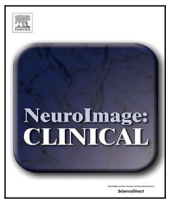
NeuroImage: Clinical 46 (2025) 103795

Contents lists available at ScienceDirect

NeuroImage: Clinical

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



Mitigating catastrophic forgetting in Multiple sclerosis lesion segmentation using elastic weight consolidation

Luisana Álvarez [a,b,*], Sergi Valverde [b], Àlex Rovira [c], Xavier Lladó [a]

[a] Vicorob Institute, University of Girona, Girona, Spain
[b] Tensor Medical, Girona, Spain
[c] Section of Neuroradiology, Department of Radiology (IDI), Vall d'Hebron University Hospital, Spain, Universitat Autònoma de Barcelona, Barcelona, Spain

ARTICLE INFO

Keywords:
Multiple sclerosis
Lesion segmentation
Continuous learning
Transfer learning
Catastrophic forgetting

ABSTRACT

Multiple sclerosis (MS) lesion segmentation is crucial for monitoring disease progression. Deep learning methods have shown promising results but suffer from domain shift problems when evaluated in data from different protocols or scanners. Transfer learning (TL) achieves successful domain adaptation, but can lead to catastrophic forgetting, resulting in a significant performance drop on the source domain. Continuous learning aims to address this issue by retaining knowledge from previous domains while adapting to new ones. This work applies Elastic Weight Consolidation (EWC) for the first time in the context of domain-incremental learning for MS lesion segmentation. The approach was evaluated using a 3D U-Net trained on public datasets (WMH2017 and Shifts) and fine-tuned on an in-house dataset using both TL and EWC, in both full training and few-shot scenarios. Results show that with only 3 training images from the target domain, EWC leads to a 10% improvement in F-score, while using 5 images achieves similar results to using all available training images. Catastrophic forgetting was reduced by 8%–19% compared to standard TL, where performance drops ranged from 20 to 37%. This work demonstrates that EWC enables models to adapt to new domains while preserving previous knowledge, with minimal data requirements, advancing towards more generalizable deep learning models for clinical MS applications.

1. Introduction

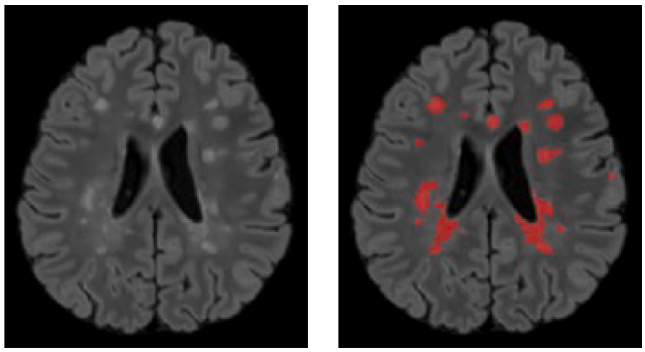
Multiple sclerosis (MS) is a chronic autoimmune and degenerative disease characterized by inflammation, demyelination, and axonal damage, resulting in the formation of lesions throughout both white and gray matter of the central nervous system, with pathology affecting cortical and subcortical structures (Compston and Coles, 2008). It is the most common cause of non-traumatic disability among young adults. MS is a very prevalent disease, reaching 2.9 million diagnosed patients in 2023, which means a prevalence of 36 per 100.000 people around the world.

Magnetic resonance imaging (MRI) is one of the main tools for diagnosis and monitoring of people with MS. MRI scans can provide quantitative information such as the number and volume of lesions (see Fig. 1). This quantitative analysis allows the assessment of disease progression and evaluation of therapies (Lladó et al., 2012; Chertcoff et al., 2024). McDonald criteria (Filippi et al., 2022) states that MRI scans should show evidence of damage in at least two separate areas of the central nervous system, including brain, spinal cord and optic nerves (dissemination in space) and at different points in time (dissemination in time) in order to diagnose a patient with MS. For this reason, segmentation of lesions becomes an important tool in clinical practice. Several automatic segmentation methods have been developed, with a strong focus in recent years on deep learning approaches, particularly Convolutional Neural Networks (CNNs). However, one of the main drawbacks of deep learning models is their lack of adaptability (Pianykh et al., 2020) when tested on data that differs from the one they were trained on (see Fig. 2(a)). This phenomenon, known as domain shift, occurs when the statistical distribution of the inference data (data the model is applied to) differs from the source data (data the model was trained on) (Guan and Liu, 2022). This is particularly important for commercial solutions intended for clinical practice, where variations in image acquisition, scanner, contrast, noise level, magnetic field strength (e.g., 1.5T vs. 3T) and the presence of bias field (intensity inhomogeneity) often result in poor generalization capabilities (Valverde et al., 2019).

Transfer learning (TL) emerges as a possible solution to the above mentioned problem by using the knowledge gained in solving a specific

* Corresponding author at: Tensor Medical, Girona, Spain.
E-mail address: lalvarez@tensormedical.ai (L. Álvarez).

https://doi.org/10.1016/j.nicl.2025.103795
Received 11 November 2024; Received in revised form 4 April 2025; Accepted 30 April 2025
Available online 20 May 2025



(a) FLAIR (b) Ground truth

Fig. 1. Example MRI of a MS patient from the White Matter Hyperintensity Challenge (Kuijf et al., 2019). (a) FLAIR modality. (b) Ground truth overlayed on FLAIR.

problem to improve the performance on a target task with a different underlying data distribution (Karimi et al., 2021). Ghafoorian et al. (2017) analyzed the effect of the training set size and the number of unfreezed layers when performing TL for domain adaptation applied to MS lesion segmentation. Valverde et al. (2019) also studied the impact of the amount of unfreezed parameters but in one-shot domain adaptation scenarios, assessing how did the lesion load of the chosen subject impact in the TL results. However, the main focus of TL is to leverage prior knowledge, rather than retaining it, leading to an abrupt loss in performance in the source dataset once the model is retrained on the target dataset (Pianykh et al., 2020). This phenomenon is known as catastrophic forgetting (see Fig. 2(b)) and it is a significant limitation for MS lesion quantification, where patients undergo periodic follow-up MRI scans, usually acquired in different scanners, potentially leading to performance degradation of the models over time.

Continuous learning (CL) arises as a solution to this issue, with the objective of retaining knowledge from previous tasks while adapting to new tasks. In essence, CL aims to continuously expand the model's capacity in an incremental way, allowing it to learn and integrate new information without forgetting past knowledge (see Fig. 2). Therefore, a model that can learn continuously without forgetting previous knowledge is especially valuable in the context of MS, where continuous adaptation to new data is crucial.

There are different CL strategies to prevent catastrophic forgetting when learning new tasks. Rehearsal-based approaches store previous tasks' data in a small memory buffer to be used while training on new tasks. The stored data can be the original images (experience-replay based) (Perkonigg et al., 2021), deep features (latent replay-based) (Srivastava et al., 2021) or generated pseudo samples (generative replay-based) (Li et al., 2023). Regularization-based methods aim to control weight update within the training of the model to minimize forgetting the previous learning, either through knowledge distillation from a teacher model to a student model (data-focused regularization) (Li and Hoiem, 2018) or by penalizing large changes on important parameters for previous tasks (prior-focused regularization) (van Garderen et al., 2019). Moreover, architectural-based methods assign to each task a set of parameters, either by fixing the architecture (limited by the network's capacity) (Bayasi et al., 2021) or dynamically extending the network (increasing memory requirements with each new task) (Yan et al., 2021).

Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) is one of the most well-known prior-based regularization methods. These methods present several key advantages compared to other strategies. Firstly, unlike rehearsal-based methods, they do not require storing data from previous tasks during training on new tasks, reducing privacy concerns and memory requirements. Secondly, they avoid the substantial memory overhead incurred by some architectural-based methods

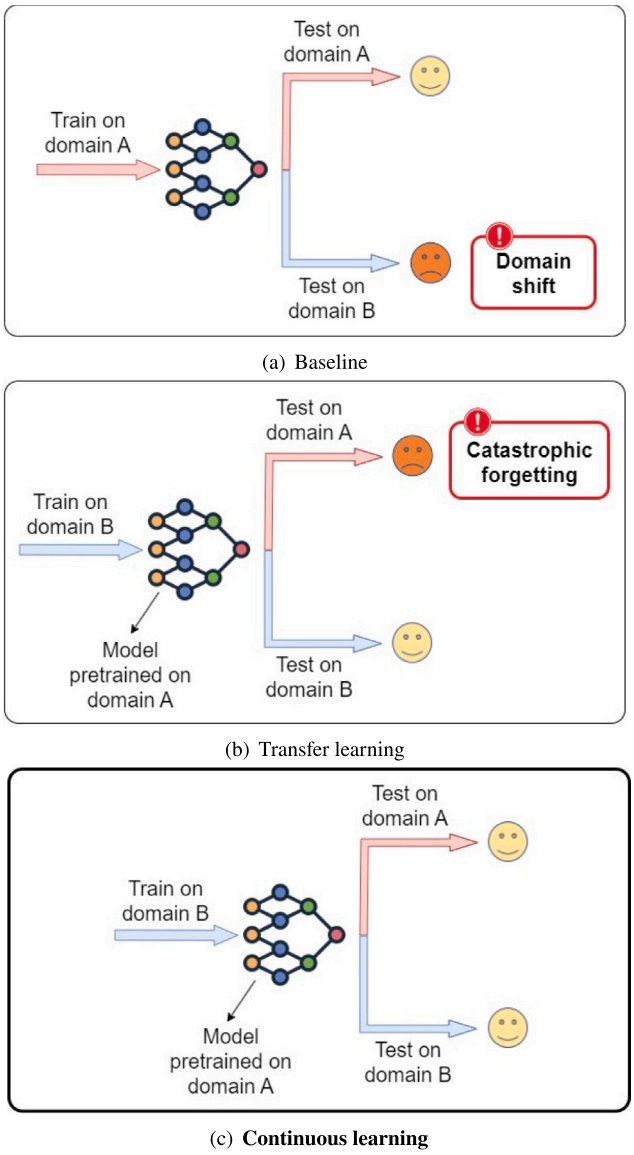


Fig. 2. Schematic representation of the problems faced in this work. (a) Deep learning models trained on domain A fail to generalize to unseen data with different distributions (domain B). (b) Transfer learning (TL): Retraining the model on domain B effectively adapts it to the target domain, but at the cost of catastrophic forgetting (drastic drop of performance) on the source domain A. (c) Continuous learning (CL): This approach aims to adapt the model to the target domain B while simultaneously preserving knowledge from domain A, thus mitigating catastrophic forgetting.

