DEEPLESIONBRAIN: TOWARDS A BROADER DEEP-LEARNING GENERALIZATION FOR MULTIPLE SCLEROSIS LESION SEGMENTATION

ARXIV VERSION

Reda Abdellah Kamraoui[1], Vinh-Thong Ta[1], Thomas Tourdias[2],
Boris Mansencal[1], José V Manjon[3], and Pierrick Coupé[1]

[1]Univ. Bordeaux, Bordeaux INP, CNRS, LaBRI,, UMR5800, PICTURA, F-33400 Talence, France
[2]Univ. Bordeaux, INSERM, Neurocentre Magendie,, U1215, F-3300 Bordeaux, France,
[3]ITACA, Universitat Politècnica de València,, 46022 Valencia, Spain,

June 13, 2022

ABSTRACT

Recently, segmentation methods based on Convolutional Neural Networks (CNNs) showed promising performance in automatic Multiple Sclerosis (MS) lesions segmentation. These techniques have even outperformed human experts in controlled evaluation conditions such as Longitudinal MS Lesion Segmentation Challenge (ISBI Challenge). However, state-of-the-art approaches trained to perform well on highly-controlled datasets fail to generalize on clinical data from unseen datasets. Instead of proposing another improvement of the segmentation accuracy, we propose a novel method robust to domain shift and performing well on unseen datasets, called DeepLesionBrain (DLB). This generalization property results from three main contributions. First, DLB is based on a large group of compact 3D CNNs. This spatially distributed strategy aims to produce a robust prediction despite the risk of generalization failure of some individual networks. Second, we propose a hierarchical specialization learning (HSL) by pre-training a generic network over the whole brain, before using its weights as initialization to locally specialized networks. By this end, DLB learns both generic features extracted at global image level and specific features extracted at local image level. Finally, DLB includes a new image quality data augmentation to reduce dependency to training data specificity (*e.g.*, acquisition protocol). DLB generalization was validated in cross-dataset experiments on MSSEG'16, ISBI challenge, and in-house datasets. During experiments, DLB showed higher segmentation accuracy, better segmentation consistency and greater generalization performance compared to state-of-the-art methods. Therefore, DLB offers a robust framework well-suited for clinical practice.

Keywords: Multiple Sclerosis SegmentationDeep LearningDomain Generalization

1 Introduction

1.1 Problem Description

In recent years, the medical imaging community has witnessed a rapid increase in image processing methods based on Deep Learning (DL). These novel techniques came with remarkable performance in many tasks including Multiple Sclerosis (MS) lesion segmentation. Some automated algorithms have even reached human-level performance in controlled evaluation conditions (see [6]). Unlike over-controlled conditions where most DL approaches have been validated, real-world data exhibit high diversity. Consequently, clinical use of MS lesion segmentation based on DL is still limited mainly because of their poor generalization on new data coming from medical sites that have not been covered during training (unseen domains). This lack of generalization of DL methods can result from several factors such as the selected solution during optimization, the diversity of the training dataset or the genericity of the extracted features.

DL is based on the assumption that training and test data are independent but come from the same distribution. This assumption is usually not respected in medical imaging data especially for Magnetic Resonance Imaging (MRI) where acquisition protocols, MRI scanner, patient populations, and software processing may vary depending on the clinical center or the cohort. As a result of these differences of data distribution (covariate shift), a decrease in performance is observed between the training data (source domain) and the test data derived from different distributions (target domain). This is known as the domain shift.

An intuitive method to reduce this problem is to train on a wider and more heterogeneous dataset (as shown by [28]). However, this requires a large dataset annotated by experts which are rarely available and tedious to produce. Some deal with this phenomenon by applying extensive data augmentation (such as [43]). Others use few available labeled images from the target domain to reduce the covariate shift, such as few-shot and one-shot learning strategies (see [34, 37]).

Besides, DL requires the tuning of a large number of parameters relative to the number of training data samples. Thus, it usually ends up converging to one of the many possible local minima as opposed to the theoretical best parameter configuration which leads to the global minimum. Consequently, the generalization ability of the model depends on the selected solution. The selection of the best generalizing local minimum is still an open question. On one hand, some works have proposed to select it using the local characteristics (*e.g.*, flatness) of the loss function (see [23, 38]). On the other hand, an alternative strategy consists in combining several local minima to improve the generalization capability of the method. This can be done by averaging several local minima of one model (*e.g.*, snapshot ensemble [19]) or by combining outputs of different models trained independently (*e.g.*, classical ensemble [41] and spatially distributed networks [12, 20]).

Unlike classical methods that use hand-crafted features, Convolutional Neural Networks (CNNs) automatically extract the most suitable set of features for a particular task. Although this strategy is very efficient to extract relevant features for a particular source domain, this set of features may not generalize well for the target domain. Some works proposed to learn invariant features that coexist across different source domains [29, 30, 40]. They tried to apply a regularization to learn an abstract representation of the specific computer vision task (*i.e.*, just like humans understand high-level concepts). Indeed, the extraction of generalizing features lies between the freedom of the optimization process to find the optimal combination from data and the constraints used for minimizing domain bias.

The successful deployment of DL based methods for MS lesion segmentation requires generalization capabilities that can guarantee high performance for unseen domains. First, such methods should ensure the convergence of the DL model to generalizing minima. Second, the training process should anticipate the reduction of the covariate shift. Moreover, the method should be enforced to learn MS lesion generalizing features from the source domain, to effectively delineate lesions despite the target domain distribution. Finally, this solution should not require additional annotation in case of processing unseen domains.

1.2 Related Works

Recently, many works have been proposed for MS lesion segmentation using CNNs.
First, [5] proposed a deep 3D encoder-decoder network, with joint training of the encoder and the decoder. The authors used shortcut connections between the two interconnected pathways for integrating high and low-level features. This pioneering work demonstrated the high potential of deep learning for MS lesion segmentation.
[36] proposed a cascade of two patch-wise 3D CNNs, composed of a first sensitive network to reveal possible lesion candidates followed by a second network to reduce misclassified voxels. This cascade allows refined segmentation but it uses a small receptive field that prevents capturing the global context. Later, the authors [37] improved their method by proposing a one-shot domain adaptation model which uses transfer learning and partial fine-tuning. However, this domain adaptation needs a labeled example from the new domain. Moreover, such strategies lead to different versions of the method after each adaptation, this results in discrepancies in the segmentation.
[18] considered the problem of data imbalance (*i.e.*, the under-sampling of the lesion class) by using an asymmetric similarity loss function based on Tversky index to train a 3D CNN that performed better than Dice or cross-entropy measures. This result suggests that further work should be done on choosing an adequate loss function. Although the proposed loss can be tuned for the optimal trade-off between precision and recall in a particular domain, the generalization to unseen domains has to be proven.
[41] used a fully convolutional densely connected network for MS lesion segmentation. They stacked adjacent 2D slices of different modalities with a channel-wise concatenation, before forwarding this stack through a 2D CNN. The final segmentation is based on a majority vote along different orientations. While this method showed competitive performance on a well-controlled challenge, the stacking using only the two directly adjacent slices gives a weak insight

into the 3D nature of the data. Moreover, 2D features may not be considered as generalizing features when processing 3D volumes and can result in the limited generalization of the method.
[2] proposed an end-to-end encoder-decoder 2D network with multiple downsampling branches, one for each input modality, and a decoder part where features from the different modalities are put together at multiple scales. This separation in encoder branches enables the model to encode information efficiently from each modality, before combining them in a later stage. However, this 2D approach does not combine features based on axial, coronal, and sagittal orientations that may greatly reduce its generalization on 3D images.
[15] considered MRI modality unavailability during segmentation by introducing sequence dropout. This is an important point since the availability of all the modalities is not always ensured between datasets that can greatly reduce the generalization capacity of a method. This framework randomly drops specific MRI sequences during training, with the intent to learn the intrinsic information of each sequence. This technique showed it can produce acceptable segmentation even in the absence of one or two modalities. Nonetheless, it is less efficient than other state-of-the-art methods when all modalities are available (will be detailed in section 3.3).
[3] tackled the problem of generalization to new domains by integrating a regularization network to the traditional encoder-decoder network. The regularizer penalizes the network when the latter learns features that allow the prediction of MRI scanning sites. However, [25] have argued that such strategies suffer from overfitting, the obtained representation could well generalize for all the source domains but poorly for the unknown target domains.

All the cited MS methods ([5, 36, 15, 18, 41, 2, 15]) focused on obtaining accurate segmentation within a same domain evaluation. However, the use of out-of-domain datasets is essential to ensure a good evaluation of the generalization capabilities of a method. This question is right now a hot topic (see [28], and [4]) and an important recommendation from the clinical world (see [32]). Experiments using training and testing images derived from the same domain are known to be biased [32] and do not ensure generalization. Therefore, a model used in clinical conditions should produce accurate segmentation for new domain images without the need of retraining with expert segmentation on the new domain.

1.3 Proposals

In this paper, we propose DeepLesionBrain (DLB), a novel method for MS lesion segmentation robust to domain shift, validated on out-of-domain testing (cross-dataset testing).

First, we use a large group of compact 3D CNNs spatially distributed over the brain with overlapping receptive fields between regions.

By associating a distinct network with each region of the brain, the spatially distributed networks (see [12, 20]) strategy simplifies the MS lesion segmentation from a single complex task on the whole brain to multiple simpler sub-tasks on each region. Moreover, the overlapping regions ensure consistent and stable consensus.

Applied to brain segmentation, this strategy demonstrated good generalization on unseen domains (*e.g.*, child's brain or patients with Alzheimer's disease) when trained on healthy adult brains [12].

Second, to extract more relevant features for MS that may lead to better generalization, we focus on feature learning strategy. We consider that a generalizing model should learn two types of features: first global and generic features, and second local and specific features. Therefore, we propose Hierarchical Specialization Learning (HSL) to efficiently extract those features in two-steps. In the first step, a single network (the generic network) is trained on all brain regions. In the second step, each network of the spatially distributed networks is initialized with the generic network weights and specialized for a specific region of the brain.

