Longitudinal detection of new MS lesions using deep learning

Reda Abdellah Kamraoui, Boris Mansencal, José V. Manjón, Pierrick Coupé

frontiers | Frontiers in Neuroimaging

TYPE Methods
PUBLISHED 25 August 2022
DOI 10.3389/fnimg.2022.948235

OPEN ACCESS

EDITED BY
Benoit Combès,
Inria Rennes-Bretagne Atlantique Research Centre, France

REVIEWED BY
Kalavathi Palanisamy,
The Gandhigram Rural Institute, India
Niharika S. D'Souza,
IBM Research Almaden, United States

*CORRESPONDENCE
Reda Abdellah Kamraoui
reda-abdellah.kamraoui@u-bordeaux.fr

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroimaging

RECEIVED 19 May 2022
ACCEPTED 11 July 2022
PUBLISHED 25 August 2022

CITATION
Kamraoui RA, Mansencal B, Manjon JV and Coupé P (2022) Longitudinal detection of new MS lesions using deep learning.
Front. Neuroimaging 1:948235.
doi: 10.3389/fnimg.2022.948235

Longitudinal detection of new MS lesions using deep learning

Reda Abdellah Kamraoui[1]*, Boris Mansencal[1], José V. Manjon[2] and Pierrick Coupé[1]

[1]PICTURA, Univ. Bordeaux, Bordeaux INP, CNRS, LaBRI, UMR5800, Talence, France, [2]ITACA, Universitat Politècnica de València, Valencia, Spain

The detection of new multiple sclerosis (MS) lesions is an important marker of the evolution of the disease. The applicability of learning-based methods could automate this task efficiently. However, the lack of annotated longitudinal data with new-appearing lesions is a limiting factor for the training of robust and generalizing models. In this study, we describe a deep-learning-based pipeline addressing the challenging task of detecting and segmenting new MS lesions. First, we propose to use transfer-learning from a model trained on a segmentation task using single time-points. Therefore, we exploit knowledge from an easier task and for which more annotated datasets are available. Second, we propose a data synthesis strategy to generate realistic longitudinal time-points with new lesions using single time-point scans. In this way, we pretrain our detection model on large synthetic annotated datasets. Finally, we use a data-augmentation technique designed to simulate data diversity in MRI. By doing that, we increase the size of the available small annotated longitudinal datasets. Our ablation study showed that each contribution lead to an enhancement of the segmentation accuracy. Using the proposed pipeline, we obtained the best score for the segmentation and the detection of new MS lesions in the MSSEG2 MICCAI challenge.

KEYWORDS

new lesion detection, new lesions segmentation, data augmentation, transfer learning, data synthesis

1. Introduction

Multiple sclerosis (MS) is a chronic autoimmune disease of the central nervous system. The pathology is characterized by inflammatory demyelination and axonal injury, which can lead to irreversible neurodegeneration. The disease activity, such as MS lesions, can be observed using magnetic resonance imaging (MRI). The detection of new MS lesions is one of the important biomarkers that allow clinicians to adapt the patient's treatment and assess the evolution of this disease.

Recently, the automation of single time-point MS lesion segmentation has shown encouraging results. Many techniques showed performance comparable to clinicians in controlled evaluation conditions (refer to Commowick et al., 2016; Carass et al., 2017). These methods use a single time-point scan to segment all appearing lesions at the time of the image acquisition. However, these cross-sectional techniques are not adapted to the longitudinal detection of new lesions. Indeed, using these methods requires repeatedly

running the segmentation process for each time-point independently to segment MS lesions before detecting new ones. Unlike the human reader, these methods are not designed to jointly exploit the information contained at each time point. Consequently, single-time MS lesion segmentation methods performance is not optimal for the detection of new lesions between two time-points. Moreover, inconsistencies may appear between segmentations of both time-points since they are processed independently.

To specifically address this detection task using both time-points at the same time, some detection methods have been proposed. In one of the earliest studies, Bosc et al. (2003) used a nonlinear intensity normalization method and statistical hypothesis test methods for change detection. Elliott et al. (2013) used a Bayesian tissue classifier on the time-points to estimate lesion candidates followed by a random-forest-based classification to refine the identification of new lesions. Ganiler et al. (2014) used image subtraction and automated thresholding. Cheng et al. (2018) integrated neighborhood texture in a machine learning framework. Salem et al. (2018) trained a logistic regression model with features from the image intensities, the image subtraction values, and the deformation field operators. Schmidt et al. (2019) used lesion maps of different time-points and FLAIR intensities distribution within normal-appearing white matter to estimate lesion changes. Krüger et al. (2020) used a 3D convolutional neural network (CNN) where each time-point is passed through the same encoder. Then, the produced feature maps are concatenated and fed into the decoder.

