Pattern Recognition Letters 200 (2026) 115–122

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Towards robust and reliable multi-modal 3D segmentation of multiple sclerosis lesions

Edoardo Coppola [a,*], Mattia Savardi [b], Alberto Signoroni [b]

[a] Dept. of Information Engineering, University of Brescia, Via Branze, 38, Brescia, 25123, Italy
[b] Dept. of Medical and Surgical specialties, Radiological Sciences and Public Health, University of Brescia, Viale Europa, 11, Brescia, 25123, Italy

ARTICLE INFO

Editor: Maria De Marsico

Keywords:
Multiple sclerosis
Magnetic resonance imaging
Multi-modal segmentation
Domain adaptation
Reliability estimation

ABSTRACT

Accurate 3D segmentation of multiple sclerosis lesions is critical for clinical practice, yet existing approaches face key limitations: many models rely on 2D architectures or partial modality combinations, while others struggle to generalise across scanners and protocols. Although large-scale, multi-site training can improve robustness, its data demands are often prohibitive. To address these challenges, we propose a 3D multi-modal network that simultaneously processes T1-weighted, T2-weighted, and FLAIR scans, leveraging full cross-modal interactions and volumetric context to achieve state-of-the-art performance across four diverse public datasets. To tackle data scarcity, we quantify the *minimal* fine-tuning effort needed to adapt to individual unseen datasets and reformulate the few-shot learning paradigm at an "instance-per-dataset" level (rather than traditional "instance-per-class"), enabling the quantification of the *minimal* fine-tuning effort to adapt to *multiple* unseen sources simultaneously. Finally, we introduce *Latent Distance Analysis*, a novel label-free reliability estimation technique that anticipates potential distribution shifts and supports any form of test-time adaptation, thereby strengthening efficient robustness and physicians' trust.

1. Introduction

Multiple Sclerosis (MS) is a chronic autoimmune disease that damages the myelin sheath in the central nervous system, leading to a range of neurological symptoms that often worsen over time and significantly impact quality of life [1]. Magnetic Resonance Imaging (MRI) is the gold standard for the diagnosis and monitoring of MS, as it enables neurologists to visualize inflammatory lesions that reflect the disease activity and the response to treatment [2]. While other methods, such as Optical Coherence Tomography (OCT), are increasingly integrated into the clinical management of MS [3], they do not provide direct access to these lesions. For this purpose, multiple MRI sequences are used: T1-weighted (T1w), T2-weighted (T2w) and Fluid-Attenuated Inversion Recovery (FLAIR). FLAIR scans, in particular, are the most sensitive to MS lesions, which appear as hyperintense regions whose interpretation requires careful scrutiny. Accurate lesion delineation on MRI volumes is invaluable for both monitoring and treatment planning. However, manual delineation, still common in clinical workflows, is time-consuming and inconsistent between raters [4]. This inconsistency arises from the natural variability of MS lesions - ambiguous borders, diverse sizes/shapes, similarity to healthy structures - underscoring the critical need for robust computer-aided segmentation. Research in this area spans from semi-automated tools [5,6], limited by scalability and generalisability [7], to deep learning approaches facing data-centric limitations: scarcity of large, multi-modal MRI datasets [8], inconsistencies from differing annotation protocols across these often modest collections [9], and significant inter-site variability [1]. Consequently, many approaches are restricted to single-modality inputs or 2D architectures, sacrificing vital cross-modality and volumetric information. This limitation ultimately leads to suboptimal performance when handling the heterogeneity of MS lesions and data. To improve robustness, some have trained models on multicentre datasets [8], while others adapted models to new distributions via fine-tuning [10,11], domain adaptation [12], or fully-test-time training [1]. However, these strategies present fundamental downsides. Multicentre approaches assume access to relatively large amounts of labelled data in the target domain, which is often infeasible in clinical practice. Conversely, adaptation methods risk sacrificing global generalisability by specialising the model to a single new site. This dilemma is exacerbated in the context of MS, where datasets are limited in both size and availability.

The MRI heterogeneity issues are an instance of the broader problem of distribution shift [13], often framed by the distinction between in-distribution and out-of-distribution (OOD) samples [14–16]. However, to our knowledge, no prior work has addressed how to estimate the

* Corresponding author.
E-mail addresses: edoardo.coppola@unibs.it (E. Coppola), mattia.savardi@unibs.it (M. Savardi), alberto.signoroni@unibs.it (A. Signoroni).

https://doi.org/10.1016/j.patrec.2025.12.008
Received 18 June 2025; Received in revised form 14 November 2025; Accepted 20 December 2025
Available online 24 December 2025

reliability of segmentation predictions before or after adaptation. This is a key need to build clinical trust.

To address these challenges, we present a study of cross-domain MS lesion segmentation under minimal supervision. Unlike many previous works, our approach leverages a 3D full-volume architecture with complete multi-modal input (T1w, T2w, FLAIR). We focus on three main goals: (i) quantifying how much labelled data is needed to adapt a model to multiple, unseen sites, (ii) exploring cross-domain generalisation via a revised few-shot learning paradigm, where t instances are drawn from each target dataset, and (iii) proposing a novel, online, differentiable method for test-time reliability estimation based on latent space analysis. Our main contributions are:

- **Minimal fine-tuning quantification**: we empirically quantify the *minimal* fine-tuning effort to improve performance on multiple, diverse MS datasets, offering insights into adaptation feasibility under real-world annotation constraints;
- **Few-shot cross-dataset adaptation under multiple shifts**: we propose and evaluate a realistic few-shot adaptation paradigm where a minimal number of labelled instances is sampled from each target dataset, enabling joint adaptation across multiple domains with significant distributional shifts;
- **Latent Distance Analysis (LDA)**: we introduce an online, differentiable, test-time reliability estimation technique based on latent feature statistics, enabling uncertainty-aware deployment without requiring additional annotations.

The remainder of this paper is organised as follows. Section 2 reviews related work in MS lesion segmentation, model transfer under distribution shift, and out-of-distribution detection. Section 3 details our proposed methodology. Section 4 presents the experimental results and discusses the limitations of our approach. Finally, Section 5 summarises our findings and outlines the directions for future work.

2. Related work

Segmentation of Multiple Sclerosis Lesions. Early MS lesion segmentation methods relied on atlas-based [5] or semi-automated pipelines [6] with limited generalisability and scalability [7]. Deep learning approaches introduced more scalable solutions, starting with 2D single-modality models [17,18], followed by multi-view CNNs approximating 3D context [19]. Bi- and tri-modal setups have been explored [20–22], though often constrained to 2D data processing. Patch-based 3D architectures improved spatial coherence [23,24], and models such as LST-AI [2] and dual-path 3D CNNs [25,26] added bi-modal capacity. Still, most works either process sequences of small 3D patches [27–29], instead of full volumes, or leverage partial modality combinations. Our work combines full-volume modelling with tri-modal input (T1w, T2w, FLAIR) for maximal lesion visibility and structural context exploitation.