Third, DLB is trained with a novel Image Quality Data Augmentation (IQDA) method, which mimics real-world data diversity by adding realistic alterations to the training images. As shown in the works of [42], such a type of regularization technique aims to reduce covariate shift. IQDA proposes specific augmentations that constrain task learning to be independent from source data acquisition resolution, data contrast, or data quality. Consequently, the proposed augmentation strategy enables domain shift robustness.

Finally, we propose a method using only two modalities (T1w and FLAIR) to ensure its compatibility with a large number of datasets. Most of the methods ([15, 5, 36, 41]) optimize their segmentation using T1w, FLAIR, PD, and T2 modalities. However, in clinical conditions, not all these sequences are always available. Therefore we focused our work on developing a robust approach using only two modalities.

2 Method and Material

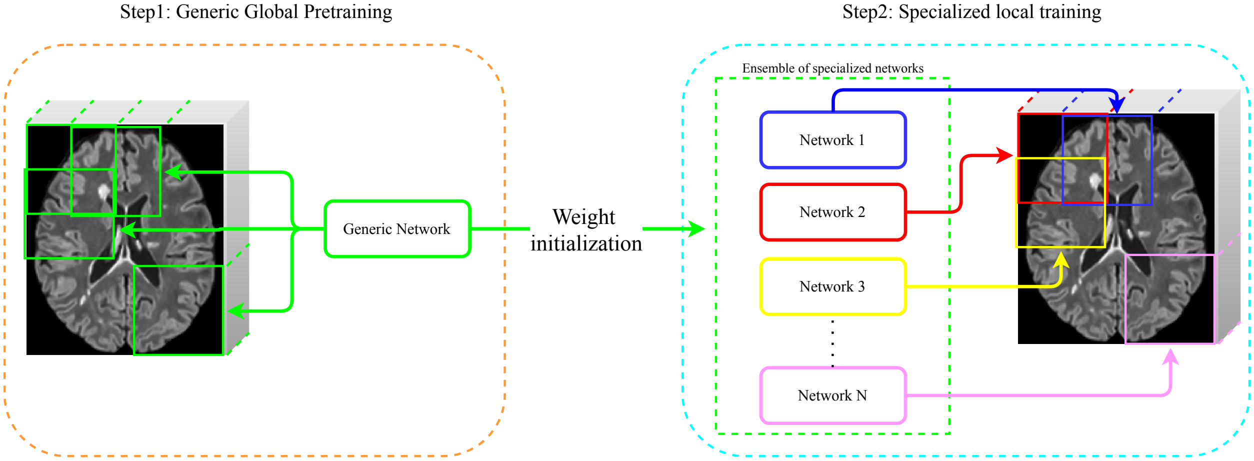
Step1: Generic Global Pretraining
Generic Network
Weight initialization
Step2: Specialized local training
Ensemble of specialized networks
Network 1
Network 2
Network 3
Network N

Figure 1: The two-steps training process of DeepLesionBrain (see Sect. 2.1.1 for more details)

2.1 Method Overview

2.1.1 Spatially Distributed Networks with Hierarchical Specialization

Spatially Distributed Networks Strategy

DLB is based on a spatially distributed networks strategy, proposed by [12] and [20]. Such a strategy uses a group of compact networks, where each network is specialized in a particular region of the brain, and processes a sub-volume of the global volume. The receptive fields of the neighboring networks overlap with one another, and the final segmentation of the whole volume is obtained with a majority vote of the local predictions. Employing our spatially distributed compact networks is equivalent to a big network with more filters and a higher receptive field (see Fig. 1). This particular configuration with overlapping receptive fields aims to produce a more robust segmentation compared to an individual network. As shown in the work of [20] that compared between 8 networks configuration with no-overlapping and 27 networks with overlapping, having an average prediction over overlapping regions led to significantly better performance on the evaluation data-sets including out-of-domain (on infants) data (*i.e.* better generalization). Similarly, [12] confirmed that larger overlapping led to better performance until the limit of 125 networks where performance peaked and stayed stable for 125, 216, and 343 networks. Additionally, we assume that averaging the prediction of a group of networks that have been trained separately on different sub-volumes is less likely to collapse than having a single network trained on the same set of data.

Hierarchical Specialization Learning (HSL)

To improve generalization for the task of MS lesion segmentation, we take inspiration from MS lesion features and propose a better learning strategy. MS lesions features can be grouped into two categories:
First, some lesion characteristics are considered generic and shared among lesion types. Such features are independent of lesion localization. They have a common and inherent significance at the global scale of the brain volume, we will refer to them in this paper as “generic global features”.
Second, other relevant features for MS lesions depend on brain structure and some distinct regions (see [16]). In this work, we refer to these features as “specialized local features”.

On the one hand, training each specialized network on a specific sub-region of the brain (see Fig. 1 right) would prevent the efficient learning of “generic global features”, since each specialized member of our group would be trained on a particular region of the whole brain. On the other hand, using a single 3D CNN to learn "specialized local features" over the whole brain volume would require a large model which may not fit into memory and a large dataset to train it.

To overcome this limitation, we propose a novel Hierarchical Specialization Learning (HSL). Fig. 1 shows our two-step learning process. First, the “generic network” is trained with data samples from all over brain regions to learn general knowledge about lesions by extracting “generic global features”. Second, each network in the spatially distributed strategy is specialized over a specific sub-volume of the brain.

The generic network is used as an initialization for each network of our spatially distributed networks, by transferring the generic network weights to each individual specialized network. The knowledge gained from this transfer learning transmits the ability to extract “generic global features”, while the specialized network training will specialize them in extracting local “specialized local features”.

In our ablation study, we will show that this hierarchical specialization learning of the specialized networks performs better than training a single network over the whole brain, or training the specialized networks without HSL.

2.1.2 Image Quality Data Augmentation (IQDA)

The quality of the MRI greatly varies between datasets. In fact, the quality of the images depends on several factors such as signal-to-noise ratio, contrast-to-noise ratio, resolution, or slice thickness. To address this issue, we propose a data augmentation strategy that considers image quality disparity. During training, we simulate “on the fly” altered versions of 3D patches. We randomly introduce at each iteration either blur, edge enhancement, or axial subsampling distortion (2D FLAIR are usually acquired along the axial direction). For the blur, a gaussian kernel is used with a randomly selected standard deviation ranging between $[0.5, 1.75]$. For edge enhancement, we use unsharp masking with the inverse of the blur filter. For axial subsampling distortion, we simulate acquisition artifacts that can result from the varying slice thickness. We use a uniform filter (a.k.a mean filter) on the axial direction with a size of $[1 \times 1 \times sz]$ where $sz \in 2, 3, 4$. Ground truth is kept the same as the original version. This process reduces the domain bias when learning to extract relevant features caused by data variability.

2.1.3 Selection of the Required Modalities

To use a trained model for MS lesion segmentation with optimal performance, it usually requires the use of the same set of modalities that have been chosen during training. DLB proposes a method that needs only T1w and FLAIR sequences to be compatible with all benchmark MS datasets and most already available MS patients data.

Our method is built with the purpose to generalize on unseen datasets, thus it uses the minimum necessary modalities. Indeed, increasing the number of sequences requires a longer scan acquisition time. Besides, it needs more complex processing which may be prone to error, such as multimodal image registration. Furthermore, the use of more sequences during the training on a dataset may reduce the generalization to other image domains.

In addition to the wide use of T1w and FLAIR for MS diagnosis, the choice of these modalities has also been motivated by the fact that FLAIR is the most relevant sequence for revealing most MS lesions (see [31]), while T1w can provide complementary structural information needed for accurate segmentation.

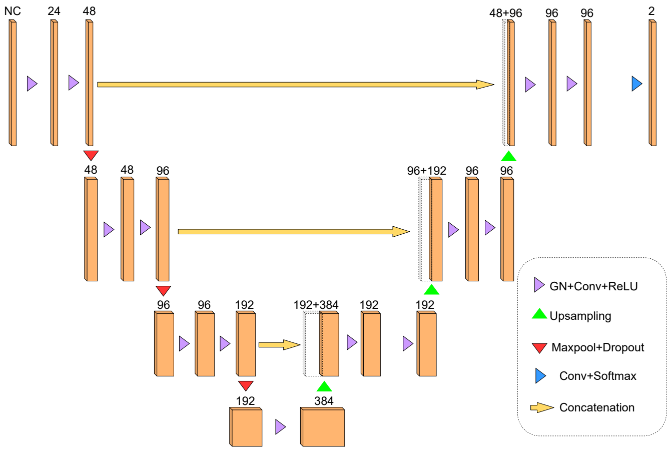

Figure 2: Illustration of the considered U-Net architecture. The number of input channels (NC) depends on the modality number (*i.e.*, NC= 2, for using T1w and FLAIR). Each block is composed of group normalization (GN), Convolution (Conv) and Rectified Linear Unit (ReLU) activation.

2.2 Implementation Details

The network architecture used in our spatially distributed strategy is based on 3D U-Net composed of a downsampling part and an upsampling one, linked with one another by skip connections at the multiple scales. This 3D CNN architecture, shown in Figure 2, has been used for all the networks in our approach. Dropout with 0.5 rate is used after max-pooling layers to prevent the overfitting of our model to the training data. Due to GPU memory constraints, we trained with a batch size of 1, and so we used Group Normalization (GN) [39] with 8 groups before each convolution. We have chosen Rectified Linear Units (ReLu) to introduce non-linearity after convolution layers. DLB is optimized with Adam [24] using a learning rate of 0.0001 and a momentum of 0.9. The experiments have been performed with Keras 2.2.4 [8] and Tensorflow 1.12.0 [1] on NVIDIA Titan Xp 12 GB GPU.

2.3 Method Description

To obtain sub-volumes for each image, we first divide our whole MRI registered into the MNI space into multiple overlapping 3D patches. We perform a cropping operation over the whole image using a sliding window of the size (Px, Py, Pz), and a stride of (Sx, Sy, Sz). We take $Sx < Px$, $Sy < Py$, $Sz < Pz$ to ensure the overlapping.