Training learning-based methods for the task of new lesions detection require a dataset specifically designed for the task. The most obvious form of the training data would be a longitudinal dataset of MS patients (with two or more successive time-points) with new appearing lesions carefully delineated by experts in the field. However, the construction of such a dataset is very difficult. To begin, new lesions may take several months or even years to appear and be visible in a patient's MR image. Moreover, a time-consuming and costly process is necessary for several experts to annotate new lesions from the two time-points and to obtain an accurate consensus segmentation. Although the organizers of the MICCAI Longitudinal Multiple Sclerosis Lesion Segmentation Challenge (MSSEG2-challenge MICCAI, 2021) provided such a dataset, the training set is severely impacted by class imbalance (refer to Section 2.5.3 for more details) due to the difficulty of finding new lesions in the follow-up scan. This under-representation of new lesions in longitudinal datasets is limiting the training of state-of-the-art deep learning algorithms from scratch on this complex task. Besides, achieving generalizing results on unseen domains (refer to Mårtensson et al., 2020; Bron et al., 2021; Omoumi et al., 2021) may require more data diversity.

Several studies tackled the problem of training data scarcity. First, transfer learning is a strategy used to create high-performance learners trained with more widely available data from different domains when the target domain/task data are expensive or difficult to collect (refer to Torrey and Shavlik, 2009; Weiss et al., 2016). Second, synthetic data generation is performed by using a model able to simulate realistic artificial data that can be used during training (refer to Tremblay et al., 2018; Tripathi et al., 2019; Khan et al., 2021). Third, data-augmentation is a set of techniques used to handle the variability in real-world data by enhancing the size and quality of the training dataset (refer to Shorten and Khoshgoftaar, 2019). Recently, Zhang et al. (2020) showed that applying extensive data augmentation during training also enhances the generalization capability of the methods.

In this article, we propose an innovative strategy integrating these three strategies into a single pipeline for new MS lesion segmentation to tackle data rarity for our task. First, we use transfer-learning to exploit the larger and more diverse datasets available for the task of single-point MS lesion segmentation which does not require longitudinal data. Second, we propose a novel data synthesis technique able to generate two realistic time-points with new MS lesions from a single FLAIR scan. Third, we use a data-augmentation technique to simulate a large variety of artifacts that may occur during the MRI acquisitions. This technique aims to enhance both the variability and size of the training data and to improve the generalization of our model.

2. Methods and materials

2.1. Method overview

To deal with data rarity for new MS lesion segmentation, we proposed a three stage pipeline as shown in Figure 1. In Stage One, an encoder-decoder network is trained on the task of single time-point MS lesions segmentation. This step aims to train the encoder part of the network to extract relevant features related to MS lesions that can be used in the next steps. Stage One enables to indirect use of large datasets dedicated to single time-point MS lesion segmentation for the task of new lesions segmentation. This stage is detailed in Section 2.2. In Stage Two, the new lesions segmentation model composed of the previous task encoder is pretrained with synthetic data. To this end, we trained external models able to generate two realistic time-points from a single image also taken from single time-point MS datasets. It combines the effects of lesion inpainting and lesion generating models to simulate the appearance of new lesions. This strategy is detailed in Section 2.3. In Stage Three, the decoder is fine-tuned with real longitudinal data from the new MS lesion training-set of the MSSEG2 MICCAI challenge.