Transferring and Adapting Models under Distribution Shifts. Domain shifts caused by scanner and site variability often degrade performance on unseen domains [9,13]. Adaptation is further hampered by the inherent data scarcity of MS scans. Therefore, building intrinsically robust and easily adaptable models is crucial. [8] aimed at robustness via multi-site training, while [26] and [23] opted for data augmentation and hierarchical modelling. Regarding adaptation to unseen sources, test-time optimisation [1] and few-shot fine-tuning have been proposed. [10] showed that single-instance fine-tuning is effective. Similarly, [11] included a few volumes from longitudinal series to enhance segmentations. Differently, [12] generated target-style inputs trying to bridge domain gaps under annotation constraints. We extend these efforts by quantifying the *minimal* number of target instances necessary to adapt across multiple datasets, simulating practical scenarios with only a few labelled instances per domain available. We also reformulate few-shot learning as instance-per-dataset rather than instance-per-class, enabling scalable multi-domain adaptation under constrained supervision and multiple simultaneous shifts.

Table 1
Our nnUNetV2 3D full-volume architecture for tri-modal (T1w, T2w, FLAIR) input. All `Conv3d` layers have bias enabled and are followed by `InstanceNorm3d` (eps = $1e^{-5}$, affine = True) and `LeakyReLU` activations. The Decoder mirrors the encoder. Heterogeneous inputs are resized to a standard shape of 128^3 and then remapped to the original shape.

Stage	Feature maps	Kernel size	Stride	# Convs
1	32	$3\times3\times3$	$1\times1\times1$	2
2	64	$3\times3\times3$	$2\times2\times2$	2
3	128	$3\times3\times3$	$2\times2\times2$	2
4	256	$3\times3\times3$	$2\times2\times2$	2
5	320	$3\times3\times3$	$2\times2\times2$	2
6	320	$3\times3\times3$	$1\times2\times2$	2
Tri-modal input size		$3\times(128\times128\times128)$		
Total trainable parameters		≈ 30.4 million		

Detecting OOD samples before adaptation. Prior to adaptation, detecting whether new samples are out-of-distribution is crucial. Post hoc methods like Maximum Softmax Probability [15,16] estimate OOD scores from softmax outputs, although often relying on erroneous, overconfident classifiers. To counter this, others leverage latent features [30,31]. Alternatively, Outlier Exposure methods use auxiliary OOD data to penalise confident predictions on unseen distributions [32], but assume access to known and representative OOD samples, which is often unrealistic in medical settings. In contrast, we introduce *Latent Distance Analysis*, a test-time, differentiable, label-free reliability estimator that scores new samples based on their distance to training distribution clusters in latent space.

3. Methods

To experiment with adaptation across datasets and setups, it is first necessary to pre-train a segmentation model that could serve as both experimental baseline and starting point for future adaptations. We detail pre-training in Section 3.1, minimal-instance adaptation on individual target datasets in Section 3.2, and cross-dataset few-shot adaptation in Section 3.3. We eventually describe our LDA technique in Section 3.4.

3.1. Pre-training a segmentation model

To build a baseline MS lesion segmentation model, we pre-train in a supervised manner a 3D, multi-modal version of the nnUNetV2 [33], a description of which is provided in Table 1. When pre-training on an arbitrary labelled dataset D_0, we optimise the sum of the following loss functions:

$$\mathcal{L}_{Dice}(\hat{y}, y) = \lambda_1 \cdot \frac{2\sum(y\cdot\hat{y})+\epsilon}{\sum y^2+\sum\hat{y}^2+\epsilon} \tag{1}$$

$$\mathcal{L}_{BCE}(\hat{y}, y) = \lambda_2 \cdot [y\cdot\log(\hat{y}) + (1-y)\log(1-\hat{y})] \tag{2}$$

where the first function is a soft Dice loss, the second one is a standard binary cross-entropy, λ_1, λ_2 are learnable parameters and $\epsilon = 1\cdot10^{-8}$ is used for numerical stability. Model weights are updated with SGD ($\eta = 0.01$, $\rho = 0.99$, $w = 3\cdot10^{-5}$), using a batch size of 2 volumes for 1000 epochs. To enhance generalisation, we apply a set of independent neurologist-validated data augmentations during training, including vertical flipping and $\pm10°$ rotation, both performed with a probability of 0.2.

3.2. Minimal-instance adaptation

To assess the feasibility and effectiveness of model adaptation under severe data and annotation constraints on new target datasets $D_1, \dots D_M$, we fine-tune the pre-trained model on each dataset individually. For each target source, fine-tuning is performed using 1, 3, or 5

randomly sampled training instances until training loss (Eqs. (1) and (2)) or metrics cease to improve by a minimum of $\Delta = 1 \cdot 10^{-3}$. We do not use any validation set, skip any hyperparameter tuning and utilise the high Δ to avoid overfitting on fine-tuning instances. The adapted model is then evaluated on the precomputed testing partition of the corresponding dataset. As a baseline, we also evaluate the pre-trained model without any adaptation, representing direct transfer across domains. This series of experiments quantifies the *minimal* amount of supervision required to achieve meaningful performance gains under diverse domain shifts in realistic downstream settings.

3.3. Cross-domain few-shot adaptation

In contrast to isolated domain adaptations, we simulate a realistic deployment scenario in which a model must generalise across multiple unseen sites simultaneously with limited supervision from each. To this end, following the protocol of the previous paragraph, we fine-tune the pre-trained nnUNet on pooled sets created by sampling $t = 1, 3, 5, 10$, or 15 instances from each target dataset $(D_1, \dots D_M)$ and merging them into a single adaptation training set of size $t \times M$. This setting revises classical few-shot learning by focusing on distributional diversity rather than class diversity, enabling us to investigate simultaneous cross-domain adaptation with minimal annotation effort. This few-shot adaptation technique is also extensible to an arbitrary number of new datasets, i.e. M can be arbitrary large, by simply sampling t (with $t \in [1, \min_{0 \le i \le M} \{\text{size}(D_i)\}]$) labelled instances from each, merging them, and fine-tuning jointly. Noteworthy, the inclusion of D_0 in the adaptation pool serves to prevent catastrophic forgetting of the original training domain - a factor often overlooked in domain adaptation works, but crucial for practical deployment. After each of these five adaptation experiments, the adapted model is evaluated on the test sets of all participating datasets to measure robustness and generalisation across heterogeneous domains.