Sub-volumes from different images, that represent the same receptive field into the MNI space (the same sub-volume region of the whole volume), are grouped together. They are used for training a network specialized for that particular region.

In this work, we explored many combinations of sub-volume sizes and numbers. We chose a configuration with 125 sub-volumes by taking experimentally $Px = Py = Pz = 96$, $Sx = Sy = 76$, and $Sz = 67$ as a good trade-off between the overall performance and computation resources.

2.3.1 Symmetrical Training

To limit redundant training and to use the most possible data for a particular brain region, we choose to train specialized networks only on one hemisphere. By flipping (mirroring) the sub-volumes of the second hemisphere, it is possible to train a single specialized network for both sides. Thus, we can use twice the amount of data for each region while reducing the number of networks to train to nearly a half (due to sub-volume overlapping, the plane of symmetry cuts through the median networks which cover equivalent symmetrical regions from both hemispheres). Consequently, unlike previous works with spatially distributed networks (*i.e.*, [12]), instead of using 125 networks we use only 75 specialized networks. We experimentally verified that using 125 networks or only 75 trained with twice the number of patches, produced similar segmentation accuracy.

2.3.2 Loss Function

MS lesion segmentation task suffers from class imbalance since lesion volume is considerably low compared to healthy volume. Thus, we use a smooth version of the Generalized Jaccard Loss (GJL), which considers this issue [27].

$$GJL = 1 - \frac{\sigma + \sum_{c=1}^{Nc} w_c \sum_{i=1}^{N} p_{ci} t_{ci}}{\sigma + \sum_{c=1}^{Nc} w_c \left(\sum_{i=1}^{N} (p_{ci} + t_{ci}) - \sum_{i=1}^{N} p_{ci} t_{ci} \right)} \tag{1}$$

Where $w_c = 1/(1 + \sum_{i=1}^{N} t_{ci})$, σ is the smoothness factor, N is the number of voxels, Nc is the number of classes, p_{ci} and t_{ci} are respectively the predicted probability and the ground truth probability of the voxel i for the class c.

During inference, we combine the overlapping predictions in a straightforward majority vote technique. The class of each voxel (either lesion or healthy tissue) is chosen based on the most predicted class among the networks which cover that voxel.

2.4 Datasets

To assess the robustness of a model, it is crucial to test its ability to generalize on unseen domains. Therefore, DLB has been trained and validated using different datasets to assess its domain generalization ability (see 3.2). These datasets exhibit high heterogeneity in terms of resolution, data processing, acquisition sites, delineation protocols, and they also cover a large variety of clinical scenarios.

2.4.1 ISBI Longitudinal Multiple Sclerosis Lesion

The ISBI dataset [6] consists of five subjects for training, fourteen subjects for testing, with a mean of 4.4 time-points per subject (21 images for training and 61 images for testing). Two human expert raters delineated MS lesions, from the four available modalities acquired on 3.0 Tesla MRI scanner: 3D MPRAGE T_1-weighted (T1w) of $0.82 \times 0.82 \times 1.17$ mm^3 voxel size, 2D T_2-weighted (T2), 2D T_2-weighted fluid attenuated inversion recovery (FLAIR), and 2D Proton Density weighted (PD), of $0.82 \times 0.82 \times 2.2$ mm^3 voxel size each.

For the training, we used the ISBI training dataset with available annotations from the two experts. For test and evaluation, we segmented the test data with no available expert annotation, and submitted our results to the ISBI challenge website[1]. The ISBI pipeline already included preprocessing. Each first time-point T1w was inhomogeneity-corrected using N4 [35], skull-stripped [7], dura stripped [33], followed by a second N4 inhomogeneity correction, and rigid registration to a 1 mm^3 isotropic MNI template. Then, this image was used as a target for the remaining T1w time-points and all modalities for the same subject. These images were N4 corrected and then rigidly registered to the T1w in the MNI space. The skull and dura-stripped mask from the target T1w was applied, which were then N4 corrected again. We added an intensity normalization step using kernel density estimation for all images.

2.4.2 MICCAI2016 MS Challenge Dataset

The MSSEG'16 training dataset [9] contains 15 patients from 3 different clinical sites. Five modalities are available for each patient: 3D FLAIR, 3D T1w, 3D T1w GADO, 2D PD, and 2D T2. The images were acquired on 1.5T and 3T MRI scanners with multiple resolutions: 3D FLAIR modalities ranging from $1 \times 0.5 \times 0.5$ to $1.25 \times 1.04 \times 1.04$ mm^3, and 3D T1w sequences between $0.85 \times 0.74 \times 0.74$ and $1.08 \times 1.08 \times 0.9$ mm^3.

Seven human experts have manually segmented the multiple sclerosis lesions. Each patient modalities have been preprocessed with the same pipeline. First, each sequence was denoised using the non local means algorithm [14]. Second, a rigid registration of each modality on the FLAIR was performed [11]. Then, skull stripping of T1w was performed using the volBrain platform [26], the same mask is applied to other modalities. Finally, bias field correction was applied using the N4 algorithm [35]. In addition to these steps that have been performed on the available images, each modality was registered to the MNI space for our experiments. Similarly, to the ISBI images, we used kernel density estimation for the normalization step.

2.4.3 In-house Dataset

For further evaluation of our approach, we used an In-house 3D MRI dataset, with 3D T1w and 3D FLAIR modalities [13]. This dataset contains 43 subjects diagnosed with MS. The images were acquired with different scanners and multiple resolutions ($0.6 \times 0.6 \times 0.65$ mm^3, $0.5 \times 0.5 \times 0.9$ mm^3, and $1 \times 1 \times 1$ mm^3).

The dataset lesion masks have been obtained by human experts manual delineation. The images were pre-processed using the lesionBrain pipeline from the volBrain platform [26]. First, it included denoising of each modality [14]. Second, an affine registration to MNI space was performed on the T1w , then the FLAIR was registered to the transformed T1w. Skull stripping and bias correction have been performed on the modalities, followed by a second denoising. Finally, the intensities have been normalized.

Table 1: Description of datasets used in this work.

	FLAIR resolution	Site	# Subjects	# Raters	Modalities
ISBI train-set	2D	Mono	5, multiple time-points	2	T1w, FLAIR, T2, PD
MSSEG'16	3D	Multi	15	7	T1w, T1w GADO, FLAIR, PD, T2
In-house	3D	Multi	43	1	T1w, FLAIR

2.4.4 Datasets Summary

Table 1 summarizes the main differences between the 3 datasets. We focused specifically on the resolution of FLAIR due to its known relevance in MS lesion segmentation. To summarize, ISBI train-set contains multiple time points of only five subjects, acquired in a single clinical site with two human expert segmentations. Except for 3D MPRAGE T1w, the other three modalities are in 2D. MSSEG'16 dataset is a multi-site database comprising 15 patients, with seven human segmentations. This dataset contains 5 available modalities with 3D FLAIR. Finally, the In-house dataset is the

[1] https://smart-stats-tools.org/lesion-challenge-upload-results

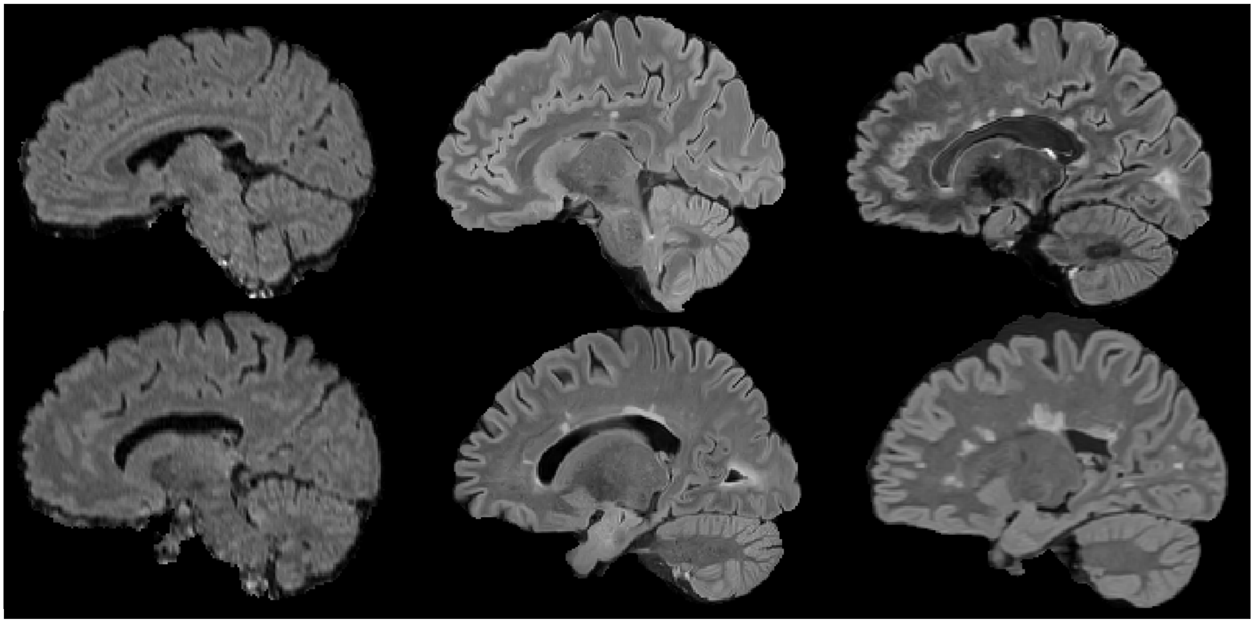

Figure 3: FLAIR examples from the considered three datasets in the MNI space and after intensity normalization. From left to right, the two images are from ISBI, MSSEG'16, our in-house data, respectively.

largest dataset with 43 patients, and multi-site 3D modalities, segmented by a single human rater and validated by a second one.