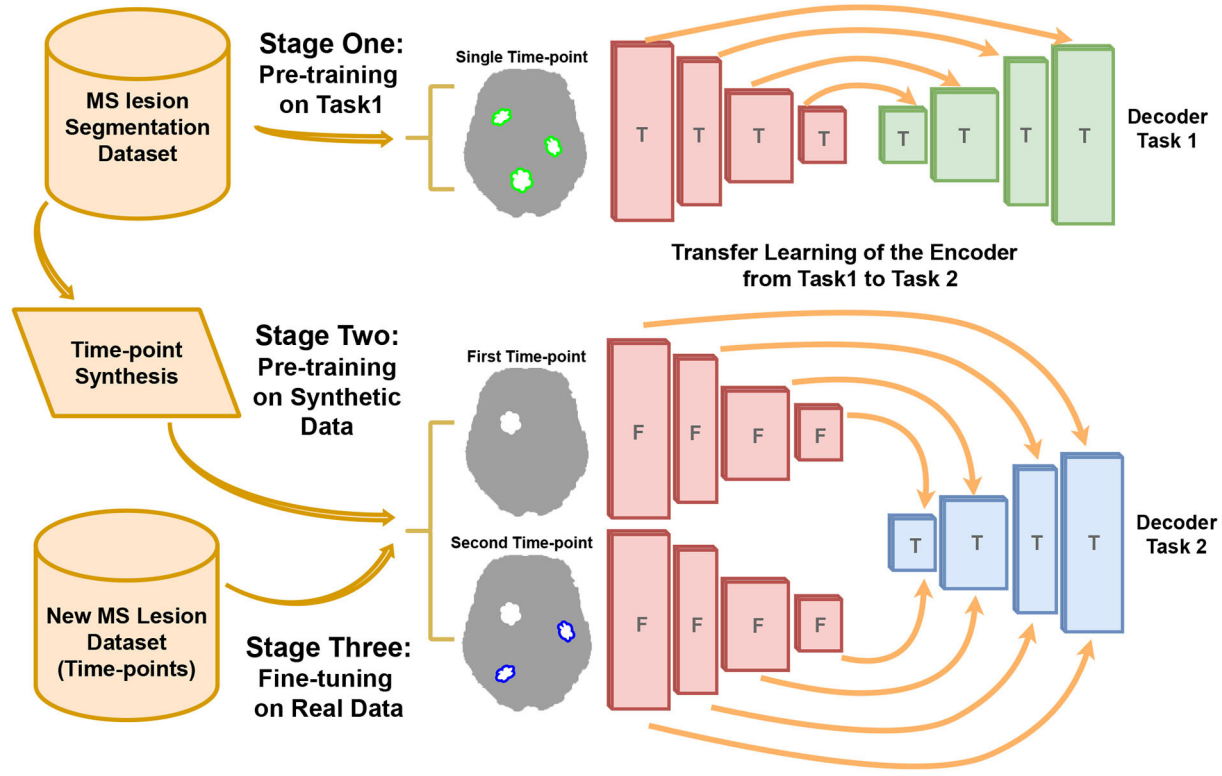

FIGURE 1
The pipeline of our new MS lesion segmentation method. The three stages include: First, the pre-training on the task of single-time-point MS lesion segmentation (Task 1). Second, pre-training on the task of new MS lesions segmentation (Task 2) with synthetic data. Third, fine-tuning the model with real data. The encoder weights are trained (T) in Stage One and freezed (F) in Stage Two and Stage Three.

2.2. Transfer-learning from single time-point MS lesion segmentation task

The encoder used for new MS lesion segmentation is first trained on single time-point lesion segmentation (refer to Figure 2, from Stage One to Stage Two). This choice is motivated by two reasons. First, we consider that datasets for MS lesion segmentation with lesion mask segmentation by experts are more diverse and larger than available datasets for new lesion segmentation (which requires a longitudinal study). Second, the task of MS lesion segmentation is tightly close to the one of new MS lesion segmentation. By learning to segment lesions, the model implicitly learns the concept of a lesion, either the lesion is considered new or was already existing in the first time-point. To conclude, since there is a proximity between the two tasks, there is likely a gain from exploiting a large amount of training data for the first task to improve the second task's performance.

2.2.1. Model architecture design

Our method is based on the transfer learning from the task of "Single time-point MS lesion segmentation" to the task of "new lesions segmentation from two time-points." Thus, two different architectures are used but with the same building blocks for each task. For the first task, a 3D U-Net shape architecture is used, as shown in Figure 3A. This kind of architecture has been very effective and robust for MS lesion segmentation (Isensee et al., 2021; Kamraoui et al., 2022). It is composed of an encoder and a decoder linked with one another by skip connections.

For the second task, a siamese-encoder followed by a single decoder is used, as shown in Figure 3B. The shared-weights encoders are chosen to extract the same set of features from both time points. Then, these features resulting from the different levels of both encoder paths are aggregated (refer to Figure 3B). The aggregation module is composed of concatenation and a convolution operation. Feature maps are first concatenated by channels (i.e., the result channel size is two times the original size), then the convolution operation aggregates the information back to the original channel size. Finally, the aggregated features are passed through the decoder.

2.3. Time-points synthesis

The data synthesis method is based on the simulation of new MS lesions between two time-points using single time-point

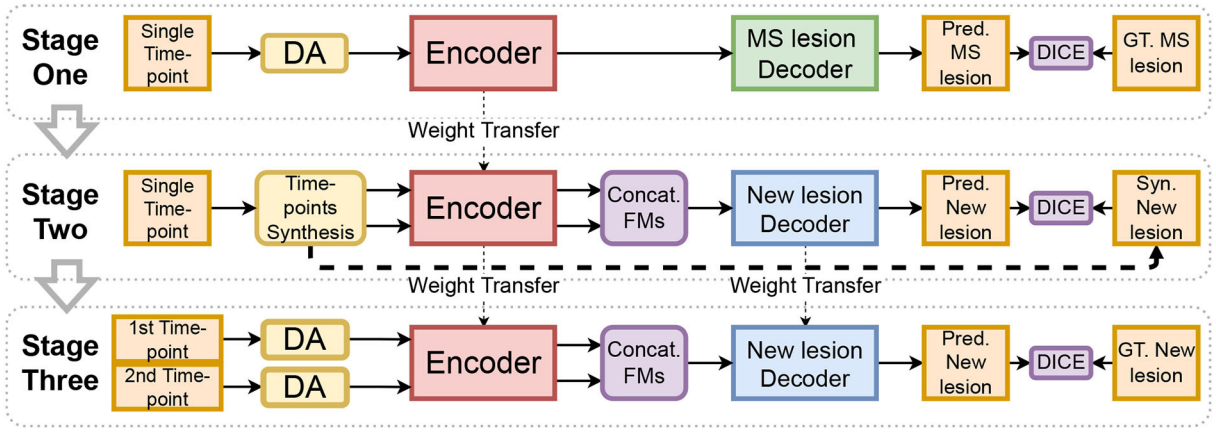

FIGURE 2
The diagram represents our training method. Input images are augmented with the proposed method (DA). The encoder trained in Stage One is used in Stage Two and Stage Three to extract feature maps (FMs) of the two-time points. The aggregation block (Concat. FMs) is used to combine features.