3.4. Latent distance analysis

Defining the out-of-distribution score. To estimate test-time prediction reliability under potential distribution shifts, we introduce Latent Distance Analysis, an online, differentiable, geometry-based method for measuring the deviation of individual test samples from the training distribution in latent space. Our approach operates entirely at inference time and does not require target-domain labels nor training examples, allowing fully-test-time adaptation [34]. Moreover, working in latent space enables lightweight adaptation of only the model components (or part of them) that lead to the target latent features. We first extract latent representations from the pre-trained encoder for each training instance. Let $\mathbf{z}_i \in \mathbb{R}^d$ denote the latent feature vector for the i-th training volume. We aggregate all such vectors $\mathbf{z}_i, i = 1 \dots N$ and cluster them into K groups using either: (i) density-based approaches (e.g., DBSCAN), suitable when no a-priori knowledge of acquisition protocols is available; (ii) centroid-based methods (e.g., K-Means), where K is set based on known factors such as the number of sites or scanners, allowing domain-informed partitioning. This results in K clusters, each with centroid $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}^{N \times N}$ estimated from the cluster members. When a new individual test instance x_t is acquired, we compute its latent representation $\mathbf{z}_t$ and the Mahalanobis distance to each cluster centroid:

$$d_{t,k} = \sqrt{(\mathbf{z}_t - \boldsymbol{\mu}_k)^\intercal \boldsymbol{\Sigma}_k^{-1} (\mathbf{z}_t - \boldsymbol{\mu}_k)} \tag{3}$$

To aggregate the relative importance of each cluster, we divide each distance by the cluster cardinality w_k: $\tilde{d}_{t,k} = \frac{d_{t,k}}{w_k}$. We then aggregate the distances using a softmin-ispired function to take into account each cluster's contribution:

$$s_t = -\log\left(\frac{1}{K} \sum_{k=1}^{K} \exp(-\tilde{d}_{t,k})\right) \tag{4}$$

This yields an OOD score $s_t \in \mathbb{R}_0^+$ for the test instance x_t. In particular, scores near 0 indicate strong alignment with the training distribution while scores that deviate from it indicate a completely unfamiliar instance. To better identify strong deviations from cluster typicality, we also present training volume scores. Because real in-distribution samples can vary around these centroids, we do not expect zero OOD scores for training examples. This aligns with the intuition that in-distribution variability should be preserved in uncertainty estimation.

LDA Operational Properties and Sensitivity. Here we present some notable properties of our LDA framework. First, it is fully differentiable, allowing s_t to be directly optimised as a loss function. This opens the door to any kind of test-time adaptation, especially fully-test-time adaptation [34], where only access to test instances and pre-computed cluster statistics is required. Second, in contrast to several OOD detection methods that require batch access, LDA can operate in a fully streaming mode on individual samples. Third, since $\mathbf{z_t}$ is computed by a model encoder, adaptation can be restricted to the encoder weights only, which facilitates optimisation.

Along with operational properties, it is important to acknowledge LDA's sensitivity to two key factors: the number of clusters K and the number of training instances N. To analyse how K impacts the OOD score s_t, we begin by defining the minimum normalised distance $m_t = \min_k(\tilde{d}_{t,k})$ between the test representation $\mathbf{z}_t$ and the centroids $\boldsymbol{\mu}_k$. We can refactor the exponential sum: $\sum_{k=1}^{K} e^{-\tilde{d}_{t,k}} = \sum_{k=1}^{K} e^{-\tilde{d}_{t,k}} \cdot \frac{e^{m_t}}{e^{m_t}} = e^{-m_t} \sum_{k=1}^{K} e^{-(\tilde{d}_{t,k} - m_t)} := e^{-m_t}\mathbf{S}$ where $\mathbf{S} \in [1, K]$ since all exponential terms lie in $[0, 1]$. Then the OOD score becomes: $s_t = -\log\left(\frac{1}{K} e^{-m_t}\mathbf{S}\right) = m_t - \log\left(\frac{\mathbf{S}}{K}\right)$. Given that $\mathbf{S}/K \in [1/K, 1]$, we can derive tight bounds:

$$m_t \le s_t \le m_t + \log(K) \tag{5}$$

Thus, the score is always at least m_t, and at most $\log(K)$ higher. Increasing K cannot reduce s_t below m_t, but can increase its spread. In practice, over-partitioning may inflate scores unnecessarily (small clusters imply small w_k, large $\tilde{d}_{t,k}$). Conversely, if the data is truly multi-cluster, more clusters help keep m_t low by ensuring at least one centroid remains close to $\mathbf{z}_t$. To set K, prior knowledge is helpful, although density-based clustering remains a valid, agnostic, and outlier-robust alternative. Next, we examine the effect of dataset size $N = \sum_{k=1}^{K} w_k$. We introduce cluster proportions $p_k = \frac{w_k}{N}$ and assume they remain fixed as N grows (e.g., the scanners workload does not change). We also reasonably assume there is a stable generative process behind the training observations. Then, after training and fixing encoder weights Θ_{enc}, we can claim that the centroids are substantially stable as N grows. To relate s_t, collecting large amounts of data (i.e., large values of N that push $\tilde{d}_{t,k}$ toward 0) enables the following Taylor expansion: $e^{-\tilde{d}_{t,k}} \approx 1 - \tilde{d}_{t,k} + O(N^{-2})$. Averaging over k, we get: $\frac{1}{K}\sum_{k=1}^{K} e^{-\tilde{d}_{t,k}} \approx 1 - \frac{1}{KN}\sum_{k=1}^{K} \frac{d_{t,k}}{p_k} + O(N^{-2})$. Taking the logarithm and recalling that $-\log(1-\epsilon) \approx \epsilon$ for small ϵ, we obtain:

$$s_t \approx \frac{1}{KN} \sum_{k=1}^{K} \frac{d_{t,k}}{p_k} + O(N^{-2}) \tag{6}$$

Therefore, as N increases, the OOD score s_t decays predictably as $O(1/N)$. This reflects the intuition that with more data, centroids stabilise, distances normalise by larger w_k, and the model becomes more confident about in-distribution membership. In summary, LDA emerges as a robust estimator: its OOD scores are tightly bounded between interpretable limits, governed by key parameters (K, N), and converge reliably with increasing data.

