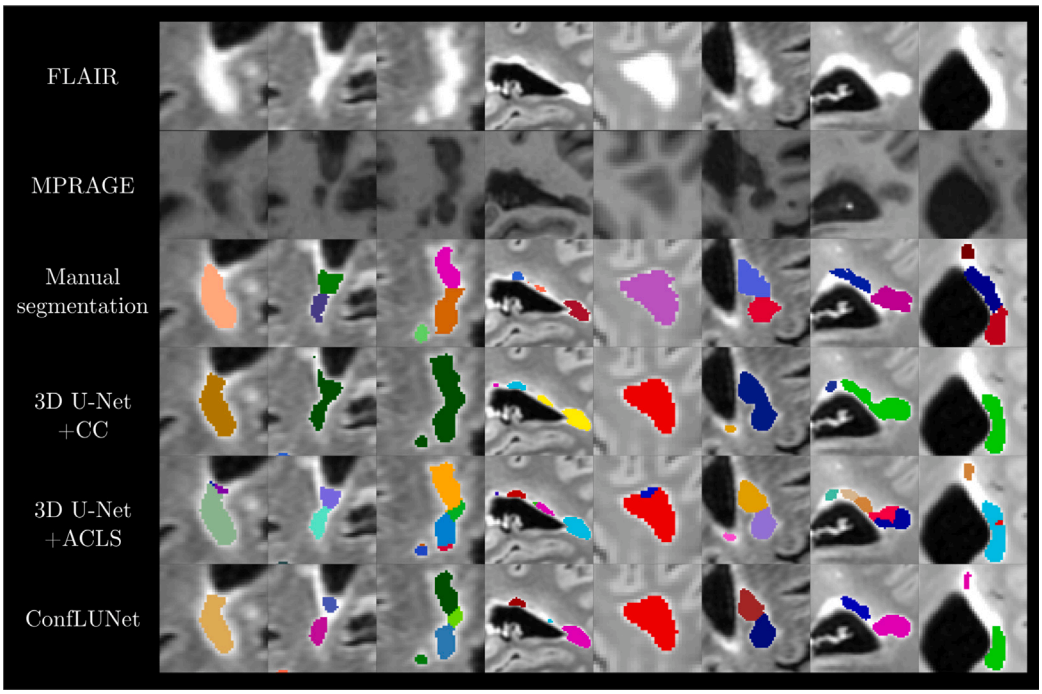

Fig. 9. Illustrative examples of predictions made by connected components (CC), automated confluent splitting (ACLS) and ConfLUNet on the test set.

6. Discussion

Call for an evaluation framework adapted to lesion instance segmentation in MS. Despite the current need for lesion-level delineation in MS, both for clinical [2,3] and technical [15,16,44–46] applications, current evaluation practices remain poorly aligned with the goals of instance segmentation. Most reported detection metrics are derived from semantic masks processed with CC, which lack the granularity needed to assess individual lesions within confluent regions. Although confluent lesions have been clinically described [17,18], no prior work has formally defined them or the distinct units they contain, making their systematic evaluation particularly difficult. To address these limitations, our work introduces formal definitions and proposes CLU-aware metrics grounded in mathematical definitions. By incorporating these metrics alongside instance segmentation metrics, the proposed framework moves toward evaluation practices that are more consistent with clinical and technical needs. In parallel, progress in developing new instance segmentation methods is also hindered by the lack of public benchmarks. This is likely related to the complexity and time required for manual segmentation, which can be infeasible in advanced disease stages or when image quality is insufficient. To drive progress in this area, the community must prioritize the development of reliable public benchmarks and adopt evaluation protocols that reflect the clinical and technical demands of instance-level lesion analysis.