FLAIR images. As shown in Figure 4, our pipeline generates "on the fly" synthetic 3D patches that represent longitudinal scans of the same patient with evolution in their lesion mask. The synthetic data is generated in three steps. In the first step, a 3D FLAIR patch and its MS lesion segmentation mask are randomly sampled from different MS lesion segmentation datasets (refer to Section 2.5.1). Then, the patch and lesion mask are randomly augmented with flipping and rotations. A copy of the FLAIR patch is performed to represent the two time-points. Then, both identical patches are altered with the described data augmentation (refer to Section 2.4) to differentiate the two patches. At this point, the lesion masks of the two synthetic time-points are still identical. Thus, there are no new lesions. In the second step, a connected component operation is used to separate each independent lesion from the lesion mask. Each lesion is either inpainted (i.e., removed) from one of the two time-points or both of them, or it can be kept in both of the time-points. The lesion inpainting model is used to inpaint the lesion region with hallucinated healthy tissue (refer to Section 2.3.1). Next, the new lesion mask is constructed from lesion regions that have been kept in the second time-point but not the first one. In the third step, the lesion generator model is used to simulate new synthetic lesions at realistic locations (using white/gray matter segmentation and a probabilistic distribution of MS lesions on the brain in the MNI space). Synthetic lesions are generated for one of the time-points or both of them (refer to Section 2.3.2). Similar to the previous step, the new lesion mask is updated to include only the generated lesions on the second time-point.

2.3.1. Lesion inpainting model

The lesion inpainting model is trained, independently and priorly to our proposed pipeline, with randomly selected 3D FLAIR patches which do not contain MS lesions or white matter hyperintensities. Similar to Manjón et al. (2020), A 3D U-Net network is optimized to reconstruct altered input images. Specifically, the input patch is corrupted with Gaussian noise (i.e., with a mean and a standard deviation of the image intensities) in lesion-like areas at random locations. When the model is trained, it can be used to synthesize healthy regions in lesion locations that are replaced with random gaussian (refer to Manjón et al., 2020 for details).

2.3.2. Lesion generator model

The lesion generator is trained before our proposed pipeline to simulate realistic lesions. The generator is a 3D U-Net network with two input channels and one output channel. The first input channel receives an augmented version of 3D FLAIR patches containing MS lesions where lesions are replaced with random noise. The second input channel receives the MS lesion mask of the original 3D FLAIR patch. The output channels predict the original 3D FLAIR patch with lesions. Thus, the trained model can simulate synthetic MS lesions from a 3D patch of FLAIR and its corresponding lesion mask.

2.4. Data augmentation

The quality of the MRI greatly varies between datasets. The quality of the images depends on several factors such as signal-to-noise ratio, contrast-to-noise ratio, resolution, or slice thickness. Since our training set is limited, it does not reflect the diversity of real-world images. To make our training stages robust to the large variety of artifacts that may occur during the MRI acquisitions, an extensive Data