Figure 3 shows examples from the three presented datasets, each image represents a sagittal section of the FLAIR modality in the MNI space after intensity normalization. The two images on the left are examples from the ISBI dataset. We notice blurring effects which makes it hard to distinguish precisely brain structures. This blur comes from 2D low-resolution acquisitions. In the middle, the two examples come from the MSSEG'16 dataset. These 3D FLAIR are noticeably of higher resolution than the other images. Therefore, lesion boundaries are more easily delineated and main structures are apparent. The final two images on the right are from our In-house 3D dataset. The 3D resolution enables the differentiation of white matter, gray matter, and shows the lesions clearly. In terms of FLAIR images, we notice that both MSSEG'16 and In-house dataset (3D FLAIR) propose better visual quality than ISBI dataset (2D FLAIR).

2.5 Validation Framework

2.5.1 Evaluation Metrics

The assessment of a segmentation method is usually measured by a similarity metric between the predicted segmentation and the human expert ground truth.

First, we use several complementary metrics to assess segmentation performance. Namely, we use the Dice similarity coefficient, the Positive Predictive Value (PPV or the precision), true positive rate (TPR, known as recall or Sensitivity), and Pearson's correlation coefficient (CORR).

$$PPV = \frac{TP}{TP + FP}, \qquad TPR = \frac{TP}{TP + FN}, \tag{2}$$

$$Dice = \frac{2 \times TP}{(TP + FN) + (TP + FP)}, \tag{3}$$

where TP, FN, FP represent respectively true positives, false negatives, and false positives.

Second, recent works (*i.e.*, [10]) question the relevance of classic metrics (Dice) compared to detection metrics, which are used for MS diagnostic and clinical evaluation of the patient evolution. Thus, in addition to the voxel-wise metrics, we also use lesion-wise metrics that focus on the lesion count. Such as Lesion False Positive Rate (LFPR) and Lesion True Positive Rate (LTPR).

$$LTPR = \frac{LTP}{LTP + LFN}\,, \qquad LFPR = \frac{LFP}{LTP + LTN}\,, \tag{4}$$

where LTP, LFN, LFP represent respectively lesion true positives, lesion false negatives, and lesion false positives.

Moreover, [17] pointed out that even though Dice is commonly used and simplifies method comparison, multiple complementary metrics are needed to provide a better understanding of the performance. Recently, international challenges took into consideration several metrics ([6] and [10]). Consequently, we decided to evaluate our methods using Hybrid score proposed by [6]. This metric combines voxel-wise segmentation, lesion-wise detection, and volumetric metrics. It is defined as:

$$Hybrid = \frac{Dice}{8} + \frac{PPV}{8} + \frac{(1 - LFPR)}{4} + \frac{LTPR}{4} + \frac{CORR}{4} \tag{5}$$

Finally, we also use the ISBI Submission Score for the evaluation of ISBI test-set segmentations. [6] defined it as the average of the hybrid scores of all image examples with the different human raters and with inter-rater variability taken into consideration. This score is computed after submitting the segmentation to ISBI's challenge website [2]. Obtaining an ISBI score of 90 or higher with a segmentation technique indicates that this method is similar to the human raters.

2.5.2 Reference Methods

During experiments, our method was compared to three publicly available state-of-the-art approaches. We performed training and validation for all three compared methods, in the same conditions regarding datasets and preprocessing. The reference methods are nicMSlesion by [37], DeepMedic by [22], and 2.5D Tiramisu by [41]. These methods have been selected for the availability of the authors source code and the relevance of their contributions in the MS segmentation community.

nicMSlesion: This method is based on a cascade of two 3D patch-wise CNNs. The first one is trained to be sensitive to reveal lesion candidates. The second one is trained to reduce the misclassified voxels from the first network. Training is performed on $11 \times 11 \times 11$ patches randomly augmented with flipping and rotations. Therefore, nicMSlesion involves classical data augmentation. In the first network, the negative class is under-sampled to the same number of existing lesion voxels. It is composed of patches extracted from all of the available lesion voxels and a random selection of normal-appearing tissue voxels. Afterwards, an evaluation of the first CNN model is computed by performing inferences on the same train-set and identifying negative voxels that have been misclassified as lesions (False Positives). Finally, the second model is trained using a balanced set composed of all the lesion voxels and a random selection from the identified False Positives in the previous step.

DeepMedic: This method is based on an 11-layers deep dual pathway 3D CNN designed for brain lesion segmentation. to incorporate both local and larger contextual information, the dual pathway architecture processes the input images at multiple scales simultaneously. To overcome the computational burden, the authors use a hybrid dense training scheme processing adjacent image patches into one pass through the network. To refine the network segmentation and remove false positives, a 3D fully connected conditional random field is used. The training includes data augmentation with sagittal reflections.

2.5D Tiramisu: This method is based on a fully convolutional densely connected network. The model uses stacked slices from all three anatomical planes to achieve a 2.5D based method. Individual slices from a given orientation provide global context along the plane and the stack of adjacent slices adds local context. The training also includes flipping and rotations of the 2D patches for data augmentation. Therefore, 2.5D Tiramisu involves classical data augmentation. For each stack of 2D 128×128 slices composed of a center slice and its 2 adjacent slices, the model produces the segmentation of the center slice. Then, the inference results along the different orientations are combined via a majority vote to output the final segmentation.
For both these methods, we use the implementations provided publicly by the authors (see [3] and [4]).

[2] https://smart-stats-tools.org/lesion-challenge

[3] https://github.com/sergivalverde/nicMSlesions

[4] https://github.com/MedICL-VU/LesionSeg

2.5.3 Statistical Test

To assert the advantage of a technique obtaining the highest average score, we conducted a two-sided Wilcoxon test (*i.e.*, paired statistical test) over the lists of hybrid scores measured at image level (for the consistency of the segmentation section we took the lists of dice indices between the two segmentations). The significance of the test is established for a p-value below 0.05. In the following tables, * indicates a significantly better average score when compared with the rest of the other approaches.

3 Results

3.1 Ablation Study

To demonstrate the impact of each proposed contribution on domain generalization, we measured separately their effects on different metrics. To show both the effect on accuracy improvement and the domain shift robustness, we propose an out-of-domain and in-domain ablation study. First, we trained each method configuration on ISBI challenge train-set, then we validated on both ISBI test-set (see Table 2) and MSSEG'16 (see Table 3). To ensure a fair comparison, each configuration is trained until convergence. Specifically, we used an early stopping criterion of 50 epochs (*i.e.*, the training stops if the loss function does not improve on the validation set during 50 epochs) with a maximum number of 500 epochs. We verified that none of the configurations reached this maximum number.

Table 2: Ablation study results with different variants of our approach trained on ISBI challenge train-set and tested on ISBI test-set. DeepLesionBrain (DLB) refers to using our spatially distributed specialized networks, each network in charge of segmenting a sub-volume. The generic network represents the variant of DLB with a single network (without the spatially distributed strategy). Hierarchical Specialized Learning (HSL) indicates that we initialized the "specialized networks" with the "generic Network". To evaluate the performance of the proposed Data Augmentation, we compared variants with IQDA (previously defined in 2.1.2) and without IQDA. For each metric, the bold values indicate the best result. In the hybrid score column, * indicates a significantly better score than the other approaches using the two-sided Wilcoxon test.

Method	Hybrid Score	Dice	PPV	TPR	LFPR	LTPR	Submission Score
DLB with HSL and IQDA	**0.747***	0.646	0.888	0.545	0.131	0.486	**92.849**
DLB with HSL and without IQDA	0.732	**0.677**	0.849	0.603	0.192	**0.489**	92.383
DLB without HSL and with IQDA	0.710	0.576	**0.892**	0.453	**0.121**	0.360	91.713
DLB with models genesis init. and IQDA	0.718	0.621	0.867	0.513	0.187	0.438	91.885
DLB with AssemblyNet init. and IQDA	0.723	0.628	0.885	0.515	0.140	0.406	92.109
The generic network with IQDA	0.736	0.668	0.859	0.585	0.178	**0.489**	92.491
The generic network without IQDA	0.688	0.654	0.502	**0.869**	0.162	0.468	92.425

Table 2 shows the effect of each contribution to segmentation accuracy, when trained on ISBI challenge train-set and tested on ISBI test-set. First, the best performing combination is DLB with HSL and IQDA, it obtained an ISBI Score of 92.849. Second, both the versions of DLB without IQDA and DLB without HSL are less accurate. They obtained respectively ISBI scores of 92.383 and 91.713. The later comparison shows the impact of HSL on the accuracy of segmentations. Moreover, the generic network is less accurate than our spatially distributed approach used in DLB. The variant of generic Network with IQDA obtained a score of 92.491, whereas the variant without IQDA obtained a hybrid score of 92.425. Finally, we compare HSL with other weight initialization strategies. Specifically, HSL is compared with the neighbor transfer learning from AssemblyNet proposed by [12] and models genesis proposed by [44]. Although both variants obtained a better score compared to DLB without HSL, both initialization strategies gave a lower score than DLB with HSL and IQDA.

Table 3: Ablation study results with different variants of our approach trained on ISBI challenge train-set and tested on MSSEG'16 (see caption of Table 2 for details).

Method	Hybrid Score	Dice	PPV	TPR	LFPR	LTPR
DLB with HSL and IQDA	**0.684***	0.639	0.768	0.608	**0.319**	0.700
DLB with HSL and without IQDA	0.673	**0.669**	0.728	0.671	0.416	0.725
DLB without HSL and with IQDA	0.648	0.562	**0.806**	0.489	0.320	0.629
DLB with models genesis init. and IQDA	0.623	0.593	0.737	0.576	0.436	0.665
DLB with AssemblyNet init. and IQDA	0.610	0.541	0.708	0.537	0.466	0.705
The generic Network with IQDA	0.672	0.665	0.721	**0.673**	0.413	**0.727**
The generic network without IQDA	0.626	0.625	0.763	0.588	0.449	0.611

Table 3 shows the effect of each contribution to domain shift robustness, when trained on ISBI challenge train-set and tested on MSSEG'16. First, the most robust combination is DLB with HSL and IQDA, it obtained a hybrid score of 0.684. Second, both the variants of DLB without IQDA and DLB without HSL are less accurate. They obtained hybrid scores of 0.673 and 0.648 respectively. Moreover, the generic network is less robust than our spatially distributed approach with DLB. The variant of the generic network with IQDA obtained a score of 0.672, whereas the variant without IQDA obtained a hybrid score of only 0.626. The later comparison shows the impact of IQDA on robustness even without the spatially distributed networks. Finally, HSL is compared with Assembynet [12] and model genesis [44] initialization strategies. The variants with model genesis and AssemblyNet initialization methods obtained respectively hybrid scores of 0.623 and 0.6103.