CC underestimates CLU count and lacks sensitivity. The results from Experiments 1, 2 and 3 demonstrate that frequently adopted CC strategy systematically underestimates CLU counts, particularly in cases with higher amounts of confluent lesions. This limitation stems from CC's inherent inability to distinguish spatially connected and/or overlapping lesions, as it can only detect CLUs when lesions are perfectly separated. Whether this spatial separation is correctly done to this purpose is heavily dependent on the predicted lesion semantic segmentation mask. To illustrate this difficulty, none of the three semantic segmentation tools (LST-LPA, SAMSEG, FLAMeS) evaluated in this study produced adequately disconnected instances, as evidenced by CC's CLU recall dropping below 50% across all tools — yielding $\text{Recall}^{\text{CLU}}$ scores of 41.2%, 30.6%, and 36.2% for LST-LPA, SAMSEG, and FLAMeS, respectively — when restricted to patients with confluent lesions. While CC achieved extraordinarily high precision in CLU and CLU+ detection across all experiments, this largely reflects its conservative nature, which limits oversplitting and aligns well with how these metrics penalize false positives. Experiment 3 further confirms that these results hold true even when the model was trained on qualitatively conservative annotations. Altogether, these findings suggest that CC, despite its widespread use in current methods and evaluation frameworks (Suppl. Appendix E), is ill-suited for accurate lesion instance segmentation in the presence of confluent lesions and should be reconsidered as a default post-processing choice.

ACLS overestimates lesion count and oversplits lesions. Originally designed to split confluent lesions, our study systematically evaluate ACLS within an instance segmentation framework for the first time, with a particular focus on CLU detection—addressing the original method's limitation of only considering global lesion counts. ACLS consistently and significantly outperforms CC in instance segmentation (PQ: +1.6%) and shows markedly higher recall for both overall lesion detection (+17.7%) and CLU-wise detection ($\text{Recall}^{\text{CLU}}$: +20.7%; $\text{Recall}^{\text{CLU+}}$: +25.8%). However, this increased sensitivity comes at a cost: ACLS systematically overestimates total lesion counts (bias: 73.3, Fig. 6) and exhibits low precision across all segmentation tools (47.5%, 50.4%, and 53.9% for LST-LPA, SAMSEG, and FLAMeS, respectively), with approximately half of the predicted lesions being false positives. These findings, corroborated by visual inspection, suggest that ACLS frequently oversplits lesions. This behavior and its excessive sensitivity exhibited provides evidence against ACLS's original assumption, stating local maxima in probability maps correspond to true lesion centers. Interestingly, results from Experiment 1 show virtually the same effects in applying ACLS to either LST-LPA, SAMSEG or FLAMeS, which indicates that ACLS can be applied regardless to segmentation methods with different methodological roots.

This study could not replicate the performances reported in Dworkin et al. [20], especially regarding the correlation between predicted and reference lesion count. Methodological choices might explain these differences, such as chosen connectivity structure or probability threshold—more conservative in Dworkin et al.'s study (0.3), provoking CC to create larger connected regions and thus underperform.

ConfLUNet represents a balanced precision–recall alternative. The proposed ConfLUNet is the first end-to-end framework for instance segmentation of MS lesions with a particular focus on improving detection of CLUs. Our results indicate that ConfLUNet achieves significantly higher instance segmentation performance than both CC and ACLS (resp. +4.5% and +5.2% on the test set). Compared to CC, which favors Precision, and ACLS, which prioritizes Recall, ConfLUNet achieves a more balanced lesion-wise and CLU-wise detection performance, combining excellent precision with high sensitivity (F1: +5.7%/+7.4%; $F1^{CLU}$: +1.7%/+18.1%; $F1^{CLU+}$: +10.7%/+17.4%). As further indicated by Fig. 8, ConfLUNet mitigates both CC's undercounting bias and ACLS's oversplitting tendency.