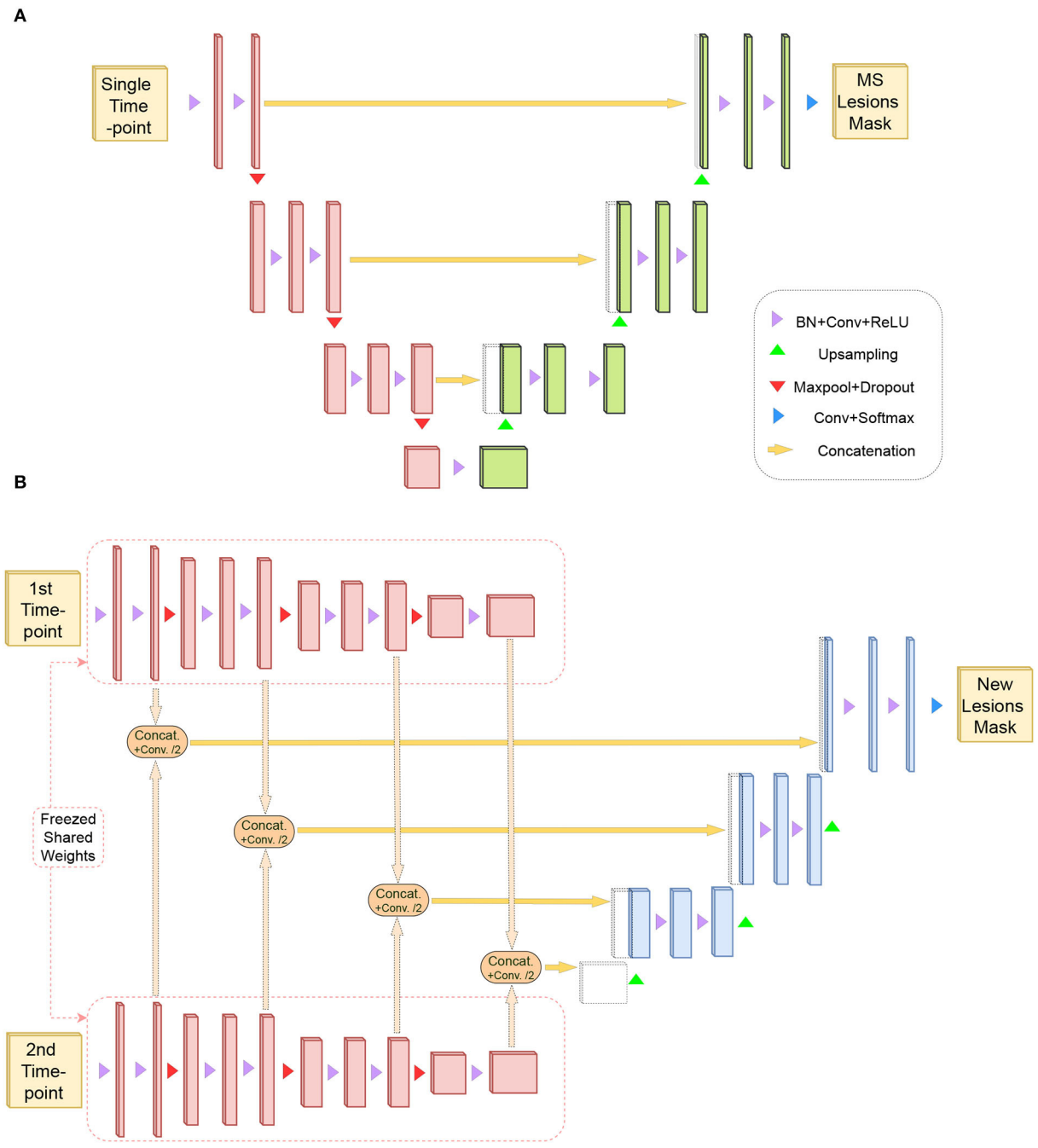

FIGURE 3
(A) Represents U-Net like architecture composed of an encoder (in red) and a decoder for the task of MS lesion segmentation (in green). This task requires a single time-point as input and produces the MS lesion mask. **(B)** Shows a siamese-encoder (in red) to extract the same sets of features from the two time-points. Same-level features are aggregated with a combination module and are forwarded to a decoder for the task of new lesions segmentation (in blue).

Augmentation (DA) is used (refer to "DA" in Figure 2 and "Data Augmentation" in Figure 4). Such DA technique also helps to better oversample the scarce samples with new lesions (refer to Section 2.5.3).

We use an improved version of the data augmentation strategy proposed in Kamraoui et al. (2022), which simulates MRI quality disparity. During training, we simulate "on the fly" altered versions of 3D patches. We randomly introduce a set of

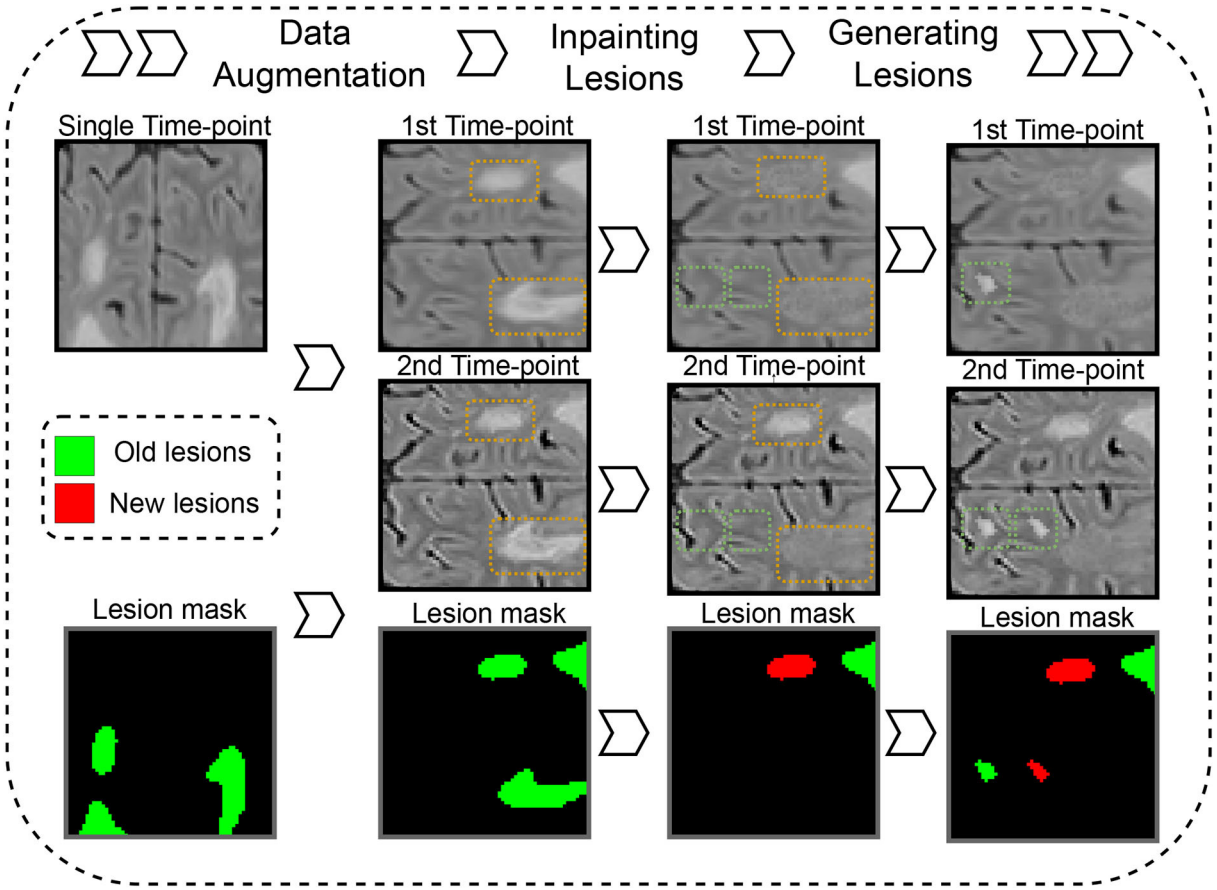

FIGURE 4
Synthetic time points with new MS lesion generation pipeline. Dashed orange and green rectangles on images represent areas where lesions are inpainted or generated.