that dynamically expand the network for each new task. Finally, they offer more flexibility compared to strictly assigning network parts to each task in other architectural approaches. Within the realm of regularization-based methods, data-focused methods require an additional model for the new task (student model) to perform knowledge distillation from the previous task's model (teacher model). On the other hand, prior-focused methods, such as EWC, leverage information from the model's previous state directly, making them more lightweight and efficient.

EWC aimed to emulate synaptic consolidation of human brains to reduce catastrophic forgetting by adding a penalty term in the loss function that slowed down learning on specific weights that were important for previous tasks. EWC has been adapted to different medical tasks. For instance, van Garderen et al. (2019) applied it for a glioma segmentation problem, in order to perform TL from a public dataset containing

Table 1
Description of the datasets used for MS lesion segmentation. For each dataset, the table specifies MRI scanner model, sequence, spatial resolution (in mm^3), acquisition direction, number of subjects, and lesion load statistics (mean volume in ml with minimum and maximum values in brackets).

Dataset		Scanner	Sequence	Resolution (mm^3)	Acquisition direction	Number of subjects	Lesion load (ml) - mean [min, max]
WMH2017		3T Philips Achieva	2D FLAIR	$0.96 \times 0.95 \times 3.00$	Axial	20	38.9 [1.10, 98.46]
			3D T1w	$1.00 \times 1.00 \times 1.00$	Sagittal		
		3T Siemens TrioTim	2D FLAIR	$1.00 \times 1.00 \times 3.00$	Axial	20	36.46 [1.54, 118.35]
			3D T1w	$1.00 \times 1.00 \times 1.00$	Sagittal		
		3T GE Signa HDxt	3D FLAIR	$0.98 \times 0.98 \times 1.20$	Sagittal	20	14.23 [1.90, 55.80]
			3D T1w	$0.94 \times 0.94 \times 1.00$	Sagittal		
Shifts	MSSEG-1	3T Siemens Verio	3D FLAIR	$0.50 \times 0.50 \times 1.10$	Unknown	73	20.09 [0.19, 100.76]
			3D T1w	$1.00 \times 1.00 \times 1.00$			
		3T GE Discovery	3D FLAIR	$0.47 \times 0.47 \times 0.90$			
			3D T1w	$0.47 \times 0.47 \times 0.60$			
		1.5T Siemens Aera	3D FLAIR	$1.03 \times 1.03 \times 1.25$			
			3D T1w	$1.08 \times 1.08 \times 0.90$			
		3T Philips Ingenia	3D FLAIR	$0.74 \times 0.74 \times 0.70$			
			3D T1w	$0.74 \times 0.74 \times 0.85$			
	ISBI	3T Philips Medical	2D FLAIR	$0.82 \times 0.82 \times 2.20$			
			3D T1w	$0.82 \times 0.82 \times 1.17$			
VH		3T Siemens TrioTim	2D FLAIR	$0.49 \times 0.49 \times 3.00$	Axial	57	4.01 [0.13, 23.52]
			3D T1w	$1.00 \times 1.00 \times 1.20$	Sagittal		

low and high-grade glioma to an in-house dataset containing non-enhancing low-grade glioma. Baweja et al. (2018) used EWC to learn sequentially two different tasks: brain tissue segmentation and white matter lesions segmentation. Finally, Chen and Tang (2022) developed a CL pipeline with EWC penalty for breast tumor classification.

While a review of state-of-the-art CL methods revealed promising techniques for alleviating catastrophic forgetting in general, its application to domain-incremental MS lesion segmentation remains an under-explored area. Karthik et al. (2022) were the first to apply CL in this specific scenario, through a rehearsal-based approach. Data from previous tasks was stored in a memory buffer and interleaved with the current domain data. The main disadvantage is the necessity of having data from previous tasks available for training which, as mentioned before, not only leads to high storage requirements but also can provoke privacy violation issues.

This work proposes EWC as a CL approach to address catastrophic forgetting present in TL strategies in domain-incremental scenarios for MS lesion segmentation. The experimental framework specifically focuses on domain adaptation from public challenge datasets to clinical in-house data, representing the most common real-world implementation scenario where models trained on large standardized datasets are adapted to local clinical environments. This potential solution allows deep learning models to continuously adapt to data coming from new hospitals or scanners, a very common scenario in MS, making them more generalizable with time. EWC is integrated in a deep learning model based on a 3D U-Net for segmentation. Additionally, this work demonstrates the effectiveness of EWC in both one-shot and few-shot learning scenarios, minimizing the amount of necessary labeled images for training, which are very expensive to obtain in terms of effort and time.

2. Materials and methods

2.1. Datasets

2.1.1. White Matter Hyperintensity MICCAI challenge 2017 (WMH2017)

The WMH2017 dataset contained multimodal 3D brain MRI scans from 60 subjects acquired from five scanners of three different vendors (Simenes, Philips and General Electric) in three hospitals in the Netherlands and Singapore, as it can be seen in Table 1 (Kuijf et al., 2019). While the original challenge included 110 test cases, the ground truth annotations for these were not publicly available. Therefore, only the 60 training cases with available ground truth were used in our study. For each subject, T1w and Fluid-Attenuated Inversion Recovery (FLAIR) modalities were provided. The pre-processing steps included affine registration to the MNI $1 \times 1 \times 1$ mm^3 template (Fonov et al., 2009, 2011), skull stripping using HD-BET (Isensee et al., 2019) and bias field correction using the N4 algorithm (Tustison et al., 2010).

The lesions in this dataset were manually segmented following the STandards for ReportIng Vascular changes on nEuroimaging (STRIVE) criteria (Wardlaw et al., 2013). The segmentations were performed by an expert observer with extensive prior experience with manual segmentation and were peer-reviewed by a second one with eleven years of experience in quantitative neuroimaging and clinical neuroradiology. Any segmentation that did not conform to the STRIVE criteria was corrected in a consensus meeting between the two observers.

For our study, the images of the 60 patients were split into 41 images for training, 7 for validation and 12 for testing, ensuring that in each split there was approximately the same number of images from each scanner. In this case, 5 different splits were chosen to perform a 5-fold cross-validation.

2.1.2. Shifts challenge 2023

The Shifts dataset[1] aimed to simulate real-world scenarios in which there is a distributional shift between training and testing data (Malinin et al., 2022). As it can be seen in Table 1, it was composed by several public datasets such as ISBI, MMSSEG-1 and PubMRI, coming from different institutions and scanners. In this work, only the in-domain data available in the challenge was employed (MSSEG-1 and ISBI),

[1] Data were generated by participating neurologists in the framework of Observatoire Français de la Sclérose en Plaques (OFSEP), the French MS registry (Vukusic et al. 2020). They collect clinical data prospectively in the European Database for Multiple Sclerosis (EDMUS) software (Confavreux et al. 1992). MRI of patients were provided as part of a care protocol. Nominative data are deleted from MRI before transfer and storage on the Shanoir platform (Sharing NeurOImagingResources, shanoir.org). Vukusic S, Casey R, Rollot F, Brochet B, Pelletier J, Laplaud D-A, et al. Observatoire Français de la Sclérose en Plaques (OFSEP): A unique multimodal nationwide MS registry in France. Mult Scler. 2020;26(1):118–22. Confavreux C, Compston DAS, Hommes OR, McDonald WI, Thompson AJ. EDMUS, a European database for multiple sclerosis. J Neurol Neurosurg Psychiatry 1992; 55: 671-676. Andrey Malinin, Andreas Athanasopoulos, Muhamed Barakovic, Meritxell Bach Cuadra, Mark JF Gales, Cristina Granziera, Mara Graziani, Nikolay Kartashev, Konstantinos Kyriakopoulos, Po-Jui Lu, Nataliia Molchanova, Antonis Nikitakis, Vatsal Raina, Francesco La Rosa, Eli Sivena, Vasileios Tsarsitalidis, Efi Tsompopoulou, Elena Volf. Shifts 2.0: Extending The Dataset of Real Distributional Shifts, arxiv preprint https://arxiv.org/abs/2206.15407

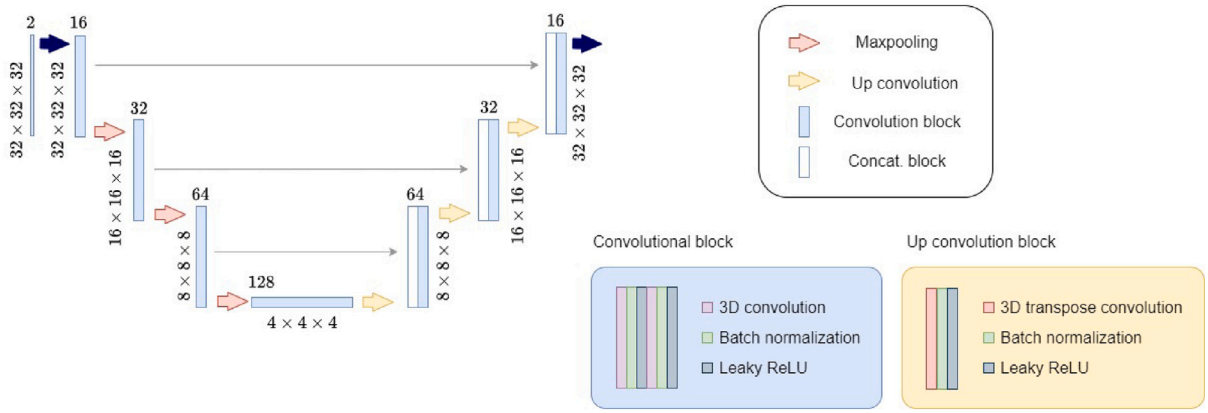


Fig. 3. Schematic representation of the 3D U-Net architecture employed for MS lesion segmentation. This model served as a baseline to study TL and CL techniques.

since it was preferred to use the in-house dataset as a domain-shift example (see Section 2.1.3), because more images were available.