4. Experimental results

In this section, we describe the datasets used in our experiments (Section 4.1) and present the results obtained by the pre-trained segmentation model (Section 4.2). We then evaluate its performance when minimally adapted to new target datasets, both individually (Section 4.3)

Table 2
Results and comparisons with the literature across datasets and adaptation setups. Results preceded by ∼ symbol indicates estimates taken by looking at graphs. - symbol denotes experiments not performed by previous studies; / symbol indicates an experiment not included in a given adaptation setup. "FT t" denotes fine-tuning the pre-trained nnUNet with t labelled samples from a single target dataset. "J-FT t" denotes joint fine-tuning with t labelled samples per dataset, pooled across all target datasets. * denote our solutions. Values are reported with 95 % confidence intervals.

	ICPR2024				ISBI2015				MICCAI2016				3DMRMS			
	Dice	Recall	Precision	F1	Dice	Recall	Precision	F1	Dice	Recall	Precision	F1	Dice	Recall	Precision	F1
[12]	–	–	–	–	∼0.60	∼0.70	∼0.55	∼0.58	–	–	–	–	–	–	–	–
[24]	–	–	–	–	0.58	0.45	0.92	0.61	0.70	0.78	0.53	0.63	–	–	–	–
[23]	–	–	–	–	0.68	–	–	–	–	–	–	–	–	–	–	–
[26].	–	–	–	–	0.48	0.35	0.88	0.50	**0.72**	**0.74**	0.73	0.74	–	–	–	–
[10]	–	–	–	–	0.58	0.48	0.84	0.61	–	–	–	–	–	–	–	–
MadSeg [a]	**0.71**	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
M3S:Meet-MIALMS [a]	**0.71**	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
nnUNet FT 0 (*)	**0.71** ± **0.04**	**0.70** ± **0.05**	0.79 ± 0.06	**0.74** ± **0.04**	0.71 ± 0.07	0.70 ± 0.11	0.82 ± 0.06	0.76 ± 0.07	0.62 ± 0.06	0.62 ± 0.06	0.92 ±0.08	0.74 ± 0.06	0.53 ± 0.10	0.47 ± 0.09	**0.95** ± **0.08**	0.63 ± 0.09
nnUNet FT 1 (*)	/	/	/	/	0.63 ± 0.15	0.34 ± 0.19	0.92 ± 0.05	0.5 ± 0.15	0.62 ± 0.06	0.63 ± 0.06	0.81 ± 0.07	0.71 ± 0.06	0.60 ± 0.08	0.54 ± 0.07	0.91 ± 0.10	0.67 ± 0.08
nnUNet FT 3 (*)	/	/	/	/	0.82 ± 0.02	0.73± 0.05	**0.94** ± **0.02**	0.82 ± 0.02	0.68 ± 0.06	0.67 ± 0.06	0.89 ± 0.08	0.77 ± 0.08	0.67 ± 0.10	0.69 ± 0.09	0.89 ± 0.12	0.78 ± 0.10
nnUNet FT 5 (*)	/	/	/	/	0.76 ± 0.07	0.60 ± 0.11	0.94 ± 0.04	0.73 ± 0.07	0.70 ± 0.05	0.72 ± 0.06	0.89 ± 0.08	**0.80** ± **0.06**	0.70 ± 0.09	**0.75** ± **0.07**	0.85 ± 0.12	0.80 ± 0.09
nnUNet J-FT 1 (*)	0.62 ± 0.06	0.55 ± 0.07	0.71 ± 0.09	0.62 ± 0.06	0.72 ± 0.08	0.52 ± 0.12	0.94 ± 0.04	0.67 ± 0.08	0.62 ± 0.05	0.47 ± 0.06	0.90 ± 0.08	0.62 ± 0.06	0.59 ± 0.08	0.50 ± 0.07	0.93 ± 0.12	0.65 ± 0.09
nnUNet J-FT 3 (*)	0.64 ± 0.04	0.68 ± 0.06	0.66 ± 0.04	0.67 ± 0.05	0.80 ± 0.05	0.81 ± 0.06	0.88 ± 0.06	0.84 ± 0.05	0.67 ± 0.06	0.71 ± 0.06	0.89 ± 0.08	0.79 ± 0.07	0.70 ±0.09	0.72 ± 0.07	0.88 ± 0.11	0.79 ± 0.09
nnUNet J-FT 5 (*)	0.64 ± 0.04	0.57 ± 0.05	**0.85** ± **0.05**	0.68 ± 0.04	0.83 ± 0.03	0.77 ± 0.04	**0.94** ± **0.02**	0.85 ± 0.03	0.65 ± 0.06	0.68 ± 0.07	0.92 ± 0.08	0.78 ± 0.06	0.72 ± 0.10	0.73 ± 0.10	0.92 ± 0.10	**0.82** ± **0.10**
nnUNet J-FT 10 (*)	0.67 ± 0.06	0.68 ± 0.07	0.78 ± 0.07	0.72 ± 0.06	**0.84** ± **0.03**	**0.82** ± **0.05**	0.92 ± 0.01	**0.87** ± **0.03**	0.68 ± 0.05	0.70 ± 0.06	**0.92** ± **0.07**	**0.80** ± **0.06**	**0.74** ± **0.08**	0.70 ± 0.08	0.88 ± 0.08	0.82 ± 0.08
nnUNet J-FT 15 (*)	0.62 ± 0.05	0.65 ± 0.06	0.71 ± 0.07	0.68 ± 0.05	0.82 ± 0.04	0.87 ± 0.02	0.84 ± 0.06	0.85 ± 0.04	0.64 ± 0.06	0.71 ± 0.05	0.84 ± 0.08	0.77 ± 0.07	0.66 ± 0.10	0.71 ± 0.09	0.84 ± 0.10	0.77 ± 0.10

[a] https://iplab.dmi.unict.it/mfs/ms-les-seg/.

and simultaneously (Section 4.4). Supplementary Figs. 3 and 4 show this performance for visual comfort. Next, we assess the ability of the LDA-derived OOD scores to rank-correlate with error metrics (Section 4.5). Finally, we outline the main limitations of our approach (Section 4.6). The experiments are conducted on a node equipped with NVIDIA A100 GPUs (forward pass speed ∼ 1 sec/scan). Training losses and durations, expressed in epochs, are reported in Supplementary Fig. 1. Performance uncertainty is estimated using 95 % confidence intervals (CI) derived from test set bootstrapping (N = 1000). During every evaluation, to remain as dataset-agnostic as possible, we use a segmentation threshold of 0.5. In addition, to probe threshold effects, we conduct sensitivity analyses with multiple thresholds and find that performance remains overall stable, underscoring the robustness of our approach. (see Supplementary Figs. 5 and 6)

4.1. Datasets

For both pre-training and adaptation, we use four publicly available, real-world MS datasets. Each source comprises multi-modal MRI scans (T1w, T2w, and FLAIR) with corresponding expert lesion annotations. These datasets vary in acquisition protocols, scanner types, and cohort characteristics, providing a diverse foundation for studying domain variability and model generalisability under distribution shifts.