3.2 Cross-dataset Testing

In this section, we assess the cross-dataset robustness and generalization ability of our proposed approach. We chose to compare our method with three state-of-the-art approaches: nicMSlesion [37], DeepMedic [22], and 2.5D Tiramisu [41].
During the proposed validation, all the methods have been trained on exactly the same dataset (*i.e.*, same preprocessing, same number of modalities, *etc.*) to ensure a fair comparison of method performance. Although reference methods have been originally proposed with a specific number of modalities (*i.e.*, Tiramisu 2.5D and nicMSlesion were tested with 4 and 3 modalities respectively), their implementation is independent of the number of modalities since all modalities are concatenated and fed to the CNN. Besides, their official open-source implementations support the usage of only T1w and FLAIR sequences. Thus, in this evaluation, all methods are trained using only these two modalities. The following cross-dataset testing (cross-domain testing) consists in training each technique on one dataset at each time. Afterward, the obtained models are evaluated on the other datasets which contain unseen domains. We verified the average inference time per image for each method on the same machine and the same preprocessed images: 57.353s for nicMSlesion, 17.547s for DeepMedic, 47.471s for 2.5D Tiramisu, and 38.014s for DLB (this time does not include image preprocessing). Unlike using a single network to segment patches coming from the entire image, DLB uses multiple networks, one network for each sub-volume. These networks are loaded one by one to enable the use of a common GPU hardware solution (*e.g.*, NVIDIA Titan Xp with 12 GB in our setup). Even though DLB requires sequential loading of multiple networks on GPU, the inference time over the whole image is similar to using a single network since network weights loading time is negligible compared to patch segmentation time. The ISBI score is returned by the challenge website only for ISBI test-set evaluation, and thus this metric is not available (NA) for testing on other datasets.

3.2.1 Trained on ISBI

Table 4: Results of the different approaches trained on the ISBI training dataset, with T1w and FLAIR modalities. For each metric, the bold values indicate the best result. In the hybrid score column, * indicates a significantly better score than the three other approaches using the two-sided Wilcoxon test. Red values indicate hybrid scores lower than 0.5 or Dice index below 0.25.

Trained on ISBI	Approach	Hybrid Score	Dice	PPV	TPR	LFPR	LTPR	CORR	Submission Score
MSSEG'16	nicMSlesion	0.537	0.442	0.614	0.423	0.504	0.629	0.495	NA
	DeepMedic	0.510	0.476	0.542	0.560	0.829	**0.850**	0.509	NA
	2.5D Tiramisu	**0.711**	**0.664**	0.741	**0.658**	**0.284**	0.695	**0.730**	NA
	DLB	0.684	0.639	**0.768**	0.608	0.319	0.700	0.650	NA
In-house dataset	nicMSlesion	0.419	0.204	0.727	0.129	0.309	0.361	0.158	NA
	DeepMedic	0.523	0.536	0.633	0.499	0.805	**0.765**	0.549	NA
	2.5D Tiramisu	0.654	0.545	**0.871**	0.410	**0.204**	0.476	0.635	NA
	DLB	**0.696***	**0.675**	0.850	**0.564**	0.342	0.644	**0.718**	NA

Table 4 shows the results of segmentation when training the different approaches using T1w and FLAIR modalities, on the ISBI training dataset (2D resolution FLAIR).

When validating the methods on MSSEG'16, we report that 2.5D Tiramisu obtained slightly better results (not significantly) than DLB, in terms of hybrid score whereas nicMSlesion and DeepMedic performed relatively worse with 0.537 and 0.51 respectively.
On our in-house dataset, DLB performed significantly better with a hybrid score of 0.696 while 2.5D Tiramisu, DeepMedic, and nicMSlesion obtained respectively 0.654, 0.523, and 0.419. We can notice that nicMSlesion offers poor cross-domain performance on 3D FLAIR when trained with a 2D FLAIR dataset.

3.2.2 Trained on MSSEG'16

Table 5: Results of the different approaches trained on the MSSEG'16 dataset, with T1w and FLAIR modalities. For each metric, the bold values indicate the best result. In hybrid score column, * indicates a significantly better score than the three other approaches using the two-sided Wilcoxon test. Red values indicate hybrid scores lower than 0.5 or Dice index below 0.25.

Trained on MSSEG'16	Approach	Hybrid Score	Dice	PPV	TPR	LFPR	LTPR	CORR	Submission Score
ISBI test-set	nicMSlesion	0.555	0.398	0.717	0.292	0.368	0.206	0.822	87,173
	DeepMedic	0.547	0.378	0.801	0.265	0.416	0.298	0.717	87.344
	2.5D Tiramisu	0.462	0.165	**0.937**	0.096	**0.075**	0.160	0.212	86,686
	DLB	**0.618***	**0.535**	0.697	**0.471**	0.353	**0.373**	**0.835**	**89.043**
In-house dataset	nicMSlesion	0.669	0.686	0.689	0.705	0.467	0.717	0.737	NA
	DeepMedic	0.597	0.645	0.647	0.670	0.721	**0.811**	0.650	NA
	2.5D Tiramisu	0.664	0.706	**0.766**	0.694	**0.432**	0.801	0.552	NA
	DLB	**0.697***	**0.746**	0.681	**0.847**	0.478	0.754	**0.799**	NA

Table 5 shows the results of segmentation when training the different approaches on the MSSEG'16 dataset comprising 3D T1w and 3D FLAIR modalities. First, we notice that our approach obtained significantly better hybrid scores for both the ISBI test and the In-house datasets. Second, when validating on ISBI, the obtained submission score is 89.043 for DLB (the closest to human performance), 87.344 for DeepMedic, 87.173 for nicMSlesion, and 86.686 for 2.5D Tiramisu (the farthest from human performance). In the same conditions, 2.5D Tiramisu obtained the average Dice of 0.165 which indicates a failure of the method and thus a lack of generalization in this scenario (when trained on high-quality 3D FLAIR and tested on low-quality 2D FLAIR). Finally, for the In-house dataset, DLB produced significantly better segmentation than other methods. DLB obtained a hybrid score of 0.697 while nicMSlesion obtained 0.669, 2.5D Tiramisu obtained 0.664, and DeepMedic obtained the lowest score of 0.597.

3.2.3 Trained on In-house

Table 6: Results of the different approaches trained on In-house dataset, with T1w and FLAIR modalities. For each metric, the bold values indicate the best result. In the hybrid score column, * indicates a significantly better score than the three other approaches, using the two-sided Wilcoxon test. Red values indicate hybrid scores lower than 0.5 or Dice index below 0.25.

Trained on In-house	Approach	Hybrid Score	Dice	PPV	TPR	LFPR	LTPR	CORR	Submission Score
MSSEG'16	nicMSlesion	0.700	0.650	**0.822**	0.586	**0.150**	0.607	0.607	NA
	DeepMedic	0.717	0.694	0.750	0.701	0.345	**0.782**	0.709	NA
	2.5D Tiramisu	**0.745**	0.665	0.741	0.687	0.164	0.720	0.722	NA
	DLB	0.741	**0.719**	0.735	**0.744**	0.209	0.671	**0.776**	NA
ISBI test-set	nicMSlesion	0.453	0.131	0.644	0.075	0.338	0.050	0.712	84,512
	DeepMedic	0.523	0.385	0.807	0.273	0.388	**0.215**	0.670	86.810
	2.5D Tiramisu	0.608	0.355	**0.938**	0.231	**0.065**	0.160	0.689	89.289
	DLB	**0.638***	**0.476**	0.877	**0.348**	0.104	0.193	**0.787**	**89.843**

Table 6 shows the results of segmentation when training on our In-house dataset with 3D FLAIR. First, the obtained results when testing on MSSEG'16 indicates a close segmentation accuracy for DLB and 2.5D Tiramisu in terms of hybrid score (0.745 and 0.741) and slightly lower performance from nicMSlesion and DeepMedic (0.7 and 0.717). Second, we notice that our approach obtained a significantly higher hybrid score when validating on the ISBI testing dataset, with a submission score of 89.843 compared to 2.5 Tiramisu, DeepMedic, and nicMSlesion with 89.289, 86.810, and 84.512 respectively. In this scenario, nicMSlesion obtained the worst score with a Dice of 0.131 indicating a failure of the method.

3.2.4 Cross-dataset Testing Summary

First, it is noteworthy that when our approach obtained a better score, the superiority was statistically significant. On the contrary, when one of the other approaches obtained a higher score, the advantage was not significant using the Wilcoxon test.

Second, it should be pointed out that in all the considered cross-domain cases, DLB did not degenerate not even once while maintaining high scores. We reported for nicMSlesion trained on ISBI and validated on the In-house dataset a hybrid score of 0.419. We also recall the low performance of 2.5D Tiramisu trained on MSSEG'16 and tested on ISBI (0.462 hybrid score). This shows the cross-domain robustness of the proposed strategy.

Table 7 sums up cross-dataset experiments results. This table presents the average score estimated over all the images obtained during the three experiments presented in Table 4, Table 5, and Table 6 (61 images for ISBI test-set, 43 images for In-house, 15 images for MSSEG'16). We notice that DLB obtains the highest hybrid score and Dice index by a large margin compared to 2.5D tiramisu, DeepMedic, and nicMSlesion.