The images were provided already pre-processed by the organization of the challenge, including denoising with non-local means (Coupe et al., 2008), skull stripping using HD-BET (Isensee et al., 2019), bias field correction with the N4 algorithm (Tustison et al., 2010) and interpolation to the 1 mm isovoxel space. In this case, as the raw images were not provided, the affine registration to the MNI $1 \times 1 \times 1$ mm^3 template (Fonov et al., 2009, 2011) was performed afterwards, in order to have all images in a common space. The ground-truth segmentation masks, also interpolated to the 1 mm isovoxel space, were obtained as a consensus of one or more expert annotators.

The Shifts dataset contained T1w and FLAIR brain MRI scans from 98 different subjects. In this case, as mentioned above, the splits were already provided by the challenge organization, containing scans from 33 subjects for training, 7 for validation and 33 for testing (in-domain samples).

2.1.3. Vall d'Hebron dataset

This in-house dataset came from the Vall d'Hebron (VH) University Hospital in Barcelona, Spain. The protocol was approved by the Vall d'Hebron Hospital (Barcelona, Spain) Research and Ethics Committee. Informed consent was obtained from each participant before enrolment in the study. It contained FLAIR and T1w MRI scans from 57 subjects acquired from a 3T Siemens scanner (see Table 1). For each patient, WM lesion masks were semi-automatically delineated from FLAIR images by an expert radiologist from the same hospital center. This ensured consistency in the annotation process across all subjects in this dataset. In this case, the images were pre-processed as stated in Valverde et al. (2019), including skull stripping, N3 bias field correction, co-registration to T1w (FSL-FLIRT) and affine registration to the MNI $1 \times 1 \times 1$ mm^3 template (Fonov et al., 2009, 2011). In this case, the images were randomly split into 25 samples for training, 5 for validation and 27 for testing.

It is important to note the differences in preprocessing pipelines across the three datasets used in this study. Even though the core preprocessing components were maintained across datasets (skull stripping, bias field correction and affine registration to the MNI space), the order in which they were applied necessarily differed due to data availability constraints. In the Shifts dataset, as the raw images were not provided, the registration step was performed at the end, after the challenge's pre-processing. In the WMH2017 and VH datasets, the registration was performed as the first step of the pre-processing pipeline.

2.2. MS lesion segmentation framework

2.2.1. Baseline architecture

In this work, a 3D U-Net architecture was implemented inspired by the original design proposed by Çiçek et al. (2016). This architecture has previously been adapted specifically for MS lesion segmentation (Wahlig et al., 2023; Greselin et al., 2024), showing good performance for this task. Even though the core principles of the 3D U-Net were maintained, it was customized for our specific application. While more advanced U-Net variations have been proposed in the literature, a simpler configuration was chosen in order to minimize the impact of architectural modifications on the evaluation of CL and TL methods. This allowed a clearer understanding of how these techniques influenced the network's performance in domain-incremental learning scenarios. Our 3D U-Net network architecture was composed by 4 layers of convolutional blocks, with 16, 32, 64 and 128 filters per layer. In the encoder side, each convolutional block contained two sequences of convolution - batch normalization - Leaky ReLU activation, followed by a maxpooling layer for downsampling. On the other hand, each decoder block was formed by an up convolution sequence (transposed convolution - batch normalization - Leaky ReLU activation) for upsampling followed by a convolution block equal to the ones in the encoder side. A diagram on the proposed architecture can be seen in Fig. 3.

2.2.2. Patch sampling

A patch-based approach was chosen for this work to perform the MS lesion segmentation. As done in related MS lesion segmentation studies (Fenneteau et al., 2021; Salem et al., 2022), 5000 patches of $32 \times 32 \times 32$ voxels were extracted from each image. To address the problem of class imbalance, an equal number of positive and negative patches were sampled, according to the class of its central voxel (Valverde et al., 2017). To ensure that the selected patches provided different information from the entire anatomy of the patient, they were not randomly sampled, but uniformly extracted from both the lesions and the healthy tissue. This uniform extraction was achieved by calculating a step size for each class, based on the number of available voxels of each class and the desired number of patches. Using these calculated step sizes, voxels were systematically sampled at regular intervals throughout both the lesion volume and healthy tissue regions, ensuring comprehensive spatial coverage of the entire brain. Patches from both FLAIR and T1w modalities were fed into the network in two separate input channels. FLAIR MRI provided good contrast between lesions and healthy tissue, while T1w sequences contributed with more structural information of the brain tissues.

2.2.3. Baseline model training strategy

The network was trained with batches of 32 patches to balance computational efficiency with gradient update quality. To address the class imbalance problem, balanced batches were constructed, ensuring an equal number of positive and negative patches within each batch. Cross-entropy was used as the loss function and the model was trained for a maximum of 300 epochs (Valverde et al., 2017). Early stopping was applied if the validation loss did not improve in the last 20 epochs, to avoid over-fitting. Adam optimizer was selected, with an initial learning rate of 10^{-4} and a weight decay of 10^{-6}. To manage the learning rate during training, a reduce-on-plateau scheduler was implemented. This scheduler automatically reduced the learning rate by a factor of 10 if the validation loss plateaued for 7 epochs, preventing the model from getting stuck in local minima. To enhance model generalization, data augmentation was implemented. This technique created variations of the original patches by applying random transformations with a 30% probability. Specifically, patches underwent random rotations within a range of $-\frac{\pi}{2}$ to $\frac{\pi}{2}$ radians, flipping along any spatial axis and affine transformations combining small rotations (−0.1 to 0.1 radians) and shearing (−0.1 to 0.1).

Three different models were trained using this strategy, one for each dataset presented in Section 2.1. These were the baseline models used as reference for each domain. Each model was evaluated on its corresponding testing set and on the other two datasets to assess the impact of domain shift.

For inference, a sliding window approach was employed with a 25% overlap between patches. To account for the overlap, the average probability was computed in the overlapping regions. The binarization threshold was optimized for each dataset to achieve the best trade-off between true positive and false positive lesion detections, with a focus on maximizing the detection F-score rather than segmentation accuracy. Lesions smaller than 3 voxels were excluded, in order to reduce the number of false positive lesions, at the cost of a reduction in sensitivity. This choice was selected based on previous works (Roura et al., 2015; Salem et al., 2018, 2020) and recent deep learning approaches for MS lesion segmentation that applied a minimum lesion size threshold of 3 voxels in 1x1x1 mm^3 resolutions (Krishnan et al., 2023; Wiltgen et al., 2024). Moreover, a comparative analysis done using different lesion size thresholds (2, 3, 5, and 7 voxels) confirmed that 3 voxels provided a good trade-off between TPF_d, FPF_d, and $F-score_d$.

2.3. Elastic weight consolidation (EWC)

Training a network consists on optimizing the value of a set of parameters θ by minimizing a loss function $\mathcal{L}$. Due to the high amount of parameters in a neural network, there should be different sets of parameter values that result in the same performance. This means that there should be a solution for task B (target domain), represented by its optimal parameters θ_B^*, that is close to the previous solution for task A (θ_A^*) (source domain). The goal of EWC is to find this particular solution, by forcing the network to learn the task B by finding θ_B^* as close as possible to θ_A^* in the parameter space. EWC achieves this by adding a penalty to the loss function for task B $\mathcal{L}_B$, that should be higher for:

- Parameters that are important for the performance in task A.
- Parameters that are getting further from the optimal values for task A.

The loss function to minimize in EWC is defined as follows:

$$\mathcal{L} = \mathcal{L}_B + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta_{A,i}^*)^2 \quad (1)$$

where F_i represents the importance of parameter i, θ_i represents the value of the parameter i in the current iteration (training on task B), $\theta_{A,i}^*$ represents the optimal value of the parameter i for the previous task A and λ controls the weight given to the old task A compared to the new one B. A higher value of λ means a stronger emphasis on preserving A's knowledge. However, very high values can lead to the network not learning the new task, so it is important to find a good trade-off.

As it can be seen in Eq. (1), this new term in the loss forces the network to minimize the difference between the current parameters and the optimal ones for the previous task A ($\theta_i - \theta_{A,i}^*$), weighted by the importance of each parameter F_i. Finally, the penalization term is computed as the sum of the penalization terms of each parameter.