- **ICPR2024** [35], available for the MSLesSeg competition[1], includes 92 training and 22 test scans from about 90 patients, taken at multiple time points and with different scanners. Although collected in a single centre, it exhibits substantial inter- and intra-subject variability.
- **ISBI2015** [36] provides 21 annotated longitudinal MRI scans from five patients. All scans follow a uniform imaging protocol and are acquired with the same scanner, offering a highly consistent setup.
- **MICCAI2016** [37] is a multicentre dataset comprising 15 training scans (one per patient across three centres) and 38 test scans from four centres. Three centres overlap between training and test sets, which introduces inter-site variability.
- **3DMRMS** [38] consists of 30 MRI series acquired at a single site using consistent scanner hardware. While homogeneous in protocol, it adds anatomical and scanner-specific variability to the dataset pool.

We exclude the MICCAI2021 dataset [39] because it targets lesion progression via differential annotations, which deviates from our focus on full lesion segmentation at each time point. To ensure reproducibility and fair comparisons across experiments, we define fixed test sets for each dataset. Specifically, we use the official test sets for ICPR2024 and MICCAI2016, and extracted test sets for ISBI2015 (5 scans stratified across patients and time points) and 3DMRMS (15 scans stratified

[1] https://iplab.dmi.unict.it/mfs/ms-les-seg/

by acquisition machine). This extraction is necessary because ISBI2015 has no available labels for the official test partition and 3DMRMS does not provide a testing split. For pre-training nnUNet, we use the entire training set of ICPR2024, as it offers a diverse and relatively large collection of multi-modal MS volumes. For minimal-instance adaptation, we sample instances from the training partitions of ISBI2015, MICCAI2016 and 3DMRMS. For cross-domain few-shot adaptation, we instead sample from all sources.

4.2. Building a state-of-the-art MS segmentation baseline

After a 9-hour pre-training on the entire ICPR2024's training set, our baseline model achieves state-of-the-art performance on this dataset and improves upon our 2nd place in the MSLesSeg Competition [35]. This competition is considered one of the most challenging benchmark for MS lesion segmentation. Also, when directly transferred to external datasets including ISBI2015, MICCAI2016 and 3DMRMS (0-instance fine-tuning configuration), we surpass or remain competitive with previous works according to several metrics (Table 2). This solid performance provides ideal foundation for the subsequent adaptation analysis.

4.3. Minimal-instance adaptation yields consistent gains across datasets

Adaptation to a specific dataset is considered effective when the adapted model outperforms its non-adapted counterpart on that source. Table 2 and Supplementary Fig. 3 show the performance of the pre-trained nnUNet without adaptation (0-instance fine-tuning), as well as results after fine-tuning on 1, 3, and 5 labelled samples. This allows us to quantify the minimum annotation effort required to achieve meaningful gains under domain shift. On MICCAI2016 and 3DMRMS, we observe consistent improvements in Dice score, Recall, and F1 with increasing numbers of fine-tuning samples, although with diminishing returns. Precision, in contrast, fluctuates slightly on MICCAI2016 and gradually decreases on 3DMRMS-likely reflecting shifts in the model's sensitivity-specificity balance during adaptation. Notably, fine-tuning with three volumes yields significant gains over both the baseline and 1-instance adaptation across all metrics-except Precision, which follows a distinct trend. These results confirm that fine-tuning with just 1–3 labelled examples consistently improves performance, demonstrating practical low-resource adaptation for clinical settings. For benchmarking purposes, we compile in Table 2 the aforementioned results (FT) and 3D segmentation performance from the literature. We highlight that, with minimal supervision, we outperform previous solutions on ISBI2015 across all metrics and surpass state-of-the-art works on MICCAI2016 in terms of Recall and F1. However, for 3DMRMS, comparisons remain challenging, as-to our knowledge-no existing work has addressed full 3D segmentation on this dataset. Examples of segmentations across datasets and various levels of adaptation are depicted in Fig. 1. The segmentation masks for the ISBI2015 patient reflect overall performance trends: adapting with 1 volume worsens results, suggesting a significant shift, while 3 samples improve alignment with the ground truth. On MICCAI2016, the baseline model detects major lesions but misses inflamed regions near the lateral ventricles. Fine-tuning with 1 volume improves these at the cost of corpus callosum segmentation, which is fully recovered with 3 instances. On 3DMRMS, both baseline and adapted models effectively segment major lesions and localize minor ones, especially with 3–5 volumes. However, tiny inflamed regions remain undetected on this slice. This overall trend is also discernible in Fig. 2, which depicts segmentation masks on the same patient volumes.

4.4. Few-shot cross-dataset adaptation significantly enhances multi-domain generalisation

Adapting a model to a specific distribution can improve performance on that particular domain but may compromise generalisation to others. Table 2 and Supplementary Fig. 4 show the results of our cross-

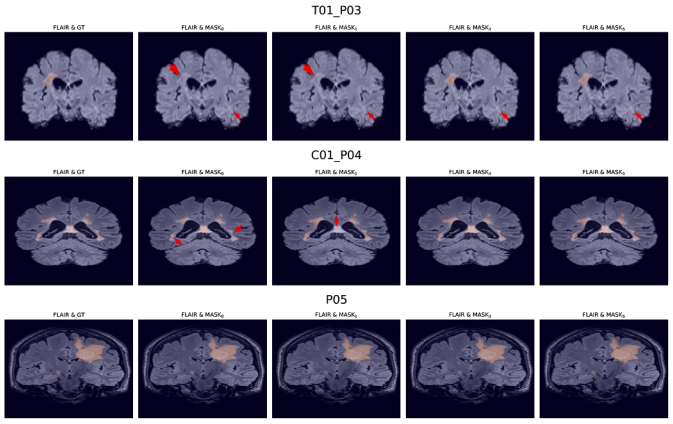

Fig. 1. Examples of segmentation by increasingly adapted models on instances from ISBI2015 (top), MICCAI2016 (centre), and 3DMRMS (bottom). Patient identifiers are reported as row titles. The number of instances used for fine-tuning are reported as subscript of `MASK`. 'GT' stands for ground-truth. Arrows point at major unsegmented/partially segmented lesions. Mid-volume slices are considered. (Zoom in for better view).