Table 7: Summary of the cross-dataset experiment. The table represent the average of cross-dataset experiment results (see Table 4, Table 5, and Table 6) based on the number of images for each dataset. For each metric, the bold values indicate the best result. In the hybrid score column, * indicates a significantly better score than the three other approaches using the Wilcoxon test.

Strategy	Hybrid Score	Dice	PPV	TPR	LFPR	LTPR	CORR
nicMSlesion	0.526	0.365	0.695	0.308	0.362	0.338	0.595
DeepMedic	0.554	0.483	0.725	0.429	0.556	**0.520**	0.649
2.5 D Tiramisu	0.608	0.443	**0.870**	0.368	**0.179**	0.402	0.537
DLB	**0.663***	**0.601**	0.775	**0.550**	0.299	0.484	**0.780**

3.3 Same Domain Validation

Despite the previously mentioned limitations of in-domain validation, we also provide experiments using the same domain as complementary results. First, Table 8 shows the results of DLB, nicMSlesion, DeepMedic, and 2.5D Tiramisu on ISBI test-set after being trained on ISBI train-set (same domain), with T1w and FLAIR modalities. The three approaches give close results with submission scores of 92.923 for 2.5D Tiramisu, 92.849 for DLB, and 92.161 for nicMSlesion. DeepMedic comes last with a submission score of 90.866.

Second, Table 9 shows the current top-performing methods on the ISBI challenge website. 2.5D Tiramisu [41] is the best-ranked method with the current highest ISBI Score of 93.21, followed in second place by nnUnet [21] with 93.09. Both approaches rely on 4 modalities (T1w, FLAIR, T2, PD). Our approach comes in third place using only 2 modalities, and obtained the ISBI submission score of 92.85. Although DLB uses a lower number of modalities, it obtained better results than IMAGINE [18], Self-adaptive network [15], and Multi-branch [2] that obtained respectively the scores of 92.49, 92.41, and 92.12.

Table 8: Results of the different approaches trained on the ISBI training dataset and tested on ISBI test-set, with T1w and FLAIR modalities. For each metric, the bold values indicate the best result. two-stepin the hybrid score column, * indicates a significantly better score than the three other approaches using the two-sided Wilcoxon test.

Approach	Hybrid Score	Dice	PPV	TPR	LFPR	LTPR	CORR	Submission Score
nicMSlesion	0.724	0.639	0.853	0.541	0.144	0.432	0.863	92.161
DeepMedic	0.649	0.643	0.827	0.557	0.408	**0.530**	**0.873**	90.866
2.5D Tiramisu	**0.750**	**0.672**	0.865	**0.592**	0.150	0.513	0.868	**92.923**
DLB	0.748	0.646	**0.888**	0.545	**0.131**	0.486	0.868	92.849

Table 9: State-of-the-art published results for the ISBI challenge.

Approach	Modalities	CNN type	Submission Score
2.5D Tiramisu [41]	T1w, FLAIR, T2, PD	2D	**93.21**
nnUnet [21]	T1w, FLAIR, T2, PD	2D and 3D	93.09
DLB [ours]	T1w, FLAIR	3D	92.85
IMAGINE [18]	T1w, FLAIR, T2, PD	3D	92.49
Self-adaptive network [15]	T1w, FLAIR, T2, PD	3D	92.41
Multi-branch [2]	T1w, FLAIR, T2	2D	*92.12*

The high accuracy of the results was expected as both the training and testing sets share the same domain (same acquisition conditions, and same scanner...). By tuning and adapting a method to this specific domain conditions, we can obtain artificially higher performance (*e.g.*, DLB with 4 modalities obtained a score of 92.92, and a 2D version of

DLB obtained 93.14 [5]). However, in our opinion, results reported in the same domain experiment do not truly reflect methods performances. For instance, the best performing method of this section (2.5D Tiramisu) failed when trained on different datasets (obtained submission scores of 89.043 and 89.289 in Table 5 and Table 6). The limitation of such a validation strategy is one of the main messages of our paper. Hence, we consider that cross-dataset evaluation with diverse images from different domains is a better alternative for method assessment.

3.4 Cross-dataset Segmentation Consistency

Finally, a usually under-investigated method property is its cross-dataset segmentation consistency. To assess the consistency of our model segmentation, we decided to compare the segmentation produced by each approach on the same data when the model is trained on different datasets. We compute the Dice between the different segmentations of a method as a similarity index to quantify the prediction consistency. Table 10 shows the segmentation consistency for each approach in our cross-dataset setting.

First, we analyzed the segmentations on In-house when the models are trained respectively on ISBI train-set and MSSEG'16. In this case, DLB obtained the best score of 0.647, followed by 2.5D Tiramisu and DeepMedic with 0.6261 and 0.602 respectively. Lastly, nicMSlesion obtained a score of 0.217. Second, we analyzed the segmentations on MSSEG'16 when the models are trained respectively on ISBI train-set and In-house. In this case, we obtained close consistency scores for 2.5D Tiramisu and DLB with Dice scores around 0.72 while DeepMedic and nicMSlesion are less consistent with 0.537 and 0.514 respectively. Finally, we analyzed the segmentations on ISBI test-set when comparing the models trained on ISBI train-set, the models trained on In-house, and the models trained on MSSEG'16. For all settings, DLB was significantly more consistent than both other methods with a Dice ranging from 0.63 to 0.649. 2.5D Tiramisu segmentation consistency index varies from 0.217 to 0.485. DeepMedic consistency index fluctuates from 0.49 to 0.602. nicMSlesion is the least consistent with scores ranging from 0.177 to 0.512.

During our cross-dataset consistency experiment, DLB was the only method capable of ensuring segmentation consistency independent of the training dataset. Both other methods failed several times as indicated with red color in Table 10.

Table 10: The consistency of the segmentations for each approach in the cross-dataset setting. The consistency index represents the test-set average of Dice values, each Dice is computed between two segmentations produced by the same method when trained on two different train-sets. Higher values indicate better consistency in the segmentations. The bold values indicate the best result and red values indicate consistency lower than 0.5. * indicates a significantly better segmentation consistency score than the three other approaches, using the two-sided Wilcoxon test.

Test-set		In-house	MSSEG'16	ISBI Test-set		
Train-sets	Dataset 1 vs. Dataset 2	ISBI Train-set MSSEG'16	ISBI Train-set In-house	ISBI Train-set MSSEG'16	In-house MSSEG'16	ISBI Train-set In-house
The consistency of the model predictions when trained on Dataset1 vs. Dataset2	nicMSlesion	0.217	0.514	0.512	0.250	0.177
	DeepMedic	0.602	0.537	0.490	0.602	0.496
	2.5D Tiramisu	0.615	**0.726**	0.217	0.485	0.460
	DLB	**0.647**	0.719	**0.630***	**0.637***	**0.649***

Figure 4 represents an image from the In-house dataset and the segmentation of the different methods when trained on the ISBI challenge and MSSEG'16 datasets. First, both nicMSlesion and 2.5D Tiramisu fail to segment the majority of lesions when trained on ISBI challenge dataset. This exhibits the limitation of the robustness of these methods to domain shift, especially for 2.5D Tiramisu currently considered as the state-of-the-art approach on the ISBI challenge. Second, DLB detects almost all the lesions in the same conditions. Third, although DeepMedic also detects most of the lesions, it is more prone to false positives compared to the other methods. Finally, when choosing MSSEG'16 as a training dataset, DLB produces the most similar segmentation to expert annotation.

Figure 5 represents an image from MSSEG'16 dataset and the segmentation of the different methods when trained on ISBI challenge and In-house datasets. First, when trained on ISBI dataset, the segmentations of 2.5D Tiramisu, DeepMedic, and DLB are more accurate than nicMSlesion segmentation, although all techniques missed a large portion of the central lesion (False Negative) located around the midsagittal plane. These common voxels misclassification can result from the subjectivity of raters between training and testing datasets. Second, when trained on In-house, DLB delineates successfully most of the lesions. Especially in the case of small lesions, DLB missed only one lesion, whereas both nicMSlesion and 2.5 D Tiramissu missed four lesions, and DeepMedic misses two lesions.

[5] https://smart-stats-tools.org/lesion-challenge

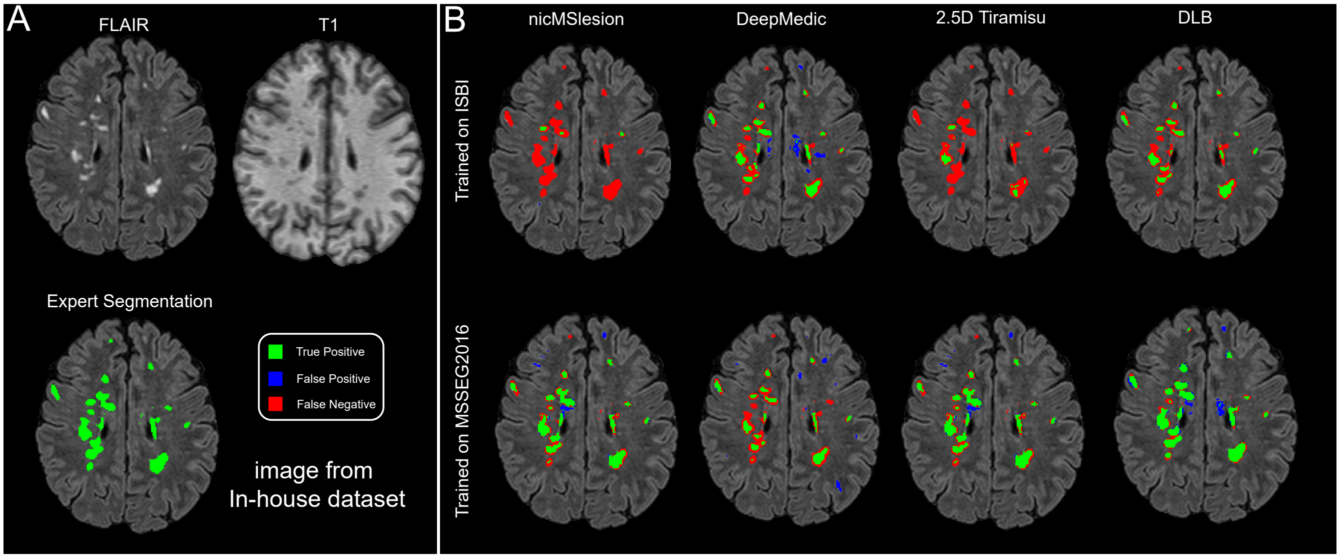

Figure 4: Part A (left) axial sections of multi-modal MRI (T1w and FLAIR) from In-house dataset, and its respective expert consensus segmentations for MS lesion segmentation. Part B (right) cross dataset segmentation of the image section shown in Part A. The first and second rows illustrate the segmentations of methods when trained respectively on ISBI dataset, and MSSEG'16 datasets. The first, second, third, and fourth columns represent respectively the segmentations of nicMSlesion, DeepMedic, 2.5D Tiramisu, and DeepLesionBrain.