The key component of this method is the parameter importance F, which quantifies how well each model parameter learned from task A reflects the characteristics of dataset A. This is calculated as the diagonal of the Fisher information matrix, which measures the amount of information about a parameter θ_i that is provided by a data sample X_j (in this case, a patch). It is computed as the average of the squared 1st derivatives of the log-likelihood function with respect to the parameters. This means that the slope of the likelihood function at the true parameter θ is a measure of the amount of information provided by the observed data regarding the parameter θ. Therefore, the Fisher information for parameter i is computed as follows:

$$F_i = \frac{1}{N} \sum_j^N \left(\frac{d}{d\theta_i} log\left[p\left(X_j|\theta\right) \right] \right)^2 \quad (2)$$

where N is the number of samples in the dataset A, X_j represents the sample j of dataset A and $p\left(X_j|\theta\right)$ represents log-likelihood, the probability of sample j given the model parameters (optimized for task A).

The advantage of this CL approach is that the parameter importance for the source domain can be computed after training on the source domain with a forward pass of the whole dataset (without parameter update). This eliminates the need to retain the source dataset for training on the target domain. Additionally, the only extra memory required is the one needed to store the importance scores and optimal parameters from the source task.

2.3.1. EWC and TL experimental setup

For evaluating EWC, a baseline model trained on a source dataset, as detailed in Section 2.2.3, was used as the base model. This model was then retrained on another dataset (target domain), both with and without the EWC penalty, representing CL and TL, respectively. After retraining, the models were evaluated on both the source and target domains to assess catastrophic forgetting and domain adaptation, respectively. This allows comparing TL and EWC. The hyperparameter configuration used during the base model training on the source dataset was maintained for all EWC and TL experiments. All the layers of the network were unfrozen during training.

To better understand how the penalization weight affected the performance of EWC on both the source and target domains, different values of the hyperparameter λ were analyzed (0.001, 0.01, 0.1, 1, 10, 10 and 1000). The optimal λ value was selected based on a trade-off between learning the target domain and preserving the source domain knowledge, and was kept fixed for the rest of the EWC experiments.

To evaluate the number of images required for adaptation to the target domain, one-shot and few-shot learning approaches were also analyzed. This involved fine-tuning the model with a limited number of images (1, 2, 3, 5, or 10) from the target domain. The images were selected based on the lesion volume of each subject to provide as much variability as possible in this regard. For the different training set sizes, the lesion characteristics were as follows: the single subject used for one-shot learning had a total lesion load of 23.51 ml with 69 lesions, the two-subject set had a combined 32.67 ml load (148 lesions), the three-subject set had 39.23 ml (196 lesions), the five-subject set measured 42.78 ml (236 lesions), while the ten-subject set contained 61.10 ml (301 lesions). The complete dataset of 32 subjects had 228.85 ml total lesion volume, with 1363 lesions. Two validation

Table 2
Segmentation and detection results of the baseline models trained on each of the available datasets. The models trained on public datasets (Shifts and WMH2017) were also tested in the other datasets to assess the impact of domain shift.

Train → Test	DSC_s	DSC_d	TPF_d	FPF_d	$F-score_d$
Shifts → Shifts	63.5 ± 15.1	65.7 ± 13.9	66.5 ± 17.1	28.6 ± 19.7	65.7 ± 13.9
WMH2017 → Shifts	49.8 ± 14.9	59.1 ± 13.3	59.7 ± 18.4	31.9 ± 21.2	59.1 ± 13.3
WMH2017 → WMH2017	74.3 ± 10.7	75.1 ± 10.1	81.0 ± 9.2	28.1 ± 11.7	75.2 ± 7.4
Shifts → WMH2017	50.3 ± 21.0	54.5 ± 16.5	76.5 ± 10.2	54.4 ± 19.1	54.5 ± 16.5
VH → VH	50.6 ± 16.8	61.0 ± 16.7	59.9 ± 18.4	33.8 ± 19.3	61.0 ± 16.7
Shifts → VH	35.9 ± 21.6	43.7 ± 19.2	60.1 ± 19.5	62.8 ± 20.4	43.7 ± 19.2
WMH2017 → VH	34.2 ± 20.2	47.2 ± 19.3	62.9 ± 19.6	59.4 ± 20.7	47.2 ± 19.3

images were also selected and kept fixed through all the experiments. Due to the small number of training images in the one-shot and few-shot experiments, a different patch extraction approach was employed. Instead of using a fixed number of patches, all possible patches of the training images were extracted centered on every positive voxel in the image. An equal number of negative patches was also extracted to maintain balanced training. The comparison between EWC and TL was also studied in one-shot and few-shot learning scenarios.

Both approaches required a similar number of epochs to converge. The training time did not differ significantly between EWC and TL, since the EWC penalty computation adds only lightweight matrix operations during the optimization process. The most computationally demanding part of EWC was the initial parameter importance calculation, which is performed only once for the source domain and can be reused across all subsequent experiments, making this approach particularly efficient for sequential domain adaptations.

2.4. Evaluation metrics

Perfect delineation of lesions is not always clinically relevant; however, accurate detection and localization of lesions are crucial for diagnosis and treatment planning. Recent MS lesion segmentation challenges, such as MSSEG-2 (Commowick et al., 2021) and Shifts (Malinin et al., 2022), have increasingly emphasized detection over segmentation. Therefore, given the clinical importance of lesion detection, lesion-wise metrics were prioritized over voxel-wise segmentation metrics, indicated by subindices d and s, respectively. The employed evaluation metrics include:

- Dice Similarity Coefficient (DSC): this metric was evaluated both voxel-wise (DSC_s) and lesion-wise (DSC_d).

$$DSC = \frac{2 \cdot TP}{FN + FP + 2 \cdot TP} \tag{3}$$

- Detection True Positive Fraction (TPF_d).
- Detection False Positive Fraction (FPF_d).

$$TPF_d = \frac{TP}{TP + FN} \quad FPF_d = \frac{FP}{FP + TP} \tag{4}$$

- Detection F-score: this metric combines the precision and sensitivity for lesion detection, providing a more balanced evaluation between TP and FP.

$$F - score_d = \frac{TP}{TP + 0.5 \cdot (FP + FN)} \tag{5}$$

To assess the statistical significance of differences in performance between models, paired t-tests were conducted on lesion-wise F-score values. Since only direct pairwise comparisons using a single performance metric were performed, correction for multiple testing was not applied. A p-value < 0.05 was considered statistically significant.

3. Results

3.1. Baseline: MS lesion segmentation

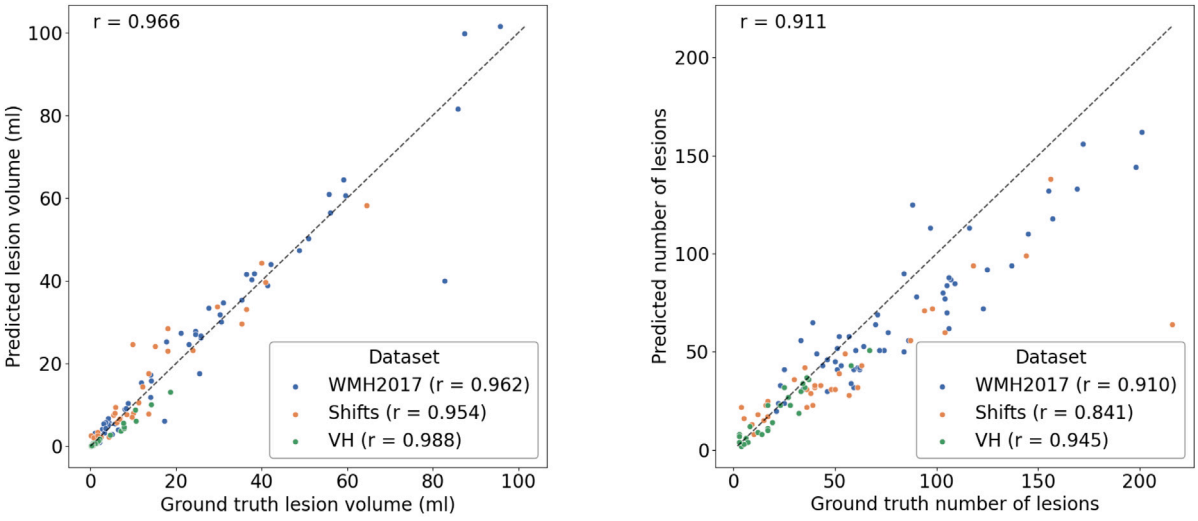
The results obtained with the baseline models trained on each individual dataset (Shifts, WMH2017 and VH) can be seen in Table 2, in which the entries are labeled as *Train* → *Test*, indicating the dataset on which the model was trained (*Train*) and the dataset used for evaluation (*Test*). The WMH2017 and Shifts models were also tested on the other two datasets. The VH dataset was used as an independent out-domain test set since it contains real-life data from clinical practice and it represents a particularly challenging domain shift, not only differing in scanners and acquisition protocols but also containing images from patients with substantially lower lesion loads compared to the challenge datasets. As expected, both base models exhibited a significant drop in performance when evaluated on datasets different from their training data. This confirmed the presence of a domain shift between the datasets.

To further validate the effectiveness of our baseline models, correlation analyses were conducted between the estimated and ground truth lesion measurements. Strong correlations were observed between the estimated and true lesion volumes (r = 0.966) and also for lesion counts (r = 0.911), as shown in Fig. 4. These high correlation values confirm that our baseline model provides reliable lesion quantification despite using a relatively simple architecture. Such an accurate quantification of lesion burden is particularly important for clinical applications in MS, where lesion count and volume are key metrics used for diagnosis and monitoring of disease progression and treatment response (Filippi et al., 2022; Oship et al., 2022). The strong correlation between predicted and ground truth measurements supports the clinical utility of our approach for quantitative MS assessment in real-world clinical settings.