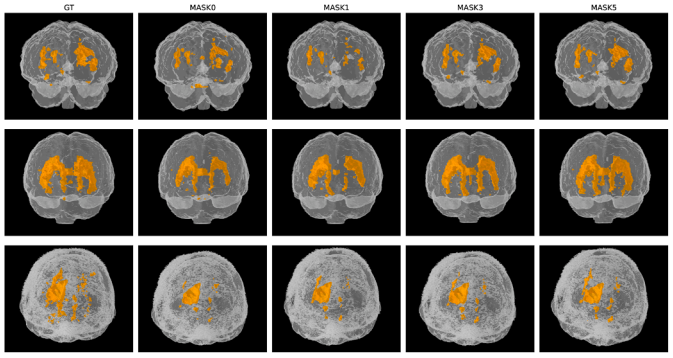

Fig. 2. Examples of 3D segmentations on instances from ISBI2015 (top), MICCAI2016 (centre) and 3DMRMS (bottom) after adaptation with 0, 1, 3, 5 volumes from each individual target dataset. (Zoom in for better view).

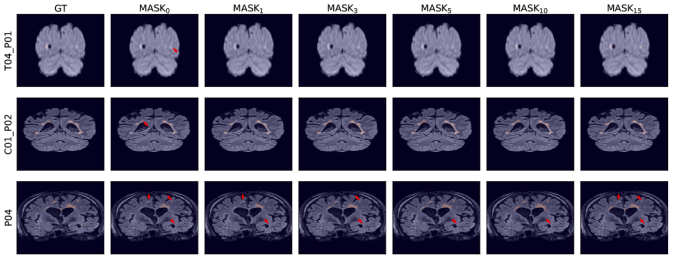

Fig. 3. Segmentation examples from ISBI2015 (top), MICCAI2016 (centre), and 3DMRMS (bottom) by increasingly adapted models on pooled sets of instances from each dataset. Patient identifiers are reported as column titles. The cardinalities of the pools used for fine-tuning are reported as subscript of `MASK`. 'GT' stands for ground-truth. Arrows point at major unsegmented/partially segmented lesions. Mid-volume slices are considered. (Zoom in for better view).

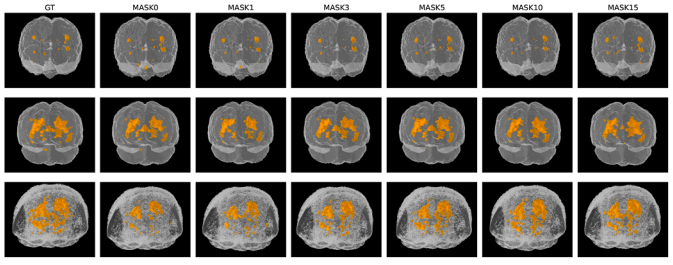

Fig. 4. Examples of 3D segmentations on instances from ISBI2015 (top), MICCAI2016 (centre) and 3DMRMS (bottom) after adaptation with pooled sets 0, 1, 3, 5, 10, 15 volumes from each target dataset. (Zoom in for better view).

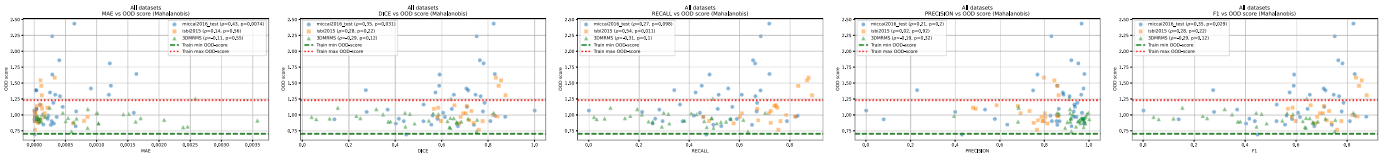

Fig. 5. Scatter plots showing the relationship between OOD scores and various prediction error metrics across all datasets. Positive rank-correlation between OOD scores and MAE is desirable, while a negative rank-correlation is expected for other metrics. Green and red dashed lines indicate the minimum and maximum OOD scores found in the training distribution, respectively. (Zoom in for better view).

dataset few-shot adaptation, where pooled sets of 1, 3, 5, 10, and 15 labelled volumes per target dataset, including ICPR2024, are used jointly for multi-dataset fine-tuning. This setup enables the quantification of the minimal cross-domain fine-tuning effort necessary for adaptation. It also allows to evaluate the adaptation's effect on generalisation across diverse sources simultaneously. The results demonstrate clear and consistent improvements across all unseen datasets and metrics. These gains are evident when compared to both the non-adapted baseline and previous adaptation settings. We also assess how performance changes on the original dataset and find a non-negligible decrease. This reduction, in contrast to previous trends, requires several samples per dataset for compensation and improvement. Overall, these findings highlight that jointly conditioning adaptation on multiple heterogeneous distributions can drastically improve cross-domain generalisation, even with *minimal* supervision. This observation significantly strengthens the findings by [8]. Notably, this strategy is especially practical in realistic low-resource scenarios where only a few annotated instances per dataset are available. In these cases, pooling limited supervision across datasets proves more effective and scalable than exhaustive per-dataset adaptations. For both comparative clarity and benchmarking purposes, we compare our results (J-FT) against those obtained in the previous adaptation setups (FT) as well as prior literature (see Table 2). On ISBI2015, we further improve all metrics, surpassing the performance of both our previous adaptations and prior works. We also strengthen our Precision and F1 scores on MICCAI2016, remaining competitive in terms of Dice and Recall. On 3DMRMS, few-shot cross-dataset adaptation delivers significant improvements over previous results in terms of Dice and F1, while only approaching prior Recall and Precision. We hypothesise that conditioning on heterogeneous sources acts as regularization, reducing dataset-specific biases and driving the observed gains in generalisation. Examples of segmentations after cross-dataset adaptation are presented in Fig. 3. For the ISBI2015 patient, cross-source fine-tuning helps identify **very** tiny lesions with just a single additional volume; for MICCAI2016 subject, pooled sets of samples help reduce false negatives in already detected lesions; for the 3DMRMS subject, increasingly larger pools enable the localisation of missed regions. This overall trend is also visible in Fig. 4, which depicts segmentation masks on the same subject volumes.