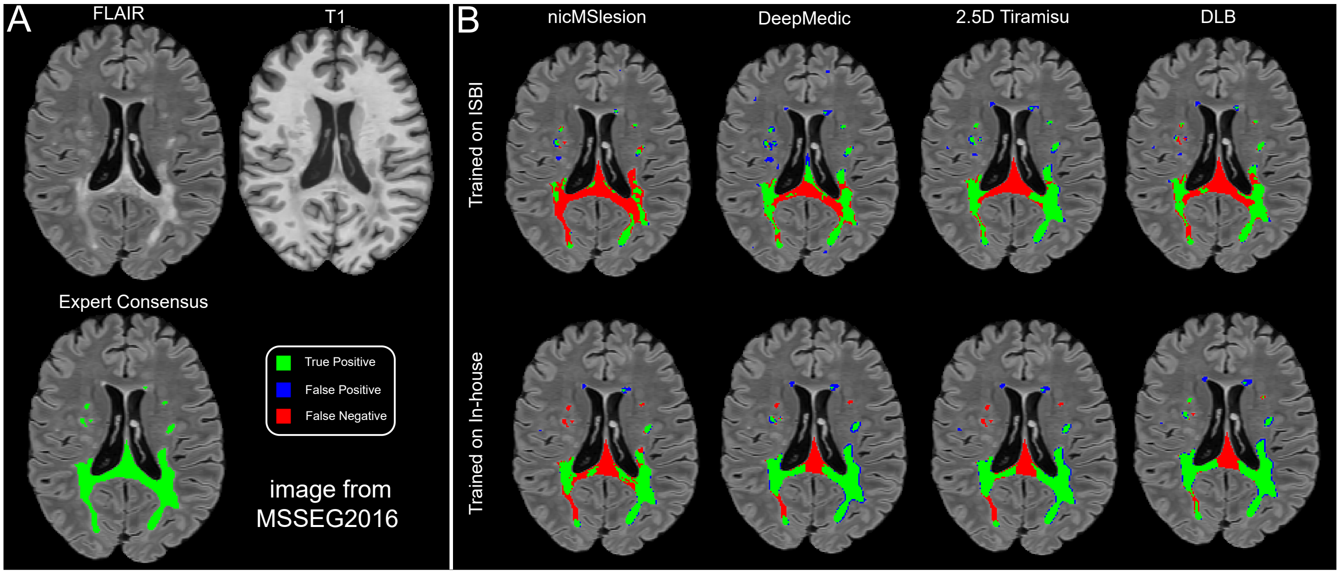

Figure 5: Part A (left) axial sections of multi-modal MRI (T1w and FLAIR) from MSSEG'16 dataset, and its respective expert consensus segmentations for MS lesion segmentation. Part B (right) cross dataset segmentation of the image section shown in Part A. The first and second rows illustrate the segmentations of methods when trained respectively on ISBI challenge, and In-house datasets. First, second, third, and fourth columns represent respectively the segmentations of nicMSlesion, DeepMedic, 2.5D Tiramisu, and DeepLesionBrain.

Figure 6 represents an image from the ISBI challenge and the segmentation of the different methods when trained on MSSEG'16 and In-house datasets. From the four methods, DLB had the most consistent segmentation across different conditions of training domains. In this case, nicMSlesion produced a decent segmentation for this example only when

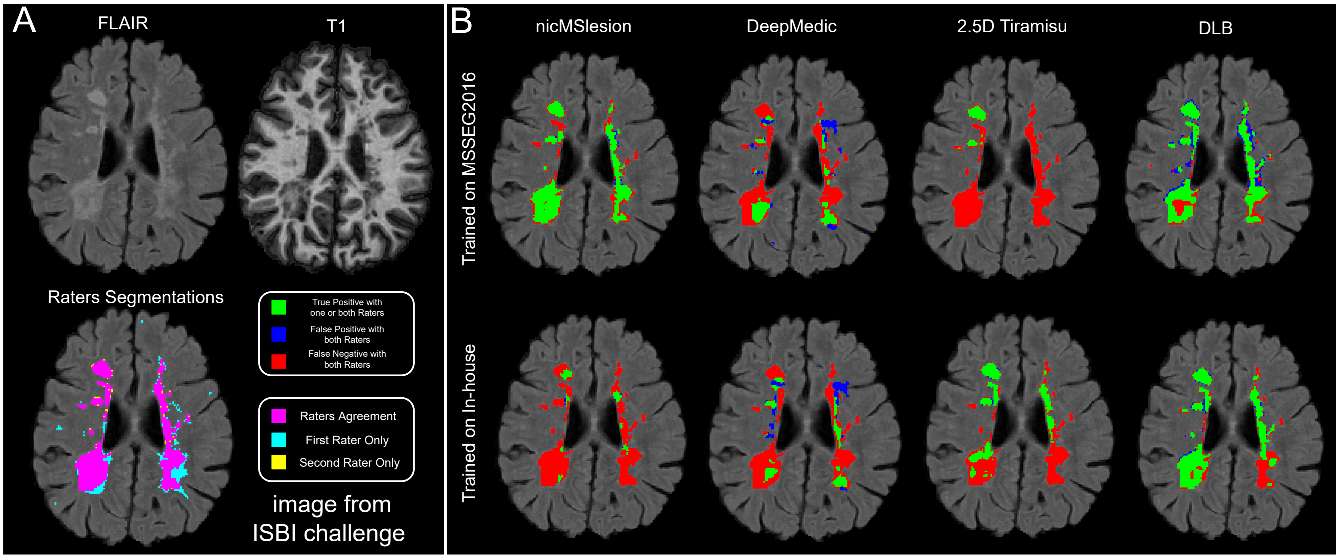

Figure 6: Part A (left) axial sections of multi-modal MRI (T1w and FLAIR) from ISBI challenge dataset, and its respective raters segmentations for MS lesion segmentation. Part B (right) cross dataset segmentation of the image section shown in Part A. The first and second rows illustrate the segmentations of methods when trained respectively on MSSEG'16, and In-house datasets. First, second, third, and fourth columns represent respectively the segmentations of nicMSlesion, DeepMedic, 2.5D Tiramisu, and DeepLesionBrain.

trained on MSSEG'16. Likewise, 2.5D Tiramisu produced better segmentation when trained on In-house than on MSSEG'16. Although DeepMedic is consistent for this case, the produced segmentation was less precise and prone to false positives compared to the other methods.

4 Discussion and Conclusion

4.1 Discussion

Deep learning-based segmentation models can be prone to generalization failure due to domain shift between training and testing data. Such a domain shift may be caused by hardware and preprocessing diversity, the difference in acquisition protocol or annotation protocol, that results in a difference between the distributions of training and testing datasets. Besides, we also have to acknowledge the subjectivity of raters in training datasets. Indeed, the disagreement between expert segmentations, both in the same dataset and across different datasets, can make it difficult to train a generalizing model. Our experiments showed the limited generalization capability of state-of-the-art approaches, whereas DLB was able to adapt across different domains. Our study emphasizes the importance of cross-dataset validation, particularly when considering the clinical application of machine learning.

DLB uses a group of several separately trained networks, each network is specialized in a particular sub-volume of the brain. In the ablation study (see Tables 2 and 3), our spatially distributed networks strategy showed better generalization and higher accuracy than using a single model.

In our work, we considered both specialized local features, and generic global features of MS lesions. The hierarchical specialization learning proposes an alternative to network cascades (*i.e.*, [36]). Instead of using cascades that are prone to error propagation, we suggested a logical hierarchy during learning based on data selection and transfer learning. The ablation study (see Tables 2 and 3) exhibits the contribution of HSL to accuracy and domain generalization compared to DLB without HSL.

In this paper, the proposed method was validated using an out-of-domain cross-dataset evaluation. This strategy ensures that the performance obtained is not biased by the training dataset domain information. Indeed, the use of testing and training images from the same domain is questionable and does not reflect the generalization ability. The community should start considering this issue for both the validation and the comparison of proposed methods.

Automated MS lesion segmentation should be able to render the most accurate segmentation with the minimum number of modalities, to be efficiently adopted in clinical conditions and to limit inter-modality dependence. Many experts agree that FLAIR is the most important modality for MS lesion delineation. Moreover, T1w modality can provide complementary information for better white-matter, gray-matter, and cerebrospinal-fluid distinction. FLAIR and T1w are the most available modalities for MS patients and in all MS benchmark datasets. Our method achieved a competitive performance using these two modalities even on unseen domains.

In this paper, we proposed a novel data augmentation technique to reduce domain shift introduced by the variability of image resolution and quality. IQDA simulates different acquisition conditions to reduce covariate shift. Our ablation study (see Tables 2 and 3) showed IQDA as a solid contribution to segmentation accuracy and cross-domain generalization. Indeed, while other methods (nicMSlesion, DeepMedic, and 2.5D Tiramisu) involve usual data augmentation (rotation and flipping), such simple strategies failed to ensure good generalization on unseen datasets.