Based on the overall performance in Table 2, the WMH2017 model was chosen as the baseline for the subsequent EWC experiments. This selection was justified by its good performance and the fact that the WMH2017 dataset was pre-processed specifically for this work, which allowed for greater control over the data. Unlike the Shifts dataset, which was provided pre-processed with voxel interpolation instead of registration to the MNI and was registered afterwards, the WMH2017 dataset underwent a more similar pre-processing pipeline to the VH dataset, including registration to MNI. This additional step in the Shifts dataset might have introduced some "noise" or distortion, potentially impacting the model's performance. Furthermore, Table 2 demonstrates a significant performance gap between *WMH2017 → VH* and *VH → VH* (upper bound) results (p-value < 0.05). More specifically, even though the sensitivity of *WMH2017 → VH* is even higher than the one of *VH → VH*, it results in a high number of false positive lesions, leading to a big gap in the detection F-score. This elevated false positive rate is a characteristic manifestation of domain shift, where the model incorrectly identifies structures or intensity patterns in the target domain as lesions due to differences in image acquisition parameters, scanner characteristics, and patient populations between the source and target domains.

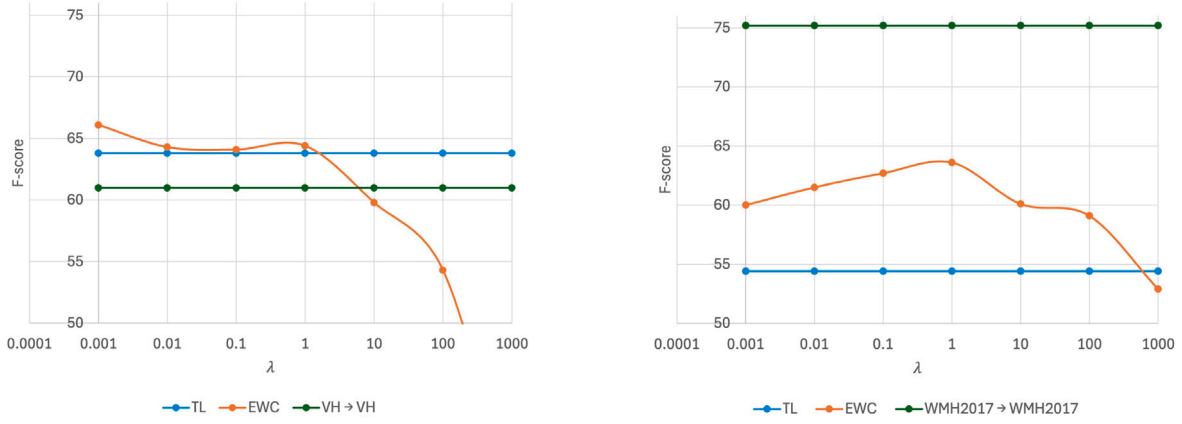
3.2. Elastic weight consolidation (EWC)

Table 3 shows the results of EWC after finetuning the WMH2017 model (source domain) with the VH dataset (target domain), compared to traditional TL. The results of the *WMH2017 → WMH2017* baseline model are also included in the Table, which also contains



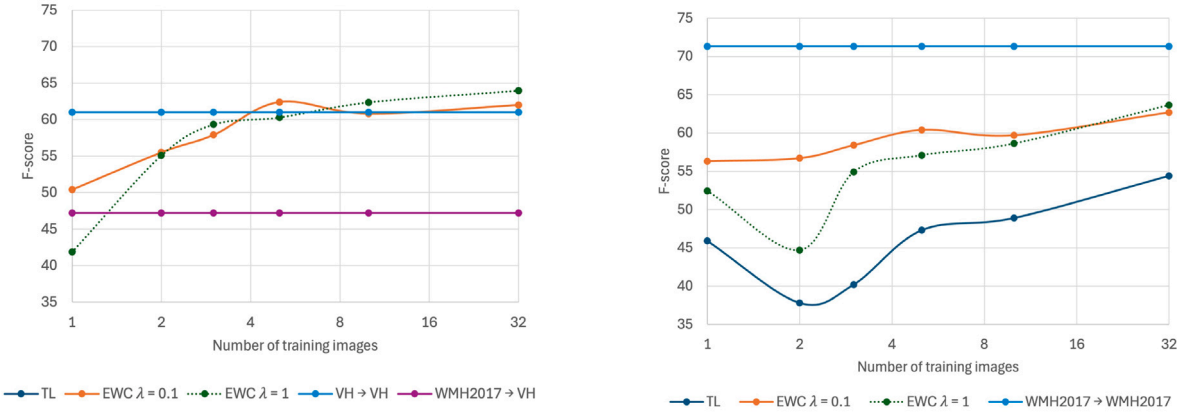
(a) Lesion volume correlation (b) Number of lesions correlation

Fig. 4. Correlation plots between predicted segmentations and ground truth.



(a) Target domain (b) Source domain

Fig. 5. Detection F-score for different values of λ in EWC penalization. λ axis is on logarithmic scale. EWC and TL are the results of retraining the source model (WMH2017) on the target domain (VH) with and without EWC penalization, respectively. Results are included for: (a) **Target domain** (VH): the VH → VH results are the ones obtained by training and testing the model on the VH dataset. (b) **Source domain** (WMH2017): the WMH2017 → WMH2017 results are the ones obtained by training and testing the model on the WMH2017 dataset.



(a) Target domain (b) Source domain

Fig. 6. Few-shot domain adaptation comparison between TL and EWC with $\lambda = 0.1$ (optimal penalization value) and $\lambda = 1$ (for comparison). Note that the x-axis is presented in a logarithmic scale (base 2) to better visualize results with smaller training set sizes. (a) **Target domain** (VH): the upper bound is the VH → VH model, and the lower bound corresponds to the WMH2017 → VH. (b) **Source domain** (WMH2017): the upper bound corresponds to the WMH2017 → WMH2017 model.

Table 3
Results of TL and EWC ($\lambda = 0.1$) approaches on both the source and target domain. The upper row shows the baseline model results (WMH2017 model).

	Target domain (VH)				Source domain (WMH2017)			
	DSC_d	TPF_d	FPF_d	$F-score_d$	DSC_d	TPF_d	FPF_d	$F-score_d$
WMH → WMH	47.2 ± 19.3	62.9 ± 19.6	59.4 ± 20.7	47.2 ± 19.3	75.1 ± 10.1	81.0 ± 9.2	28.1 ± 11.7	75.2 ± 7.4
TL	63.4 ± 15.8	63.5 ± 18.9	34.7 ± 17.5	63.4 ± 15.8	54.4 ± 12.8	58.3 ± 9.6	46.5 ± 18.8	54.4 ± 12.8
EWC	62.0 ± 15.2	59.9 ± 18.6	30.1 ± 19.5	62.0 ± 15.3	62.7 ± 9.2	61.0 ± 9.2	34.2 ± 13.3	63.0 ± 9.2

the results for both the source and target domain in all cases. In the case of EWC, the results correspond to the optimal penalization weight value $\lambda = 0.1$, which was selected because of the good trade-off provided between source domain knowledge preservation and domain adaptation to the target domain (see Fig. 5). As the performance was comparable for $\lambda = 0.1$ and $\lambda = 1$ in full-training scenarios on both the source and target domains, it was preferred to select the smaller value to avoid exploding gradients during training. Moreover, $\lambda = 0.1$ showed better performance on the source domain for smaller training set sizes compared to $\lambda = 1$ (see Fig. 6). With few training images, a stronger penalty worsens source domain preservation, probably due to competing optimization objectives (the need for significant parameter adjustments conflicts with the high regularization, creating instability and suboptimal solutions). Note that EWC provided significantly better knowledge preservation of the source domain compared to TL, as shown in a reduction of catastrophic forgetting of up to 9% in terms of F-score (p-value < 0.05). Additionally, on the target domain, the results of EWC were comparable to the ones of TL, meaning that the model also adapted to the target domain.

Fig. 7 provides qualitative results to visualize the performance differences between the baseline model, TL and EWC in both the source and target domains. The upper section of the Figure shows the source domain results. As expected, the baseline model achieved good performance on the source domain (75.2 of $F-score_d$). However, due to catastrophic forgetting, TL strategies exhibited a high number of false positive lesions in the source domain (FPF_d increased from 28.1 to 46.5 after TL). This increase in false positives occurs because the model has adapted its parameters to recognize lesions in the target domain, which differ from those in the source domain. When subsequently tested on source domain data, the model misinterprets intensities from partial volumes or brain structures as lesions based on the newly learned target domain characteristics. In the target domain, the baseline model showed poor performance due to the domain shift, again evident by the presence of numerous false positive lesions (59.4 of FPF_d). Following a TL strategy, the model successfully adapted to the target domain, achieving a segmentation comparable to the ground truth (from 47.2 to 63.4 of $F-score_d$ after TL). Finally, EWC results demonstrate the ability of this technique to balance the trade-off between source and target domain performance. EWC better preserves source domain knowledge (63.0 of $F-score_d$) compared to TL (54.4 $F-score_d$), resulting in segmentation more similar to the baseline model, while still adapting to the new target domain (62.0 $F-score_d$).