4.5. LDA-Quantified OOD score correlates with errors

Before adapting a model to a new distribution, it is crucial to assess whether incoming samples truly lie out-of-distribution. Conversely, when a model has already been adapted, it becomes important to determine whether the adaptation has successfully reduced the degree of OOD-ness, especially without ground truth labels at test time. In both these contexts, our LDA framework offers a practical, online, and label-free mechanism for estimating sample typicality. Knowing that ICPR2024 training volumes were acquired using three scanners-as supported by a t-SNE visualisation presented in Supplementary Fig. 2-we leverage this prior knowledge for clustering latent features into $K = 3$ clusters using k-means. As shown in Fig. 5, LDA assigns higher OOD scores to samples from unseen sites, effectively flagging them as atypical relative to the training distribution. These scores deviate from zero (representing centroid-based typicality) and capture how far a test instance lies from the densest regions of the latent space. Moreover, OOD scores exhibit rank-correlations with various error metrics across datasets, which strengthens the credibility of LDA as a meaningful unsupervised reliability estimator. However, we also observe that not all rank-correlations are statistically significant or consistently aligned in sign across metrics and datasets. For example, there exists a positive statistically significant rank-correlation between MICCAI2016 OOD scores and corresponding Mean Absolute Errors (MAEs), but the same relationship becomes insignificant on ISBI2015 and changes also sign on 3DMRMS. Conversely, there is a desirable negative rank-correlation between Dice, Recall, Precision and F1 metrics, and OOD scores of 3DMRMS samples. This rank-correlation is always statistically significant except for Precision.

4.6. Limitations

Despite strong performance across domains and encouraging LDA results, some limitations remain. The modest size of the public datasets, while broader than in prior works, may still constrain the model's generalisability. Single-instance adaptation is also effective but not universally so (e.g., ISBI2015 in Fig. 1), suggesting that minimal collection efforts may still be required. Furthermore, the model occasionally produces incomplete segmentations of small, challenging lesions (e.g., Fig. 3, instance `P04`), highlighting an intrinsic task complexity and a clear area for future refinement. Finally, clinical applicability of both our methodology and LDA require validation on larger and more diverse datasets before real-world deployment.

5. Conclusions and future works

In this work, we propose a 3D multi-modal model for MS lesion segmentation that, through minimal adaptation, achieves state-of-the-art performance across four diverse datasets. Using just 1-3 labelled samples, we demonstrate effective fine-tuning both per-domain and jointly across sources, enabling practical deployment in low-resource clinical settings. We also introduce Latent Distance Analysis, an online, differentiable, and label-free reliability estimator that anticipates distribution shifts at test-time. It rank-correlates with segmentation errors, potentially paving the way for lightweight, trust-aware adaptation.

Future work will include designing new loss functions to improve small lesion segmentation, scaling LDA to larger, more heterogeneous cohorts to assess its stability across acquisition conditions and, by leveraging its reliability signal, running encoder-only test-time adaptation to evaluate the quality of this lightweight fine-tuning. Both these steps are essential to enable a prospective clinical validation and ensure transferability to routine hospital data. This translation is immediately enabled by our core technical contributions. The demonstrated efficacy of minimal-instance adaptation directly addresses the critical bottleneck of data scarcity, showing that a prospective study is feasible even with a realistically small annotated cohort. Efficient adaptation, combined with LDA reliability scores designed to foster physician's trust, lays the groundwork for safe clinical translation in collaboration with our neurology and neuroradiology departments. A concrete path forward includes retrospective validation on multicentre clinical archives, followed by small-scale prospective studies embedded into neuroradiology workflows to monitor real-time performance. In this setting, we plan to evaluate the non-adapted model alongside LDA outputs to assess

whether predictions and reliability signals support or hinder clinical interpretation. We plan to repeat this process after encoder-only adaptation to examine potential shifts in trust and performance.

CRediT authorship contribution statement

Edoardo Coppola: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Mattia Savardi:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization; **Alberto Signoroni:** Writing – review & editing, Supervision, Resources, Project administration.

Data availability

All the data used in this study is publicly available and links to data sources have been provided by the authors to facilitate access.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Edoardo Coppola reports equipment, drugs, or supplies was provided by Lombardy Region. Mattia Savardi reports equipment, drugs, or supplies was provided by Lombardy Region. Alberto Signoroni reports equipment, drugs, or supplies was provided by Lombardy Region. 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

This work has been partly funded by Regione Lombardia, under the "Programme of measures for economic recovery: development of new cooperation agreements with universities for research, innovation and tech transfer" DGR n. XI/4445/2021.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.patrec.2025.12.008

References

[1] D.R. van Nederpelt, G. Pontillo, M. Barrantes-Cepas, I. Brouwer, E.M.M. Strijbis, M.M. Schoonheim, B. Moraal, B. Jasperse, H.J.M.M. Mutsaerts, J. Killestein, et al., Scanner-specific optimisation of automated lesion segmentation in MS, NeuroImage 44 (2024) 103680.

[2] T. Wiltgen, J. McGinnis, S. Schlaeger, F. Kofler, C. Voon, A. Berthele, D. Bischl, L. Grundl, N. Will, M. Metz, et al., LST-AI: a deep learning ensemble for accurate MS lesion segmentation, NeuroImage 42 (2024) 103611.

[3] E. López-Varela, N.O. Pascual, J. Quezada-Sánchez, C. Oreja-Guevara, E.S. Bueso, N. Barreira, Enhanced multiple sclerosis diagnosis using high-resolution 3D OCT volumes with synthetic slices, Pattern Recogn. Lett. 189 (2025) 99–105.

[4] P.D. Molyneux, D.H. Miller, M. Filippi, T.A. Yousry, E.W. Radü, H.J. Adèr, F. Barkhof, Visual analysis of serial T2-weighted MRI in multiple sclerosis: intra-and interobserver reproducibility, Neuroradiology 41 (1999) 882–888.

[5] N. Shiee, P.L. Bazin, A. Ozturk, D.S. Reich, P.A. Calabresi, D.L. Pham, A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions, NeuroImage 49 (2) (2010) 1524–1535.

[6] E. Roura, A. Oliver, M. Cabezas, S. Valverde, D. Pareto, J.C. Vilanova, L. Ramió-Torrentà, À. Rovira, X. Lladó, A toolbox for multiple sclerosis lesion segmentation, Neuroradiology 57 (2015) 1031–1043.

[7] K. Selvaganesan, E. Whitehead, P.M. DeAlwis, M.K. Schindler, S. Inati, Z.S. Saad, J.E. Ohayon, I.C.M. Cortese, B. Smith, S. Jacobson, et al., Robust, atlas-free, automatic segmentation of brain MRI in health and disease, Heliyon 5 (2) (2019).

[8] A.M. Hindsholm, F.L. Andersen, S.P. Cramer, H.J. Simonsen, M.G. Askløf, M. Magyari, P.N. Madsen, A.E. Hansen, F. Sellebjerg, H.B.W. Larsson, et al., Scanner agnostic large-scale evaluation of MS lesion delineation tool for clinical MRI, Front. Neurosci. 17 (2023) 1177540.