Limitations and future work. Several limitations should be noted. Our evaluation used a relatively small single-center dataset with limited cases with confluent lesions (35/63), potentially affecting statistical power and generalizability. Furthermore, among these cases, the maximum number of CLU per confluent lesion was lower ($\leq$ 4) than the dozens reported by Dworkin and colleagues [20]. Future validation should assess performance on cases with higher confluent lesion loads, though we suspect annotation complexity — and thus annotation noise — to increase with CLU density. Finally, the single-rater annotations may introduce bias, and the method's performance on multi-contrast or out-of-domain data remains untested. Despite these limitations, this work remains valuable by being the first to formally assess the shortcomings of existing methods for confluent lesion detection and by offering a comparative evaluation of established approaches (Connected Components and ACLS) against a novel strategy (ConfLUNet). Moreover, we believe that ConfLUNet constitutes a pioneering step toward end-to-end instance segmentation of MS lesions, thereby addressing confluent lesions, and lays the groundwork for broader adoption and future development in this domain. Future work includes annotating larger multi-centric datasets to extend our analysis and assess generalizability of all investigated methods, and the incorporation of test-time training to enhance ConfLUNet's generalizability [47].

7. Conclusion

Current evaluation practices for MS lesion segmentation remain misaligned with clinical needs, particularly in their inability to assess individual lesions within confluent regions. To address this gap, we introduced a formal framework for evaluating lesion instance segmentation, and proposed new CLU-aware metrics to assess their detection, grounded in precise definitions of confluent lesions and their constituent units (CLUs). While existing methods like CC and ACLS exhibit opposite failure modes — merging versus oversplitting — our end-to-end approach, ConfLUNet, jointly optimizes detection and delineation, providing a promising direction for more accurate and clinically meaningful lesion-level analysis in MS. Further validation on larger and more heterogeneous multi-centric cohorts is warranted.

CRediT authorship contribution statement

Maxence Wynen: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Pedro M. Gordaliza:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis. **Maxime Istasse:** Writing – review & editing, Supervision, Software, Methodology, Investigation, Conceptualization. **Anna Stölting:** Writing – review & editing, Visualization, Data curation. **Pietro Maggi:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Benoît Macq:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Meritxell Bach Cuadra:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization.

Compliance with ethical standards

All procedures performed in this study were in accordance with the ethical standards of the Belgian national research committee (n° B4032020000104). Written informed consent was obtained for experimentation with human subjects following the World Medical Association Declaration of Helsinki.

Funding

MW was funded by the Walloon region under grant No. 2010235 (ARIAC by digitalwallonia4.ai), and by the Swiss Government Excellence Scholarship (No. 2021.0087).

PMG acknowledges the CIBM Center for Biomedical Imaging, a Swiss research center of excellence founded and supported by Lausanne University Hospital (CHUV), Switzerland, University of Lausanne (UNIL), Switzerland, École polytechnique fédérale de Lausanne (EPFL), University of Geneva (UNIGE) and Geneva University Hospitals (HUG), Switzerland.

AS has the financial support of the Fédération Wallonie Bruxelles – FRIA du Fonds de la Recherche Scientifique – FNRS, Belgium.

PM research activity is supported by the Fondation Charcot Stichting Research Fund 2023, the Fund for Scientific Research (F.S.R, FNRS; grant 40008331), Cliniques universitaires Saint-Luc "Fonds de Recherche Clinique" and Biogen.

BM is funded by Université Catholique de Louvain, Belgium.

MBC acknowledges the CIBM Center for Biomedical Imaging, a Swiss research center of excellence founded and supported by Lausanne University Hospital (CHUV), Switzerland, University of Lausanne (UNIL), Switzerland, École Polytechnique Fédérale de Lausanne (EPFL), University of Geneva (UNIGE) and Geneva University Hospitals (HUG), Switzerland.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Pietro Maggi reports financial support was provided by Biogen Inc. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The present research benefited from computational resources made available on Lucia, the Tier-1 supercomputer of the Walloon Region, infrastructure funded by the Walloon Region under the grant agreement n° 1910247. The authors thank the study participants; the Radiology departments of the Cliniques universitaires Saint-Luc (UCLouvain, Brussels, Belgium); Colin Vanden Bulcke (Université Catholique de Louvain); Laurence Dricot (Université catholique de Louvain); Serena Borrelli (Université catholique de Louvain); François Guisset (Université catholique de Louvain); Benoît Gérin (Université catholique de Louvain); Karim El Khoury (Université catholique de Louvain) and Maxime Zanella (Université catholique de Louvain) for their precious help.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.110919.