alterations in the spatial and frequency space (k-space): Blur, edge enhancement, axial subsampling distortion, anisotropic downsampling, noise, bias-field variation, motion effect, MRI spike artifacts, and ghosting effect. Figure 5 shows augmentation samples.

For the blur, a gaussian kernel is used with a randomly selected standard deviation (SD) 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) along the axial direction with a size of $[1 \times 1 \times sz]$ where $sz \in 2, 3, 4$. For anisotropic downsampling, the image is downsampled through an axis with a random factor ranging between $[1.5, 4]$ and upsampled back again with a B-spline interpolation. For noise, we add to the image patch a Gaussian noise with 0 mean and an SD ranging between $[0.02, 0.1]$. Bias-field variation is generated using the study of Sudre et al. (2017) which considers the bias field as a linear combination of polynomial basis functions. Motion effect has been generated based on the study of Shaw et al. (2018). The movements are simulated by combining in the k-space a sequence of affine transforms with random rotation and translation in the ranges $[-5, 5]$ degrees and $[-4, 4]$ mm, respectively. Both MRI spike artifacts and the ghosting effect have been generated with the implementation of Pérez-García et al. (2021).

2.5. Data

Different datasets are used for the training and validation of the two tasks (refer to Table 1).

2.5.1. Single time-point datasets

For time-points synthesis (refer to 2.3) and encoder pretraining (refer to 2.2), we jointly used three datasets containing single time-points FLAIR and lesion masks. First, the ISBI (Carass et al., 2017) training-set contains 21 FLAIR images with expert annotation done by two raters. Although the dataset

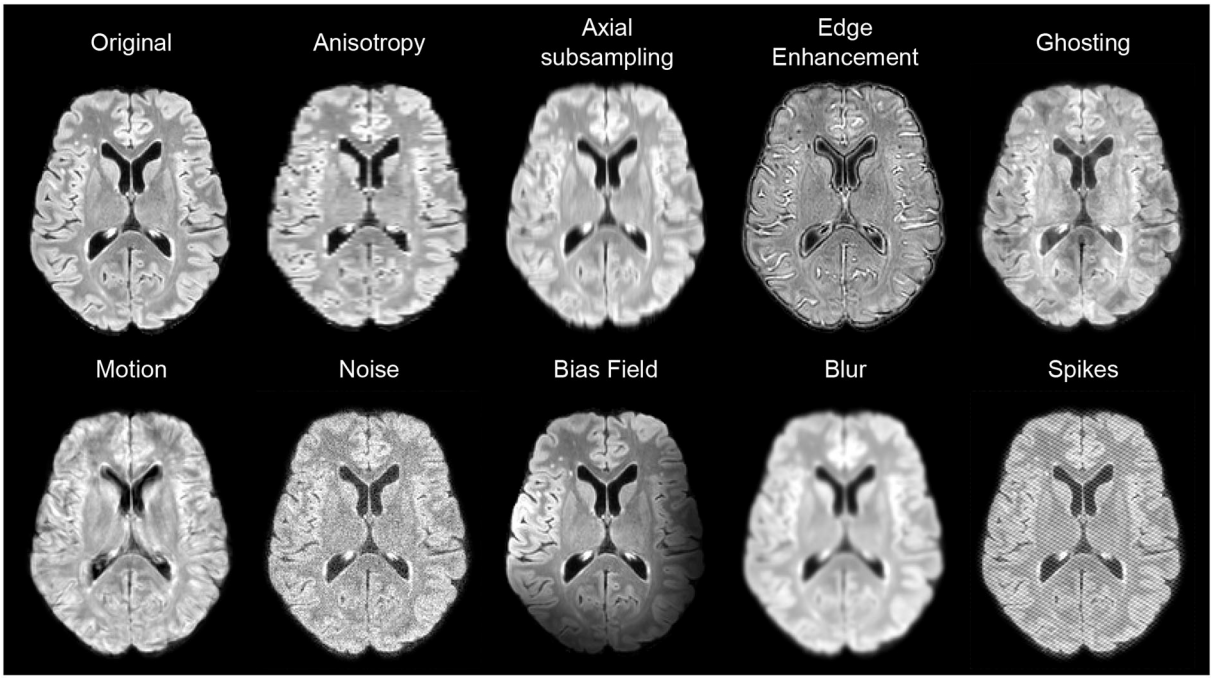

FIGURE 5
Examples of data augmentation applied on FLAIR images.

TABLE 1 Summary of the used datasets. For each dataset, the object count (Obj. Count) and the total volume (Tot. Vol. cm^3) represent, respectively, the total number and the total volume in cm^3 of lesions or new lesions (depending on the task).