[9] R. Kushol, A.H. Wilman, S. Kalra, Y.H. Yang, DSMRI: Domain shift analyzer for multi-center MRI datasets, Diagnostics 13 (18) (2023) 2947.

[10] S. Valverde, M. Salem, M. Cabezas, D. Pareto, J.C. Vilanova, L. Ramió-Torrentà, À. Rovira, J. Salvi, A. Oliver, X. Lladó, One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks, NeuroImage 21 (2019) 101638.

[11] S. Gaj, B. Thoomukuntla, D. Ontaneda, K. Nakamura, Subject-based transfer learning in longitudinal multiple sclerosis lesion segmentation, J. Neuroimaging 35 (1) (2025) e70024.

[12] J. Zhang, L. Zuo, B.E. Dewey, S.W. Remedios, S.P. Hays, D.L. Pham, J.L. Prince, A. Carass, Harmonization-enriched domain adaptation with light fine-tuning for multiple sclerosis lesion segmentation, in: Medical Imaging 2024: Clinical and Biomedical Imaging, 12930, SPIE, 2024, pp. 635–641.

[13] O. Wiles, S. Gowal, F. Stimberg, S. Alvise-Rebuffi, I. Ktena, K. Dvijotham, T. Cemgil, A fine-grained analysis on distribution shift, https://arxiv.org/abs/2110.11328 (2021).

[14] K. Fan, T. Liu, X. Qiu, Y. Wang, L. Huai, Z. Shangguan, S. Gou, F. Liu, Y. Fu, Y. Fu, et al., Test-time linear out-of-distribution detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 23752–23761.

[15] S. Liang, Y. Li, R. Srikant, Enhancing the reliability of out-of-distribution image detection in neural networks, in: International Conference on Learning Representations, 2018.

[16] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, in: International Conference on Learning Representations, 2017.

[17] A. Rondinella, E. Crispino, F. Guarnera, O. Giudice, A. Ortis, G. Russo, C. Di Lorenzo, D. Maimone, F. Pappalardo, S. Battiato, Boosting multiple sclerosis lesion segmentation through attention mechanism, Comput. Biol. Med. 161 (2023) 107021.

[18] M. Salem, S. Valverde, M. Cabezas, D. Pareto, A. Oliver, J. Salvi, À. Rovira, X. Lladó, Multiple sclerosis lesion synthesis in MRI using an encoder-decoder U-NET, IEEE Access 7 (2019) 25171–25184.

[19] S. Aslani, M. Dayan, L. Storelli, M. Filippi, V. Murino, M.A. Rocca, D. Sona, Multi-branch convolutional neural network for multiple sclerosis lesion segmentation, NeuroImage 196 (2019) 1–15.

[20] M. Hashemi, M. Akhbari, C. Jutten, Delve into multiple sclerosis (MS) lesion exploration: a modified attention U-Net for MS lesion segmentation in brain MRI, Comput. Biol. Med. 145 (2022) 105402.

[21] A. Kaur, L. Kaur, A. Singh, DeepCONN: patch-wise deep convolutional neural networks for the segmentation of multiple sclerosis brain lesions, Multimed. Tools Appl. 83 (8) (2024) 24401–24433.

[22] O. Cetin, B. Canel, G. Dogali, U. Sakoglu, Enhancing precision in multiple sclerosis lesion segmentation: a U-net based machine learning approach with data augmentation, NeuroImage 5 (1) (2025) 100235.

[23] T. Brosch, L.Y.W. Tang, Y. Yoo, D.K.B. Li, A. Traboulsee, R. Tam, Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation, IEEE Trans. Med. Imaging 35 (5) (2016) 1229–1239.

[24] S.R. Hashemi, S.S.M. Salehi, D. Erdogmus, S.P. Prabhu, S.K. Warfield, A. Gholipour, Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: application to multiple sclerosis lesion detection, IEEE Access 7 (2018) 1721–1735.

[25] N. Gessert, J. Krüger, R. Opfer, A.C. Ostwaldt, P. Manogaran, H.H. Kitzler, S. Schippling, A. Schlaefer, Multiple sclerosis lesion activity segmentation with attention-guided two-path CNNs, Comput. Med. Imaging Graph. 84 (2020) 101772.

[26] R.A. Kamraoui, V.T. Ta, T. Tourdias, B. Mansencal, J.V. Manjon, P. Coupé, DeepLesionBrain: towards a broader deep-learning generalization for multiple sclerosis lesion segmentation, Med. Image Anal. 76 (2022) 102312.

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

[28] S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, Longitudinal multiple sclerosis lesion segmentation using 3D convolutional neural networks, Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge (2015) 1–2.

[29] A. Rondinella, F. Guarnera, O. Giudice, A. Ortis, G. Russo, E. Crispino, F. Pappalardo, S. Battiato, Enhancing multiple sclerosis lesion segmentation in multimodal MRI scans with diffusion models, in: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2023, pp. 3733–3740.

[30] W. Liu, X. Wang, J. Owens, Y. Li, Energy-based out-of-distribution detection, Adv. Neural Inf. Process Syst. 33 (2020) 21464–21475.

[31] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks, Adv. Neural Inf. Process. Syst. 31 (2018).

[32] J. Chen, Y. Li, X. Wu, Y. Liang, S. Jha, Atom: robustifying out-of-distribution detection using outlier mining, in: Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21, Springer, 2021, pp. 430–445.

[33] 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, Nat. Methods 18 (2) (2021) 203–211.

[34] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, T. Darrell, Tent: fully test-time adaptation by entropy minimization, in: International Conference on Learning Representations, 2021.

[35] A. Rondinella, F. Guarnera, E. Crispino, G. Russo, C. Di Lorenzo, D. Maimone, F. Pappalardo, S. Battiato, ICPR 2024 Competition on multiple sclerosis lesion segmentation-methods and results, in: International Conference on Pattern Recognition, Springer, 2024, pp. 1–16.
[36] A. Carass, S. Roy, A. Jog, J.L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C.H. Sudre, et al., Longitudinal multiple sclerosis lesion segmentation: resource and challenge, NeuroImage 148 (2017) 77–102.
[37] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S.C. Pop, P. Girard, R. Ameli, J.-C. Ferré, et al., Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure, Sci. Rep. 8 (1) (2018) 13650.
[38] L. Ziga, P. Franjo, B. Likar, Z. Spiclin, A novel public MR image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater conensus, Neuroinformatics 16 (1) (2018) 51–63.
[39] O. Commowick, M. Kain, R. Casey, R. Ameli, J.C. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, S. Camarasu-Pop, T. Glatard, et al., Multiple sclerosis lesions segmentation from multiple experts: the MICCAI 2016 challenge dataset, Neuroimage 244 (2021) 118589.