3.2.1. Few-shot continuous learning

Fig. 6 presents the results in terms of detection F-score for both the source and target domains of the few-shot CL study using EWC. One-shot learning resulted in an improvement of 10% in DSCs and 8% in precision. This is comparable with Valverde et al. (2019) one-shot experiments, which led to an improvement of 5%–24% in DSCs and 3%–27% in precision across subjects with varying lesion load and across different numbers of retrained layers. On the target domain (VH), retraining with only 3 images led to an F-score improvement of almost 10%, and training with 5 images was enough to reach a performance comparable to the upper bound (VH → VH, see Table 2). On the other hand, on the source domain we can see that EWC consistently led to better knowledge preservation compared to traditional TL, reducing catastrophic forgetting by 8% to 19% across different training set sizes. Additionally, EWC provided more stable performance in the source domain when compared to TL, where the extent of forgetting seemed more unpredictable. Interestingly, TL exhibited less severe forgetting as the number of training images increased, likely because adapting to a larger, more representative set of target domain examples required less parameter adjustments than when generalizing from very few examples, thus allowing better preservation of source domain knowledge.

4. Discussion

This work demonstrated the potential of EWC to improve the domain shift problem in the context of MS lesion segmentation, while mitigating catastrophic forgetting in the source domain. The results confirmed the effectiveness of this approach in improving model performance on a target domain (VH) while preserving knowledge from a source domain (WMH2017), in both full-training and few-shot learning scenarios.

The baseline models trained on each public dataset (WMH2017 and Shifts) achieved detection results comparable to the state-of-the-art. For instance, on the WMH2017 dataset, our baseline model achieved a TPF_d of 81.0 and a $F-score_d$ of 75.2. These values are competitive with the winning entry in the corresponding segmentation challenge, which reported a TPF_d of 84.0 and an $F-score_d$ of 76.0 (Kuijf et al., 2019). Additionally, the baseline results of the model trained on the VH dataset (TPF_d of 59.9 and DSC_s of 50.6) are comparable to those obtained in previous works (TPF_d of 60 and DSC_s of 53) (Valverde et al., 2019). These results validate that the baseline segmentation approach provides a reliable foundation for evaluating our proposed EWC methodology. Moreover, these baseline models showed the presence of a domain shift between the datasets. This was evident from the drop in performance observed when evaluating the models on datasets different from their training data (Table 2).

As shown in Table 3, EWC emerged as an effective method for preserving knowledge from the source domain (WMH2017) while still adapting to the target domain (VH). The success of this method relies on fact that it adapts the penalization during training at each step, based on both the importance of the parameter and the difference between its current value (training on the target domain) and optimal value (trained on the source domain), leading to dynamic adaptation during training. It is worth emphasizing the importance of careful tuning of the hyperparameter λ for balancing source domain knowledge preservation and target domain learning flexibility. We observed from our optimization study (see Fig. 5) that as the λ value increased, the model allowed less flexibility for learning the target domain. Conversely, for the source domain, a higher λ value translated to better knowledge preservation. However, excessively high λ led to a decrease in the source domain performance. This was likely because a very strong penalty resulted in high loss values during training, which in turn led to significant changes in model parameters, hindering the desired behavior of preserving previous knowledge.

Moreover, few-shot learning results demonstrated that EWC allows the model to successfully adapt to the target domain with very few images, achieving similar results to TL for all training set sizes and, more importantly, it reduced catastrophic forgetting for all the number of training images compared to TL. The effectiveness of EWC has been

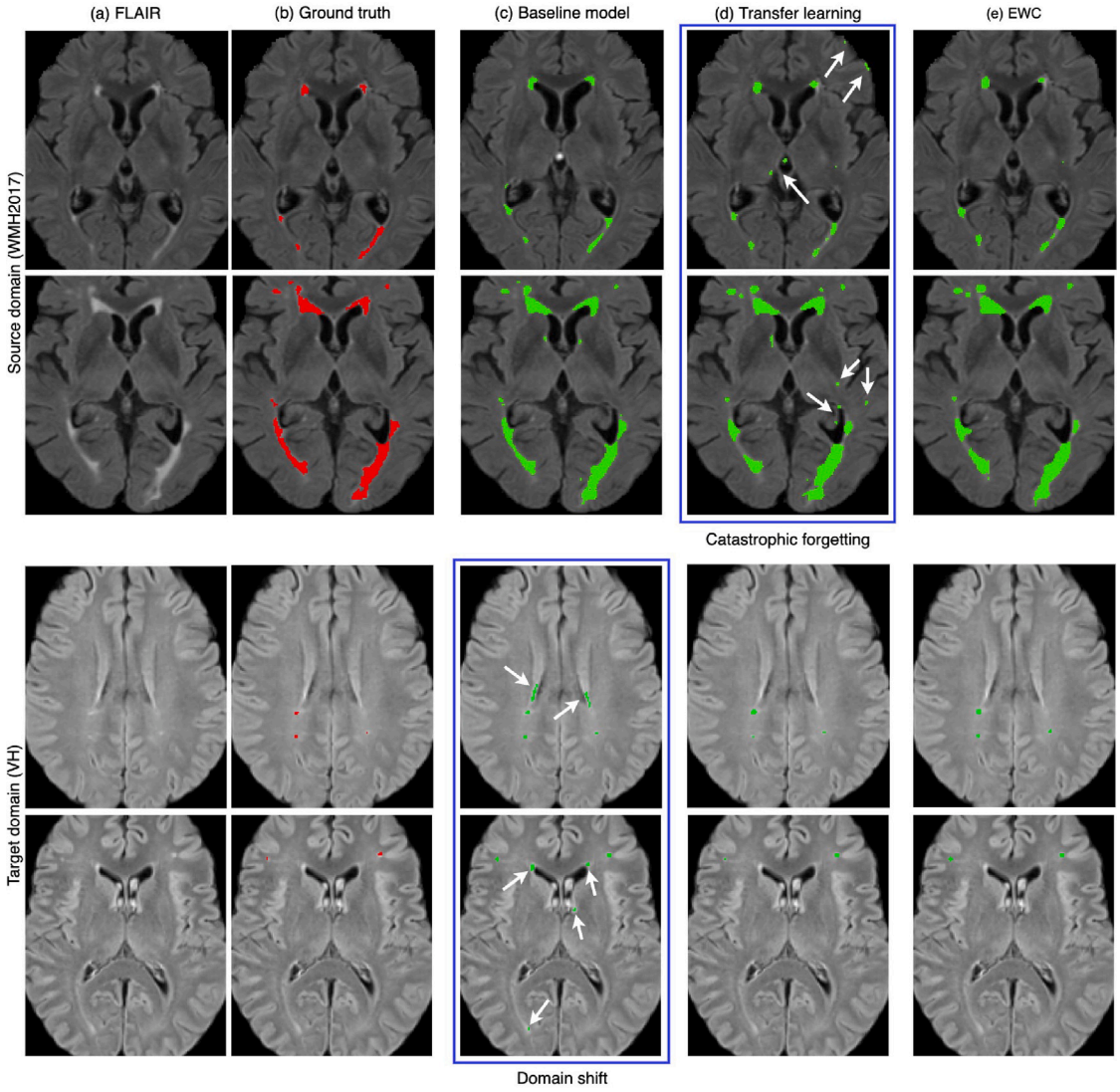


Fig. 7. Comparison of TL and CL segmentation results on the source and target domain. From top to bottom, the source domain subjects have 4.316 ml and 48.714 ml of lesion load respectively, while the target domain sample subjects have 0.966 ml and 0.517 ml. (a) FLAIR image. (b) Ground truth. (c) **Baseline model**: effective on the source domain but it fails to generalize to the target domain (domain shift). (d) **TL** achieves successful adaptation to the target domain, but suffers from catastrophic forgetting on the source domain. (e) **EWC** shows good performance in both the source and target domains.

demonstrated through internal comparison with traditional TL. However, direct comparisons with other published works are challenging since, to the best of our knowledge, this is the first application of EWC for few-shot domain adaptation in MS lesion segmentation.

The clinical relevance of this extends beyond technical improvements in domain adaptation. The strong correlations observed between estimated and ground truth measurements for both lesion volumes ($r = 0.966$) and lesion counts ($r = 0.911$) (see Fig. 4) are particularly important, as these metrics are crucial for MS diagnosis and monitoring in clinical settings. This high correlation demonstrates that our approach not only detects lesions accurately but also provides reliable volumetric quantification, which is essential for tracking disease progression and treatment response in MS patients (Filippi et al., 2022; Oship et al., 2022). Moreover, the experimental design, using an independent in-house dataset (VH) as the target domain, represents a realistic clinical scenario where models trained on public research datasets need to be adapted to local hospital environments with different scanner characteristics. Mitigating catastrophic forgetting is particularly important in clinical practice, where models may need to handle images from different scanners or even hospitals while maintaining consistent performance across all domains. Perhaps most significant for practical implementation is the EWC's few-shot learning capability, achieving good adaptation using only 3 or 5 images from the target domain. This substantially reduces the annotation burden on radiologists when deploying these models in new clinical settings, making the adaptation process feasible in resource-constrained environments.

While this work demonstrated the potential of EWC for mitigating catastrophic forgetting in domain adaptation scenarios for MS lesion segmentation, some limitations need to be addressed in future studies. A U-Net architecture was deliberately employed to isolate the effects of the CL and TL methods without architectural confounds, still achieving state-of-the-art results. Future work should explore how EWC performs when integrated with more advanced segmentation models, such as nn-Unet or attention-based networks (Basaran et al., 2022; Rondinella et al., 2023), potentially yielding even better absolute performance while maintaining the relative improvements in domain adaptation and knowledge preservation demonstrated in this study. Additionally, studying how sensitive the one-shot and few-shot learning results are to the selection of the images would provide a more robust understanding of the generalizability of the approach. It would also be interesting to apply EWC sequentially to different datasets and study how does the performance on the source domain degrade after learning several

datasets. Furthermore, exploring the application of EWC to longitudinal MS lesion segmentation represents another promising direction for future research (Valverde et al., 2017; Salem et al., 2020; Commowick et al., 2021). EWC could potentially enhance automated longitudinal analyses, supporting more accurate monitoring of disease progression and treatment response over time.