Both domain generalization and adaptation are concerned with reducing dataset bias. The difference between these strategies is that for domain adaptation, some unlabeled data or even a few labeled data from the target domain are exploited to capture properties of the target domain for model adaptation. However, in domain generalization, no samples of any kind are used from the target domain. Domain generalization has been proposed to address the problem of unavailability of target domain samples by leveraging the labeled data to learn a universal representation to generalize for any target domain and without any prior insight from that domain. In this work, we emphasize on testing the domain generalization of our approach with cross-dataset evaluation. Unlike domain adaptation such as one-shot domain adaptation (*i.e.*, [37]), DLB does not need expert segmentation from the target domain. Our testing conditions draw a clear distinction between training data containing source domains and testing data containing unseen target domains.

In section 3.2, we reported that the best performances of DLB have been obtained when using high-resolution 3D FLAIR datasets and multi-rater consensus ground truth for training. The resulting model can render more accurate segmentations for both 2D and 3D image resolution data, even across unseen domains. This observation led us to believe that to efficiently train 3D CNN-based models for domain generalization, it may be desirable to optimize the model using high-resolution training data.

With current available hardware, it is unfeasible to exploit 3D CNNs with equivalent depth and kernel size as state-of-the-art 2D CNNs. Consequently, many neuroimaging automated pipelines are still using 2D CNNs despite processing 3D data. Our results suggest that using multiple compact networks can approximate a larger and more stable model since the sum of features extracted by the group of specialized networks and the features of a hypothetical big network may be equivalent in terms of relevant information for MS lesion segmentation. In our work, we have chosen to break down the complexity of the task spatially, based on the sub-volume division of the whole brain volume. One other advantage of this distribution is the ability to train networks in parallel since network weights and images of each region are independent. It is possible to use several GPUs for parallel training.

Our full pipeline including the preprocessing, MS lesion segmentation with DLB, and an easy-to-read report is available on our repository [6].

4.2 Conclusion

DeepLesionBrain is a deep learning framework for MS lesion segmentation designed for domain generalization. First, we use a spatially distributed strategy of multiple compact 3D CNNs with large overlapping receptive fields, to produce consensus-based segmentation robust to domain shift. Our method is trained using hierarchical specialization learning to efficiently incorporate both generic and specialized features. Second, we propose a novel image quality data augmentation to increase training data variability in a realistic way. Finally, we use only T1w and FLAIR modalities to propose a method compatible with a large number of datasets.

The ablation study showed the impact of each contribution on segmentation accuracy and domain generalization. The out-of-domain cross-dataset testing is suggested as an alternative for method evaluation in areas that are sensitive to domain bias (*i.e.*, medical imaging). Our validation showed the generalization ability of our method and its robustness to domain shift. We also proved experimentally that DLB produces consistent segmentations compared to other state-of-the-art approaches regardless of the training data domain.

[6] https://github.com/volBrain/DeeplesionBrain

5 Acknowledgements

This work benefited from the support of the project DeepvolBrain of the French National Research Agency (ANR-18-CE45-0013). This study was achieved within the context of the Laboratory of Excellence TRAIL ANR-10-LABX-57 for the BigDataBrain project. Moreover, we thank the Investments for the future Program IdEx Bordeaux (ANR-10-IDEX-03-02, HL-MRI Project), Cluster of excellence CPU and the CNRS/INSERM for the DeepMultiBrain project. This study has also been supported by the DPI2017-87743-R grant from the Spanish Ministerio de Economia, Industria Competitividad. The authors gratefully acknowledge the support of NVIDIA Corporation with their donation of the TITAN Xp GPU used in this research.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

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

[3] S. Aslani, V. Murino, M. Dayan, R. Tam, D. Sona, and G. Hamarneh. Scanner invariant multiple sclerosis lesion segmentation from MRI. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 781–785. IEEE, 2020.

[4] E. E. Bron, S. Klein, J. M. Papma, L. C. Jiskoot, V. Venkatraghavan, J. Linders, P. Aalten, P. P. De Deyn, G. J. Biessels, J. A. Claassen, et al. Cross-cohort generalizability of deep and conventional machine learning for mri-based diagnosis and prediction of alzheimer's disease. *arXiv preprint arXiv:2012.08769*, 2020.

[5] T. Brosch, L. Y. Tang, Y. Yoo, D. K. Li, A. Traboulsee, and R. Tam. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE transactions on medical imaging*, 35(5):1229–1239, 2016.

[6] 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*, 148:77–102, 2017.

[7] A. Carass, M. B. Wheeler, J. Cuzzocreo, P.-L. Bazin, S. S. Bassett, and J. L. Prince. A joint registration and segmentation approach to skull stripping. In *2007 4th IEEE international symposium on biomedical imaging: from nano to macro*, pages 656–659. IEEE, 2007.

[8] F. Chollet et al. Keras. `https://keras.io`, 2015.

[9] O. Commowick, F. Cervenansky, and R. Ameli. Msseg challenge proceedings: Multiple sclerosis lesions segmentation challenge using a data management and processing infrastructure, 2016.

[10] 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. *Scientific reports*, 8(1):1–17, 2018.

[11] O. Commowick, N. Wiest-Daesslé, and S. Prima. Block-matching strategies for rigid registration of multimodal medical images. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 700–703. IEEE, 2012.

[12] P. Coupé, B. Mansencal, M. Clément, R. Giraud, B. D. de Senneville, V.-T. Ta, V. Lepetit, and J. V. Manjon. AssemblyNet: A large ensemble of cnns for 3d whole brain mri segmentation. *NeuroImage*, page 117026, 2020.

[13] P. Coupé, T. Tourdias, P. Linck, J. E. Romero, and J. V. Manjón. Lesionbrain: an online tool for white matter lesion segmentation. In *International Workshop on Patch-based Techniques in Medical Imaging*, pages 95–103. Springer, 2018.

[14] P. Coupé, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot. An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images. *IEEE transactions on medical imaging*, 27(4):425–441, 2008.

[15] Y. Feng, H. Pan, C. Meyer, and X. Feng. A self-adaptive network for multiple sclerosis lesion segmentation from multi-contrast mri with various imaging sequences. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 472–475. IEEE, 2019.

[16] M. Filippi, P. Preziosa, B. L. Banwell, F. Barkhof, O. Ciccarelli, N. De Stefano, J. J. Geurts, F. Paul, D. S. Reich, A. T. Toosy, et al. Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines. *Brain*, 142(7):1858–1875, 2019.

[17] D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, and D. L. Collins. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Medical image analysis*, 17(1):1–18, 2013.

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

[19] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

[20] Y. Huo, Z. Xu, Y. Xiong, K. Aboud, P. Parvathaneni, S. Bao, C. Bermudez, S. M. Resnick, L. E. Cutting, and B. A. Landman. 3D whole brain segmentation using spatially localized atlas network tiles. *NeuroImage*, 194:105–119, 2019.

[21] F. Isensee, J. Petersen, S. A. Kohl, P. F. Jäger, and K. H. Maier-Hein. NNU-Net: Breaking the spell on successful medical image segmentation. *arXiv preprint arXiv:1904.08128*, 1:1–8, 2019.

[22] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[23] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[24] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] H. Li, S. Jialin Pan, S. Wang, and A. C. Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018.

[26] J. V. Manjón and P. Coupé. volbrain: an online mri brain volumetry system. *Frontiers in neuroinformatics*, 10:30, 2016.

[27] J. V. Manjon, J. E. Romero, and P. Coupe. DeepHIPS: A novel deep learning based hippocampus subfield segmentation method. *arXiv preprint arXiv:2001.11789*, 2020.

[28] G. Mårtensson, D. Ferreira, T. Granberg, L. Cavallin, K. Oppedal, A. Padovani, I. Rektorova, L. Bonanni, M. Pardini, M. G. Kramberger, et al. The reliability of a deep learning model in clinical out-of-distribution mri data: a multicohort study. *Medical Image Analysis*, page 101714, 2020.

[29] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5715–5725, 2017.

[30] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

[31] P. A. Narayana, I. Coronado, S. J. Sujit, X. Sun, J. S. Wolinsky, and R. E. Gabr. Are multi-contrast magnetic resonance images necessary for segmenting multiple sclerosis brains? a large cohort study based on deep learning. *Magnetic resonance imaging*, 65:8–14, 2020.

[32] P. Omoumi, A. Ducarouge, A. Tournier, H. Harvey, C. E. Kahn, F. Louvet-de Verchère, D. P. Dos Santos, T. Kober, and J. Richiardi. To buy or not to buy—evaluating commercial ai solutions in radiology (the eclair guidelines). *European Radiology*, pages 1–11, 2021.

[33] N. Shiee, P.-L. Bazin, J. L. Cuzzocreo, C. Ye, B. Kishore, A. Carass, P. A. Calabresi, D. S. Reich, J. L. Prince, and D. L. Pham. Reconstruction of the human cerebral cortex robust to white matter lesions: method and validation. *Human brain mapping*, 35(7):3385–3401, 2014.

[34] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.

[35] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee. N4itk: improved N3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010.

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

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

[38] L. Wu, Z. Zhu, et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

[39] Y. Wu and K. He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[40] P. Y. Yang and W. Gao. Multi-view discriminant transfer learning. 2013.

[41] H. Zhang, A. M. Valcarcel, R. Bakshi, R. Chu, F. Bagnato, R. T. Shinohara, K. Hett, and I. Oguz. Multiple sclerosis lesion segmentation with tiramisu and 2.5 D stacked slices. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 338–346. Springer, 2019.

[42] L. Zhang, Z. Deng, K. Kawaguchi, A. Ghorbani, and J. Zou. How does mixup help with robustness and generalization? *arXiv preprint arXiv:2010.04819*, 2020.

[43] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, B. J. Wood, H. Roth, A. Myronenko, D. Xu, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging*, 2020.

[44] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang. Models genesis. *Medical image analysis*, 67:101840, 2021.