References

[1] D.S. Reich, C.F. Lucchinetti, P.A. Calabresi, Multiple sclerosis, New Engl. J. Med. 378 (2) (2018) 169–180, http://dx.doi.org/10.1056/NEJMra1401483.

[2] A.J. Thompson, et al., Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria, Lancet. Neurol. 17 (2) (2018) 162–173, http://dx.doi.org/10.1016/S1474-4422(17)30470-2.

[3] on behalf of the MAGNIMS study group, MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis—establishing disease prognosis and monitoring patients, Nat. Rev. Neurol. 11 (10) (2015) 597–606, http://dx.doi.org/10.1038/nrneurol.2015.157, URL https://www.nature.com/articles/nrneurol.2015.157.

[4] P.A. Brex, O. Ciccarelli, J.I. O'Riordan, M. Sailer, A.J. Thompson, D.H. Miller, A longitudinal study of abnormalities on MRI and disability from multiple sclerosis, New Engl. J. Med. 346 (3) (2002) 158–164, http://dx.doi.org/10.1056/NEJMoa011341.

[5] S.J. Khoury, C.R. Guttmann, E.J. Orav, M.J. Hohol, S.S. Ahn, L. Hsu, R. Kikinis, G.A. Mackin, F.A. Jolesz, H.L. Weiner, Longitudinal MRI in multiple sclerosis: correlation between disability and lesion burden, Neurology 44 (11) (1994) 2120–2124, http://dx.doi.org/10.1212/wnl.44.11.2120.

[6] R.A. Rudick, J.-C. Lee, J. Simon, E. Fisher, Significance of T2 lesions in multiple sclerosis: A 13-year longitudinal study, Ann. Neurol. 60 (2) (2006) 236–242, http://dx.doi.org/10.1002/ana.20883.

[7] X. Montalban, Revised mcdonald criteria 2023, in: 2024 40th Congress of the European Committe for Treatment and Research in Multiple Sclerosis, ECTRIMS 2024, 2024.

[8] P. Maggi, M. Absinta, M. Grammatico, L. Vuolo, G. Emmi, G. Carlucci, G. Spagni, A. Barilaro, A.M. Repice, L. Emmi, D. Prisco, V. Martinelli, R. Scotti, N. Sadeghi, G. Perrotta, P. Sati, B. Dachy, D.S. Reich, M. Filippi, L. Massacesi, Central vein sign differentiates multiple sclerosis from central nervous system inflammatory vasculopathies, Ann. Neurol. 83 (2) (2018) 283–294, http://dx.doi.org/10.1002/ana.25146.

[9] M. Absinta, P. Sati, F. Masuzzo, G. Nair, V. Sethi, H. Kolb, J. Ohayon, T. Wu, I.C.M. Cortese, D.S. Reich, Association of chronic active multiple sclerosis lesions with disability in vivo, JAMA Neurol. 76 (12) (2019) 1474–1483, http://dx.doi.org/10.1001/jamaneurol.2019.2399.

[10] C. Elliott, J.S. Wolinsky, S.L. Hauser, L. Kappos, F. Barkhof, C. Bernasconi, W. Wei, S. Belachew, D.L. Arnold, Slowly expanding/evolving lesions as a magnetic resonance imaging marker of chronic active multiple sclerosis lesions, Mult. Scler. (Houndmills, Basingstoke, England) 25 (14) (2019) 1915–1925, http://dx.doi.org/10.1177/1352458518814117.

[11] F. La Rosa, M. Wynen, O. Al-Louzi, E.S. Beck, T. Huelnhagen, P. Maggi, J.-P. Thiran, T. Kober, R.T. Shinohara, P. Sati, et al., Cortical lesions, central vein sign, and paramagnetic rim lesions in multiple sclerosis: Emerging machine learning techniques and future avenues, NeuroImage: Clin. 36 (2022) 103205, Publisher: Elsevier.