Task	Dataset	Patients	Time-point	Raters	Obj. count	Tot. vol. (cm^3)	Clinical site/Scanners
MS lesion segmentation	ISBI	5	4-5	2	514	243	Single-site
	MSSEG' 16	15	1	7	512	367	Multi-site: three sites
	In-house	43	1	2	2,391	1,313	Multi-site
New MS lesion segmentation	MSSEG2 Training-set	40	2	4	123	23	Multi-site: 15 MRI scanners
	MSSEG2 Test-set	60	2	4	174	60	(GE scanners only in Test-set)

is composed of longitudinal time-points from 5 patients, the provided expert annotations focus on the lesion mask of each time-point independently from the others and do not provide new lesion masks. Thus, we use the 21 images independently. Second, the MSSEG'16 training-set (Commowick et al., 2016) contains 15 patients from three different clinical sites. Each FLAIR image is along with a consensus segmentation for MS lesions from seven human experts. Third, our in-house (Coupé et al., 2018) dataset is composed of 43 subjects diagnosed with MS. The images were acquired with different scanners and multiple resolutions and their lesion masks have been obtained by two human experts.

All images were pre-processed using the lesionBrain pipeline from the volBrain platform (Manjón and Coupé, 2016). First, it includes image denoising (Manjón et al., 2010). Second, an affine registration to MNI space is performed using the T1w modality, then the FLAIR is registered to the transformed T1w. Skull stripping and bias correction have been performed on the modalities, followed by the second denoising. Finally, the intensities have been normalized with kernel density estimation.

2.5.2. Two time-points datasets

The dataset provided by the MSSEG2-challenge (MICCAI, 2021) is used to train our method. The challenge dataset features a total of 100 patients with MS. For each patient, two 3D FLAIR sequence time-points have been acquired spaced apart by a 1–3 years period. The dataset has been split into 40 patients for

training and 60 patients for testing. A total of 15 different MRI scanners were used for the acquisition of the entire dataset. However, all images from GE scanners have been reserved only for the testing set to see the generalization capability of the algorithms. Reference segmentation on these data was defined by a consensus of four expert neuroradiologists.

For preprocessing, the challenge organizers proposed a docker[1] built with the Anima scripts. It includes bias correction, denoising, and skull stripping. In addition, we added a registration step to the MNI space using a FLAIR template, (i.e., the training and inference are performed in the MNI space, then the segmentation masks are transformed-back to the native space for evaluation).

Before challenge day, the testing set (the 60 patients) was not publicly available. Thus, to test our methods (refer to Section 3.1.1), we defined an internal validation subset from the 40 challenge training data. Of the 40 patients, six cases containing confirmed new lesions were kept out from the training-set and were used as an internal test-set. For the challenge evaluation (refer to Section 3.2), the model submitted to the challenge organizers was trained on the entire MSSEG2 training-set.

2.5.3. Dataset class imbalance

Anomaly detection/segmentation tasks, such as MS lesion segmentation, suffer from class imbalance where the positive class is scarce (refer to Johnson and Khoshgoftaar, 2019). Herein, the MSSEG2-challenge (MICCAI, 2021) dataset is composed of 100 patients (40 for training and 60 for test) and all the MS Lesions Segmentation datasets combined account for 64 patients and 79 images. Therefore, the number of image is similar. However, the class imbalance is highly different when evaluating the class imbalance using the number of objects to detect/segment (which represent MS lesions for the first task and new lesions for the second one) and their total volume for each dataset (refer to Table 1). Indeed, we see that the MSSEG2-challenge datasets (especially training-set) suffer from more severe under-representation of the positive class. Consequently, it will be more difficult to train a model for New MS lesion segmentation than for the task of single time-point MS lesion segmentation. Furthermore, it shows that MS lesion segmentation datasets could significantly enrich the training of New MS lesion segmentation models.

2.6. Implementation details

First, all models are trained on 3D image patches of size [64 × 64 × 64]. For the two time-points new lesion model, an ensemble of five networks (different training/validation data-split) is used. During inference, the consensus (prediction average) of the ensemble segmentation is taken. For each voxel, the two classes, output probabilities of the five networks are averaged, and the class with the highest probability is picked (new lesion voxel or not).

Second, the Dice-loss (soft DICE with probabilities as continuous values) is used as a loss function for the training of the single time-point MS lesion segmentation and the two time-points new lesion models. The mean-squared error is used as a loss function to train time-point synthesis models (inpainting and lesion generator models).

Finally, the experiments have been performed using PyTorch framework version 1.10.0 on Python version 3.7 of Linux environment with NVIDIA Titan Xp GPU 12 GB RAM. All models were optimized with Adam (Kingma and Ba, 2014) using a learning rate of 0.0001 and a momentum of 0.9.