Future directions for MS lesion segmentation could explore complementary approaches to EWC for addressing domain shift challenges. Image harmonization techniques (Fortin et al., 2018; Pomponio et al., 2020; van Nederpelt et al., 2024), which aim to standardize images from different scanners or acquisition protocols to a common appearance, could be integrated with our EWC-based continuous learning method. Image harmonization techniques have proven to be effective at standardizing MRI data across different scanners, while preserving important morphological information. It would be valuable to investigate whether combining EWC with image harmonization could yield even better results, or whether effective harmonization might reduce the need for continuous learning by improving initial generalizability. The optimal approach might involve a hybrid solution that leverages both preprocessing techniques (harmonization) and model adaptation strategies (continuous learning) to achieve robust performance across diverse clinical environments.

5. Conclusions

This work was the first to apply EWC for domain-incremental learning in MS lesion segmentation. Incorporating EWC into a deep learning framework based on a 3D U-Net architecture has shown successful results in mitigating catastrophic forgetting suffered by TL methods in a domain-incremental learning scenario. This makes deep learning models more generalizable in different hospital environments, which makes it an interesting proposal for future clinical applications. Moreover, few-shot approaches demonstrated that EWC achieved domain adaptation with a reduced number of training images, which is of interest for real-life applications. All the methods were evaluated in several public international and in-house datasets.

The analysis of the baseline model revealed a significant drop in performance when testing the source model on the target domain, due to the domain-shift, with an average F-score decrease of around 14%. While both EWC and TL achieved full-domain adaptation with only 5 target images, using only 3 images led to an F-score improvement of almost 10%. Results showed that the source model's performance drop during adaptation to the target domain using TL (catastrophic forgetting) ranged from 20% to 37% in terms of F-score. EWC successfully reduced catastrophic forgetting by 8% to 19% across different training set sizes. Moreover, EWC demonstrated performance on the target domain comparable to that of TL techniques, in both few-shot and full-training settings, confirming that EWC does not hinder the model's ability to adapt to new domains.

Finally, a significant advantage of EWC is its efficiency. It enables adaptation without requiring source domain images during target domain training. This not only translates to memory efficiency but also addresses critical privacy concerns within the medical domain. Additionally, EWC avoids introducing new network parameters, eliminates the need for domain labels during inference, and does not require training separate neural networks. These characteristics make EWC a promising approach for real-world applications.

CRediT authorship contribution statement

Luisana Álvarez: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sergi Valverde:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization. **Àlex Rovira:** Writing – review & editing, Validation, Resources, Data curation. **Xavier Lladó:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization.

Acknowledgments

This work has been supported by DPI2020-114769RB-I00 and PID2023-146187OB-I00 from the Ministerio de Ciencia, Innovación y Universidades, Spain and also by the ICREA Academia program. This work was carried out in collaboration with The Observatoire Français de la Sclérose en Plaques (OFSEP), who is supported by a grant provided by the French State and handled by the "Agence Nationale de la Recherche", within the framework of the "Investments for the Future" program, under the reference ANR-10-COHO-002, by the Eugène Devic EDMUS Foundation against multiple sclerosis and by the ARSEP Foundation.

Data availability

The authors have chosen not to share the code.

References

Basaran, B.D., Matthews, P.M., Bai, W., 2022. New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation. Front. Neurosci. 16, http://dx.doi.org/10.3389/fnins.2022.1007453.

Baweja, C., Glocker, B., Kamnitsas, K., 2018. Towards continual learning in medical imaging. arXiv:1811.02496.

Bayasi, N., Hamarneh, G., Garbi, R., 2021. Culprit-prune-net: Efficient continual sequential multi-domain learning with application to skin lesion classification. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Springer International Publishing, pp. 165–175. http://dx.doi.org/10.1007/978-3-030-87234-2_16.

Chen, S., Tang, F., 2022. Breast cancer detection model training strategy based on continual learning. In: CAIBDA 2022; 2nd International Conference on Artificial Intelligence, Big Data and Algorithms. pp. 1–5.

Chertcoff, A., Schneider, R., Azevedo, C.J., Sicotte, N., Oh, J., 2024. Recent advances in diagnostic, prognostic, and disease-monitoring biomarkers in multiple sclerosis. Neurol. Clin. 42, 15–38. http://dx.doi.org/10.1016/j.ncl.2023.06.008, multiple Sclerosis.

Çiçek, Ö, Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D u-net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Springer International Publishing, pp. 424–432. http://dx.doi.org/10.1007/978-3-319-46723-8_49.

Commowick, O., Cervenansky, F., Cotton, F., Dojat, M., 2021. Msseg-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure. In: MICCAI 2021-24th International Conference on Medical Image Computing and Computer Assisted Intervention. Strasbourg, France.

Compston, A., Coles, A., 2008. Multiple sclerosis. Lancet 372, 1502–1517. http://dx.doi.org/10.1016/S0140-6736(08)61620-7.

Coupe, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., Barillot, C., 2008. An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images. IEEE Trans. Med. Imaging 27, 425–441. http://dx.doi.org/10.1109/TMI.2007.906087.

Fenneteau, A., Bourdon, P., Helbert, D., Fernandez-Maloigne, C., Habas, C.N., Guillevin, R., 2021. Investigating efficient CNN architecture for multiple sclerosis lesion segmentation. J. Med. Imaging 8, http://dx.doi.org/10.1117/1.JMI.8.1.014504.

Filippi, M., Preziosa, P., Meani, A., Dalla Costa, G., Mesaros, S., Drulovic, J., Ivanovic, J., Rovira, A., Tintorè, M., Montalban, X., Ciccarelli, O., Brownlee, W., Miszkiel, K., Enzinger, C., Khalil, M., Barkhof, F., Strijbis, E.M., Frederiksen, J.L., Cramer, S.P., Fainardi, E., Amato, M.P., Gasperini, C., Ruggieri, S., Martinelli, V., Comi, G., Rocca, M.A., on behalf of the MAGNIMS Study Group, Stefano, N.D., Palace, J., Kappos, L., Sastre-Garriga, J., Yousry, T., 2022. Performance of the 2017 and 2010 revised mcdonald criteria in predicting ms diagnosis after a clinically isolated syndrome: A magnims study. Neurology 98, http://dx.doi.org/10.1212/WNL.0000000000013016.

Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., 2011. Unbiased average age-appropriate atlases for pediatric studies. NeuroImage 54, 313–327. http://dx.doi.org/10.1016/j.neuroimage.2010.07.033.

Fonov, V., Evans, A., McKinstry, R., Almli, C., Collins, D., 2009. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage 47, 102. http://dx.doi.org/10.1016/S1053-8119(09)70884-5, organization for Human Brain Mapping 2009 Annual Meeting.

Fortin, J.P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., McInnis, M., Phillips, M.L., Trivedi, M.H., Weissman, M.M., Shinohara, R.T., 2018. Harmonization of cortical thickness measurements across scanners and sites. NeuroImage 167, 104–120. http://dx.doi.org/10.1016/j.neuroimage.2017.11.024.

Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Guttmann, C.R.G., de Leeuw, F.E., Tempany, C.M., Ginneken, B.van., Fedorov, A., Abolmaesumi, P., Platel, B., Wells, W.M., 2017. Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation. Springer International Publishing, pp. 516–524. http://dx.doi.org/10.1007/978-3-319-66179-7_59.

Greselin, M., Lu, P.J., Melie-Garcia, L., Ocampo-Pineda, M., Galbusera, R., Cagol, A., Weigel, M., Siebenborn, N.de.Oliveira., Ruberte, E., Benkert, P., Müller, S., Finkener, S., Vehoff, J., Disanto, G., Findling, O., Chan, A., Salmen, A., Pot, C., Bridel, C., Zecca, C., Derfuss, T., Lieb, J.M., Diepers, M., Wagner, F., Vargas, M.I., Pasquier, R.D., Lalive, P.H., Pravatà, E., Weber, J., Gobbi, C., Leppert, D., Kim, O.C.H., Cattin, P.C., Hoepner, R., Roth, P., Kappos, L., Kuhle, J., Granziera, C., 2024. Contrast-enhancing lesion segmentation in multiple sclerosis: A deep learning approach validated in a multicentric cohort. Bioengineering 11, http://dx.doi.org/10.3390/bioengineering11080858.

Guan, H., Liu, M., 2022. Domain adaptation for medical image analysis: A survey. IEEE Trans. Biomed. Eng. 69, 1173–1185. http://dx.doi.org/10.1109/TBME.2021.3117407.

Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H., Heiland, S., Wick, W., Bendszus, M., Maier-Hein, K.H., Kickingereder, P., 2019. Automated brain extraction of multisequence mri using artificial neural networks. Hum. Brain Mapp. 40, 4952–4964. http://dx.doi.org/10.1002/hbm.24750.

Karimi, D., Warfield, S.K., Gholipour, A., 2021. Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations. Artif. Intell. Med. 116, 102078. http://dx.doi.org/10.1016/j.artmed.2021.102078.