[12] G. Barquero, et al., RimNet: A deep 3D multimodal MRI architecture for paramagnetic rim lesion assessment in multiple sclerosis, NeuroImage: Clin. 28 (2020) 102412, http://dx.doi.org/10.1016/j.nicl.2020.102412, URL http://www.sciencedirect.com/science/article/pii/S2213158220302497.

[13] P. Maggi, M.J. Fartaria, J. Jorge, F.L. Rosa, M. Absinta, P. Sati, R. Meuli, R.D. Pasquier, D.S. Reich, M.B. Cuadra, C. Granziera, J. Richiardi, T. Kober, Cvsnet: A machine learning approach for automated central vein sign assessment in multiple sclerosis, NMR Biomed. 33 (5) (2020) e4283, http://dx.doi.org/10.1002/nbm.4283, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nbm.4283, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nbm.4283.

[14] C. Lou, P. Sati, M. Absinta, K. Clark, J.D. Dworkin, A.M. Valcarcel, M.K. Schindler, D.S. Reich, E.M. Sweeney, R.T. Shinohara, Fully automated detection of paramagnetic rims in multiple sclerosis lesions on 3T susceptibility-based MR imaging, NeuroImage. Clin. 32 (2021) 102796, http://dx.doi.org/10.1016/j.nicl.2021.102796.

[15] H. Zhang, T.D. Nguyen, J. Zhang, M. Marcille, P. Spincemaille, Y. Wang, S.A. Gauthier, E.M. Sweeney, Qsmrim-net: Imbalance-aware learning for identification of chronic active multiple sclerosis lesions on quantitative susceptibility maps, NeuroImage : Clin. 34 (2022) 102979, http://dx.doi.org/10.1016/j.nicl.2022.102979, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8892132/.

[16] M. Wynen, F. La Rosa, A. Sellimi, G. Barquero, G. Perrotta, V. Lolli, V. Van Pesch, C. Granziera, T. Kober, P. Sati, et al., Longitudinal automated assessment of paramagnetic rim lesions in multiple sclerosis using RimNet, in: ISMRM, 2021.

[17] R. Zivadinov, M. Zorzon, R. De Masi, D. Nasuelli, G. Cazzato, Effect of intravenous methylprednisolone on the number, size and confluence of plaques in relapsing–remitting multiple sclerosis, J. Neurol. Sci. 267 (1) (2008) 28–35, http://dx.doi.org/10.1016/j.jns.2007.09.025, URL https://www.sciencedirect.com/science/article/pii/S0022510X07006557.

[18] H. Lassmann, Multiple sclerosis: Lessons from molecular neuropathology, Special Issue: Progress in MS pathophysiology and treatment, Exp. Neurol. 262 (2014) 2–7, http://dx.doi.org/10.1016/j.expneurol.2013.12.003, URL https://www.sciencedirect.com/science/article/pii/S0014488613003610.

[19] J.O. Harris, J.A. Frank, N. Patronas, D.E. McFarlin, H.F. McFarland, Serial gadolinium-enhanced magnetic resonance imaging scans in patients with early, relapsing-remitting multiple sclerosis: implications for clinical trials and natural history, Ann. Neurol. 29 (5) (1991) 548–555, http://dx.doi.org/10.1002/ana.410290515.

[20] J.D. Dworkin, et al., An automated statistical technique for counting distinct multiple sclerosis lesions, Am. J. Neuroradiol. 39 (4) (2018) 626–633, http://dx.doi.org/10.3174/ajnr.A5556, URL http://www.ajnr.org/content/39/4/626, Publisher: American Journal of Neuroradiology Section: ADULT BRAIN.

[21] M. Wynen, M. Istasse, P.M. Gordaliza, A. Stölting, P. Maggi, B. Macq, M.B. Cuadra, Conflunet: Improving confluent lesion identification in multiple sclerosis with instance segmentation, in: 2024 IEEE International Symposium on Biomedical Imaging, ISBI, 2024, pp. 1–5, http://dx.doi.org/10.1109/ISBI56570.2024.10635707, URL https://ieeexplore.ieee.org/document/10635707, ISSN: 1945-8452.

[22] A.M. Hafiz, G.M. Bhat, A survey on instance segmentation: state of the art, Int. J. Multimed. Inf. Retr. 9 (3) (2020) 171–189, http://dx.doi.org/10.1007/s13735-020-00195-x.

[23] W. Gu, S. Bai, L. Kong, A review on 2D instance segmentation based on deep neural networks, Image Vis. Comput. 120 (2022) 104401, http://dx.doi.org/10.1016/j.imavis.2022.104401, URL https://www.sciencedirect.com/science/article/pii/S0262885622000300.

[24] B. Cheng, Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation, 2020, URL http://arxiv.org/abs/1911.10194, arXiv:1911.10194 [cs].

[25] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, 2021, ArXiv:2111.06377 [Cs], URL http://arxiv.org/abs/2111.06377, arXiv: 2111.06377 version: 1.

[26] E.S. Nasir, A. Parvaiz, M.M. Fraz, Nuclei and glands instance segmentation in histology images: a narrative review, Artif. Intell. Rev. 56 (8) (2023) 7909–7964, http://dx.doi.org/10.1007/s10462-022-10372-5.

[27] F. Isensee, P.F. Jaeger, S.A.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2) (2021) 203–211, http://dx.doi.org/10.1038/s41592-020-01008-z, URL https://www.nature.com/articles/s41592-020-01008-z, Number: 2 Publisher: Nature Publishing Group.

[28] P. Bilic, P. Christ, H.B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G.E.H. Mamani, G. Chartrand, F. Lohöfer, J.W. Holch, W. Sommer, F. Hofmann, A. Hostettler, N. Lev-Cohain, M. Drozdzal, M.M. Amitai, R. Vivanti, J. Sosna, I. Ezhov, A. Sekuboyina, F. Navarro, F. Kofler, J.C. Paetzold, S. Shit, X. Hu, J. Lipková, M. Rempfler, M. Piraud, J. Kirschke, B. Wiestler, Z. Zhang, C. Hülsemeyer, M. Beetz, F. Ettlinger, M. Antonelli, W. Bae, M. Bellver, L. Bi, H. Chen, G. Chlebus, E.B. Dam, Q. Dou, C.-W. Fu, B. Georgescu, X. Giró-i Nieto, F. Gruen, X. Han, P.-A. Heng, J. Hesser, J.H. Moltz, C. Igel, F. Isensee, P. Jäger, F. Jia, K.C. Kaluva, M. Khened, I. Kim, J.-H. Kim, S. Kim, S. Kohl, T. Konopczynski, A. Kori, G. Krishnamurthi, F. Li, H. Li, J. Li, X. Li, J. Lowengrub, J. Ma, K. Maier-Hein, K.-K. Maninis, H. Meine, D. Merhof, A. Pai, M. Perslev, J. Petersen, J. Pont-Tuset, J. Qi, X. Qi, O. Rippel, K. Roth, I. Sarasua, A. Schenk, Z. Shen, J. Torres, C. Wachinger, C. Wang, L. Weninger, J. Wu, D. Xu, X. Yang, S.C.-H. Yu, Y. Yuan, M. Yue, L. Zhang, J. Cardoso, S. Bakas, R. Braren, V. Heinemann, C. Pal, A. Tang, S. Kadoury, L. Soler, B. van Ginneken, H. Greenspan, L. Joskowicz, B. Menze, The liver tumor segmentation benchmark (LiTS), Med. Image Anal. 84 (2023) 102680, http://dx.doi.org/10.1016/j.media.2022.102680, URL https://www.sciencedirect.com/science/article/pii/S1361841522003085.