Karthik, E.N., Kerbrat, A., Labauge, P., Granberg, T., Talbott, J., Reich, D.S., Filippi, M., Bakshi, R., Callot, V., Chandar, S., Cohen-Adad, J., 2022. Segmentation of multiple sclerosis lesions across hospitals: Learn continually or train from scratch?. arXiv:2210.15091.

Kirkpatrick, James, Pascanu, Razvan, Rabinowitz, Neil, Veness, Joel, Desjardins, Guillaume, Rusu, Milan, Kieran, Quan, John, Ramalho, Tiago, Grabska-Barwinska, Agnieszka, Hassabis, Demis, Clopath, Claudia, Kumaran, Dharshan, Hadsell, Raia, 2017. Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. 114, 3521–3526. http://dx.doi.org/10.1073/pnas.1611835114.

Krishnan, A.P., Song, Z., Clayton, D., Jia, X., Crespigny, A.de., Carano, R.A.D., 2023. Multi-arm U-Net with dense input and skip connectivity for T2 lesion segmentation in clinical trials of multiple sclerosis. Sci. Rep. 13, 4102. http://dx.doi.org/10.1038/s41598-023-31207-5.

Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., Collins, D.L., Dadar, M., Georgiou, A., Ghafoorian, M., Jin, D., Khademi, A., Knight, J., Li, H., Lladó, X., Luna, M., Mahmood, Q., McKinley, R., Mehrtash, A., Ourselin, S., Park, B.Y., Park, H., Park, S.H., Pezold, S., Puybareau, E., Rittner, L., Sudre, C.H., Valverde, S., Vilaplana, V., Wiest, R., Xu, Y., Xu, Z., Zeng, G., Zhang, J., Zheng, G., Chen, C., Flier, W.van.der., Barkhof, F., Viergever, M.A., Biessels, G.J., 2019. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE Trans. Med. Imaging 38, 2556–2568. http://dx.doi.org/10.1109/TMI.2019.2905770.

Li, Z., Hoiem, D., 2018. Learning without forgetting. IEEE Trans. Pattern Anal. Mach. Intell. 40, 2935–2947. http://dx.doi.org/10.1109/TPAMI.2017.2773081.

Li, K., Yu, L., Heng, P.A., 2023. Domain-incremental cardiac image segmentation with style-oriented replay and domain-sensitive feature whitening. IEEE Trans. Med. Imaging 42, 570–581. http://dx.doi.org/10.1109/TMI.2022.3211195.

Lladó, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J.C., Quiles, A., Valls, L., Ramió-Torrentà, L., Rovira, Àlex, 2012. Segmentation of multiple sclerosis lesions in brain mri: A review of automated approaches. Inform. Sci. 186, 164–185. http://dx.doi.org/10.1016/j.ins.2011.10.011.

Malinin, A., Athanasopoulos, A., Barakovic, M., Cuadra, M.B., Gales, M.J.F., Granziera, C., Graziani, M., Kartashev, N., Kyriakopoulos, K., Lu, P.J., Molchanova, N., Nikitakis, A., Raina, V., Rosa, F.L., Sivena, E., Tsarsitalidis, V., Tsompopoulou, E., Volf, E., 2022. Shifts 2.0: Extending the dataset of real distributional shifts. arXiv:2206.15407.

Oship, D., Jakimovski, D., Bergsland, N., Horakova, D., Uher, T., Vaneckova, M., Havrdova, E., Dwyer, M.G., Zivadinov, R., 2022. Assessment of T2 lesion-based disease activity volume outcomes in predicting disease progression in multiple sclerosis over 10 years. Mult. Scler. Relat. Disord. 67, http://dx.doi.org/10.1016/j.msard.2022.104187.

Perkonigg, M., Hofmanninger, J., Herold, C.J., Brink, J.A., Pianykh, O., Prosch, H., Langs, G., 2021. Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. Nat. Commun. 12, 5678. http://dx.doi.org/10.1038/s41467-021-25858-z.

Pianykh, O.S., Langs, G., Dewey, M., Enzmann, D.R., Herold, C.J., Schoenberg, S.O., Brink, J.A., 2020. Continuous learning ai in radiology: Implementation principles and early applications. Radiology 297, 6–14. http://dx.doi.org/10.1148/radiol.2020200038.

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I.M., Satterthwaite, T.D., Fan, Y., Launer, L.J., Masters, C.L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S.C., Fripp, J., Koutsouleris, N., Wolf, D.H., Gur, R., Gur, R., Morris, J., Albert, M.S., Grabe, H.J., Resnick, S.M., Bryan, R.N., Wolk, D.A., Shinohara, R.T., Shou, H., Davatzikos, C., 2020. Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage 208, 116450. http://dx.doi.org/10.1016/j.neuroimage.2019.116450.

Rondinella, A., Crispino, E., Guarnera, F., Giudice, O., Ortis, A., Russo, G., Di Lorenzo, D., Pappalardo, F., Battiato, S., 2023. Boosting multiple sclerosis lesion segmentation through attention mechanism. Comput. Biol. Med. 161, 107021. http://dx.doi.org/10.1016/j.compbiomed.2023.107021.

Roura, E., Oliver, A., Cabezas, M., Valverde, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, A., Lladó, X., 2015. A toolbox for multiple sclerosis lesion segmentation. Neuroradiology 57, 1031–1043. http://dx.doi.org/10.1007/s00234-015-1552-2.

Salem, M., Cabezas, M., Valverde, S., Pareto, D., Oliver, A., Salvi, J., Rovira, Àlex, Lladó, X., 2018. A supervised framework with intensity subtraction and deformation field features for the detection of new t2-w lesions in multiple sclerosis. NeuroImage: Clin. 17, 607–615. http://dx.doi.org/10.1016/j.nicl.2017.11.015.

Salem, M., Ryan, M.A., Oliver, A., Hussain, K.F., Lladó, X., 2022. Improving the detection of new lesions in multiple sclerosis with a cascaded 3d fully convolutional neural network approach. Front. Neurosci. 16, 1007619. http://dx.doi.org/10.3389/fnins.2022.1007619.

Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., Rovira, Àlex, Lladó, X., 2020. A fully convolutional neural network for new t2-w lesion detection in multiple sclerosis. NeuroImage: Clin. 25, 102149. http://dx.doi.org/10.1016/j.nicl.2019.102149.

Srivastava, S., Yaqub, M., Nandakumar, K., Ge, Z., Mahapatra, D., 2021. Continual domain incremental learning for chest x-ray classification in low-resource clinical settings. In: Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health. Springer International Publishing, pp. 226–238. http://dx.doi.org/10.1007/978-3-030-87722-4_21.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4itk: Improved n3 bias correction. IEEE Trans. Med. Imaging 29, 1310–1320. http://dx.doi.org/10.1109/TMI.2010.2046908.

Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, A., Oliver, A., Lladó, X., 2017. Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. NeuroImage 155, 159–168. http://dx.doi.org/10.1016/j.neuroimage.2017.04.034.

Valverde, S., Salem, M., Cabezas, M., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Salvi, J., Rovira, Àlex, Oliver, A., Lladó, X., 2019. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. NeuroImage: Clin. 21, 101638. http://dx.doi.org/10.1016/j.nicl.2018.101638.

van Garderen, K., van der Voort, S., Incekara, F., Smits, M., Klein, S., 2019. Towards continuous learning for glioma segmentation with elastic weight consolidation. arXiv:1909.11479.

van Nederpelt, D.R., Pontillo, G., Barrantes-Cepas, M., Brouwer, I., Strijbis, E.M., Schoonheim, M.M., Moraal, B., Jasperse, B., Mutsaerts, H.J.M., Killestein, J., Barkhof, F., Kuijer, J.P., Vrenken, H., 2024. Scanner-specific optimisation of automated lesion segmentation in ms. NeuroImage: Clin. 44, 103680. http://dx.doi.org/10.1016/j.nicl.2024.103680.

Wahlig, S.G., Nedelec, P., Weiss, D.A., Rudie, J.D., Sugrue, L.P., Rauschecker, A.M., 2023. 3D u-net for automated detection of multiple sclerosis lesions: utility of transfer learning from other pathologies. Front. Neurosci. 17, http://dx.doi.org/10.3389/fnins.2023.1188336.

Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R.I., O'Brien, J.T., Barkhof, F., Benavente, O.R., Black, S.E., Brayne, C., Breteler, M., Chabriat, H., DeCarli, C., de Leeuw, F.E., Doubal, F., Duering, M., Fox, N.C., Greenberg, S., Hachinski, V., Kilimann, I., Mok, V., Oostenbrugge, R.v., Pantoni, L., Speck, O., Stephan, B.C.M., Teipel, S., Viswanathan, A., Werring, D., Chen, C., Smith, C., van Buchem, M., Norrving, B., Gorelick, P.B., Dichgans, M., 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol. 12, 822–838. http://dx.doi.org/10.1016/S1474-4422(13)70124-8.

Wiltgen, T., McGinnis, J., Schlaeger, S., Kofler, F., Voon, C., Berthele, A., Bischl, D., Grundl, L., Will, N., Metz, M., Schinz, D., Sepp, D., Prucker, P., Schmitz-Koep, B., Zimmer, C., Menze, B., Rueckert, D., Hemmer, B., Kirschke, J., Mühlau, M., Wiestler, B., 2024. Lst-ai: A deep learning ensemble for accurate ms lesion segmentation. NeuroImage: Clin. 42, 103611. http://dx.doi.org/10.1016/j.nicl.2024.103611.

Yan, S., Xie, J., He, X., 2021. Der: Dynamically expandable representation for class incremental learning. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE Computer Society, Los Alamitos, CA, USA, pp. 3013–3022. http://dx.doi.org/10.1109/CVPR46437.2021.00303.