[29] M. Wynen, P.M. Gordaliza, A. Stölting, P. Maggi, M. Bach Cuadra, Lesion instance segmentation in multiple sclerosis: Assessing the efficacy of statistical lesion splitting, in: ISMRM, ISMRM, 2024, URL https://dial.uclouvain.be/pr/boreal/object/boreal:285623.

[30] A. Malinin, et al., Shifts 2.0: Extending the dataset of real distributional shifts, 2022, arXiv:2206.15407 [cs, stat], URL http://arxiv.org/abs/2206.15407.

[31] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017, http://dx.doi.org/10.48550/arXiv.1412.6980, arXiv:1412.6980 [cs], URL http://arxiv.org/abs/1412.6980.

[32] S. Grahl, V. Pongratz, P. Schmidt, C. Engl, M. Bussas, A. Radetz, G. Gonzalez-Escamilla, S. Groppa, F. Zipp, C. Lukas, J. Kirschke, C. Zimmer, M. Hoshi, A. Berthele, B. Hemmer, M. Mühlau, Evidence for a white matter lesion size threshold to support the diagnosis of relapsing remitting multiple sclerosis, Mult. Scler. Relat. Disord. 29 (2019) 124–129, http://dx.doi.org/10.1016/j.msard.2019.01.042.

[33] B.B. Avants, C.L. Epstein, M. Grossman, J.C. Gee, Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain, Med. Image Anal. 12 (1) (2008) 26–41, http://dx.doi.org/10.1016/j.media.2007.06.004.

[34] A. Hoopes, J.S. Mora, A.V. Dalca, B. Fischl, M. Hoffmann, SynthStrip: skull-stripping for any brain image, NeuroImage 260 (2022) 119474, http://dx.doi.org/10.1016/j.neuroimage.2022.119474, URL https://www.sciencedirect.com/science/article/pii/S1053811922005900.

[35] C. Vanden Bulcke, M. Wynen, J. Detobel, F. La Rosa, M. Absinta, L. Dricot, B. Macq, M.B. Cuadra, P. Maggi, BMAT: An open-source BIDS managing and analysis tool, NeuroImage: Clin. 36 (2022) 103252, Publisher: Elsevier.

[36] P.A. Yushkevich, Y. Gao, G. Gerig, ITK-SNAP: an interactive tool for semi-automatic segmentation of multi-modality biomedical images, in: Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society, in: IEEE Engineering in Medicine and Biology Society. Annual Conference, vol. 2016, 2016, pp. 3342–3345, http://dx.doi.org/10.1109/EMBC.2016.7591443, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5493443/.

[37] L. Maier-Hein, A. Reinke, P. Godau, M.D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, M. Reyes, M.A. Riegler, M. Wiesenfarth, A.E. Kavur, C.H. Sudre, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, T. Rädsch, L. Acion, M. Antonelli, T. Arbel, S. Bakas, A. Benis, M.B. Blaschko, M.J. Cardoso, V. Cheplygina, B.A. Cimini, G.S. Collins, K. Farahani, L. Ferrer, A. Galdran, B. van Ginneken, R. Haase, D.A. Hashimoto, M.M. Hoffman, M. Huisman, P. Jannin, C.E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, A. Karthikesalingam, F. Kofler, A. Kopp-Schneider, A. Kreshuk, T. Kurc, B.A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A.L. Martel, P. Mattson, E. Meijering, B. Menze, K.G.M. Moons, H. Müller, B. Nichyporuk, F. Nickel, J. Petersen, N. Rajpoot, N. Rieke, J. Saez-Rodriguez, C.I. Sánchez, S. Shetty, M. van Smeden, R.M. Summers, A.A. Taha, A. Tiulpin, S.A. Tsaftaris, B. Van Calster, G. Varoquaux, P.F. Jäger, Metrics reloaded: recommendations for image analysis validation, Nature Methods 21 (2) (2024) 195–212, http://dx.doi.org/10.1038/s41592-023-02151-z, URL https://www.nature.com/articles/s41592-023-02151-z, Publisher: Nature Publishing Group.

[38] V. Raina, N. Molchanova, M. Graziani, A. Malinin, H. Muller, M.B. Cuadra, M. Gales, Tackling bias in the dice similarity coefficient: Introducing NDSC for white matter lesion segmentation, in: 2023 IEEE 20th International Symposium on Biomedical Imaging, ISBI, 2023, pp. 1–5, http://dx.doi.org/10.1109/ISBI53787.2023.10230755, URL https://ieeexplore.ieee.org/document/10230755/, ISSN: 1945-8452.

[39] P. Schmidt, Bayesian Inference for Structured Additive Regression Models for Large-Scale Problems with Applications to Medical Imaging [Ph.D. thesis], Ludwig-Maximilians-Universität München, 2017, http://dx.doi.org/10.5282/EDOC.20373, URL https://edoc.ub.uni-muenchen.de/id/eprint/20373, Medium: application/pdf.

[40] E. Dereskewicz, F.L. Rosa, J.d.S. Silva, E. Sizer, A. Kohli, M. Wynen, W.A. Mullins, P. Maggi, S. Levy, K. Onyemeh, B. Ayci, A.J. Solomon, J. Assländer, O. Al-Louzi, D.S. Reich, J. Sumowski, E.S. Beck, FLAMeS: A robust deep learning model for automated multiple sclerosis lesion segmentation, 2025, http://dx.doi.org/10.1101/2025.05.19.25327707, URL https://www.medrxiv.org/content/10.1101/2025.05.19.25327707v1, Pages: 2025.05.19.25327707.

[41] F. La Rosa, J. Dos Santos Silva, W.A. Mullins, H. Greenspan, J. Sumowski, D.S. Reich, E.S. Beck, FLAMeS: FLAIR lesion analysis in multiple sclerosis, 2023, http://dx.doi.org/10.5281/zenodo.7626121, URL https://zenodo.org/records/7626121.

[42] S. Cerri, O. Puonti, D.S. Meier, J. Wuerfel, M. Mühlau, H.R. Siebner, K. Van Leemput, A contrast-adaptive method for simultaneous whole-brain and lesion segmentation in multiple sclerosis, NeuroImage 225 (2021) 117471, http://dx.doi.org/10.1016/j.neuroimage.2020.117471, URL https://www.sciencedirect.com/science/article/pii/S1053811920309563.

[43] Scipy.org — Scipy.org, 2025, URL https://www.scipy.org/.

[44] M. Wynen, P.M. Gordaliza, A. Stölting, P. Maggi, M.B. Cuadra, B. Macq, Multimodal segmentation for paramagnetic rim lesion detection in multiple sclerosis, in: Medical Imaging 2024: Imaging Informatics for Healthcare, Research, and Applications, vol. 12931, SPIE, 2024, pp. 241–246.

[45] P. Maggi, C. Vanden Bulcke, E. Pedrini, C. Bugli, A. Sellimi, M. Wynen, A. Stölting, W.A. Mullins, G. Kalaitzidis, V. Lolli, et al., B cell depletion therapy does not resolve chronic active multiple sclerosis lesions, EBioMedicine 94 (2023) Publisher: Elsevier.

[46] M. Wynen, C. Vanden Bulcke, S. Borrelli, P.M. Gordaliza, A. Stölting, F. Guisset, C. Cordier, M.S. Martire, A. Tamanti, B. Macq, et al., Machine learning-based combination of the central vein sign, cortical lesions and paramagnetic rim lesions: A web-based tool for the diagnosis of multiple sclerosis, 2025, http://dx.doi.org/10.2139/ssrn.5192817.

[47] B. Gérin, M. Zanella, M. Wynen, S. Mahmoudi, B. Macq, C. De Vleeschouwer, Exploring viability of test-time training: Application to 3D segmentation in multiple sclerosis, in: 2024 IEEE Conference on Artificial Intelligence (CAI), IEEE, 2024, pp. 557–562.