2.7. Validation framework

2.7.1. 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), and the true positive rate (TPR, known as recall or Sensitivity).

$$Dice = \frac{2 \times TP}{(TP + FN) + (TP + FP)}, \tag{1}$$

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

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

Second, recent studies (i.e., Commowick et al., 2018) 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. We use the lesion detection F1 ($LesF_1$) score defined as

$$LesF_1 = \frac{2 \times S_L \times P_L}{(S_L + P_L)}, \tag{3}$$

where S_L is lesion sensitivity, i.e., the proportion of detected lesions and P_L is lesion positive predictive value, i.e., the proportion of true positive lesions. For result harmonization with challenge organizers and participants, the same evaluation tool is used, i.e., animaSegPerfAnalyzer (Commowick et al., 2018). All lesions that are smaller in size than $3mm^3$ are removed. For S_L, only ground-truth lesions that overlap at least

1 https://github.com/Inria-Empenn/lesion-segmentation-challenge-miccai21/

10% with segmented volume are considered positive. For a predicted lesion to be considered positive for P_L, it has to be overlapped by at least 65% and do not go outside by more than 70% of the volume.

Finally, to jointly consider the different metrics (i.e., segmentation and detection performance), it would be convenient to aggregate them into a single score. Thus, we propose the average of DICE and $LesF_1$ (Avg. Score) as an aggregation score for comparing different methods.

2.7.2. Statistical test

To assert the advantage of a technique obtaining the highest average score, we conducted a Wilcoxon test (i.e., paired statistical test) over the lists of metric scores. 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

Several experiments were conducted on our methods, including an ablation study and the comparison with state-of-the-art methods in competition during the challenge evaluation.

3.1. Internal validation

3.1.1. Ablation study

To evaluate each contribution of our training pipeline, Table 2 compares our full method with a baseline and other variations of our method on the internal validation dataset. The baseline in this experiment was trained with real time-points only and by using a classic data augmentation composed of orthogonal rotations and mirroring.

First, when using only transfer learning on top of the baseline, we measured an increase in DICE and TPR compared to the baseline but approximately the same $LesF_1$ and PPV. Second, when using only time-point synthesis pretraining on the top of the baseline, we obtained a significantly higher $LesF_1$ compared to the baseline and an increase in DICE. This variation also obtained the highest PPV at the expense of the lowest TPR. Third, when comparing the use of the proposed data augmentation, we see an increase in DICE and PPV but approximately the same $LesF_1$. Finally, when combining the transfer learning, time-point synthesis pre-training, and the proposed data-augmentation, we obtained the highest Avg. Score, DICE, $LesF_1$, and TPR.

3.1.2. The impact of longitudinal dataset size

Figure 6 shows the performance of our method when trained with different longitudinal dataset sizes. From the 34 patients available for the training with two time-points in Internal Validation settings (refer to Section 2.5.2), we tested the performance of our model when training on 34, 36, 17, 8, and 0 patients. In the case of 0 patients, our method performance was obtained using synthetic data only (i.e., Stage Two where only cross-sectional MS segmentation databases were used as described in Table 1). For the rest of the experiments, the reported number of patients with two time-points was used for the fine-tuning step (i.e., Stage Three).

First, for the baseline version (i.e., with neither pre-training nor data augmentation), the graph can be separated into two phases. From 0 to 17 patients, the graph shows an increase in both metrics. From 17 to 34 patients, metrics of baseline versions reach a plateau. Since the baseline is trained from scratch, its performance improves with the increase in dataset size. However, the performance increase is less significant for the second phase since it is more difficult to improve metrics when approaching their optimal value.

Second, for our method, the graph shows two phases. From 0 to 8 patients, the performance decreases slightly. From 8 to 34 patients, the graph shows a slow increase in metrics until plateauing. Since we use transfer learning and pretraining on synthetic data for our method, its performance does not depend only on the number of patients from MSSEG2 Training-set. The drop in performance in the first phase can be explained by the fact that using eight patients for fine-tuning is less effective than using the model trained on synthetic data only.

3.2. Challenge evaluation

To evaluate our method on the challenge dataset, Table 3 compares it to the leader-board state-of-the-art methods. Results of the top performing methods were reported from challenge-day results.

Besides the top-performing methods, Table 3 also includes the expert raters' performance to give an insight into human performance. Their performance is measured compared to each other, contrary to the top methods that are evaluated using consensus segmentation. Raters x vs. y means that we evaluate the performance of rater x when considering rater y segmentations as ground truth. Indeed, we consider that such a strategy can be more meaningful than the consensus segmentation in our case since the expert consensus already encodes the raters' segmentation and, thus, is unfair when compared to other strategies that did not participate in the consensus.

First, from the top five best-performing methods, LaBRI-IQDA (Kamraoui et al., 2021; our team's submission during the challenge-day) obtained the best score for the challenge. This method was similar to the proposed baseline with data augmentation. Second, the proposed method (results obtained after challenge-day) obtained the highest $LesF_1$ and Average

TABLE 2 The internal validation results for the ablation study.

Transfer learning	Time-point synthesis	Data augm.	Avg. Score	DICE	$LesF_1$	TPR	PPV
✓	✓	✓	**0.543***	**0.514***	**0.573***	**0.500***	0.546
✓	✗	✗	0.483	0.480	0.486	0.461	0.532
✗	✓	✗	0.501	0.461	0.541	0.384	**0.602***
✗	✗	✓	0.477	0.464	0.488	0.406	0.565
✗	✗	✗	0.469	0.449	0.489	0.413	0.534

✓ and ✗ symbolize using or not each contribution. Bold values indicate the best result for a metric and * indicates that the advantage is statistically significant (Wilcoxon test).

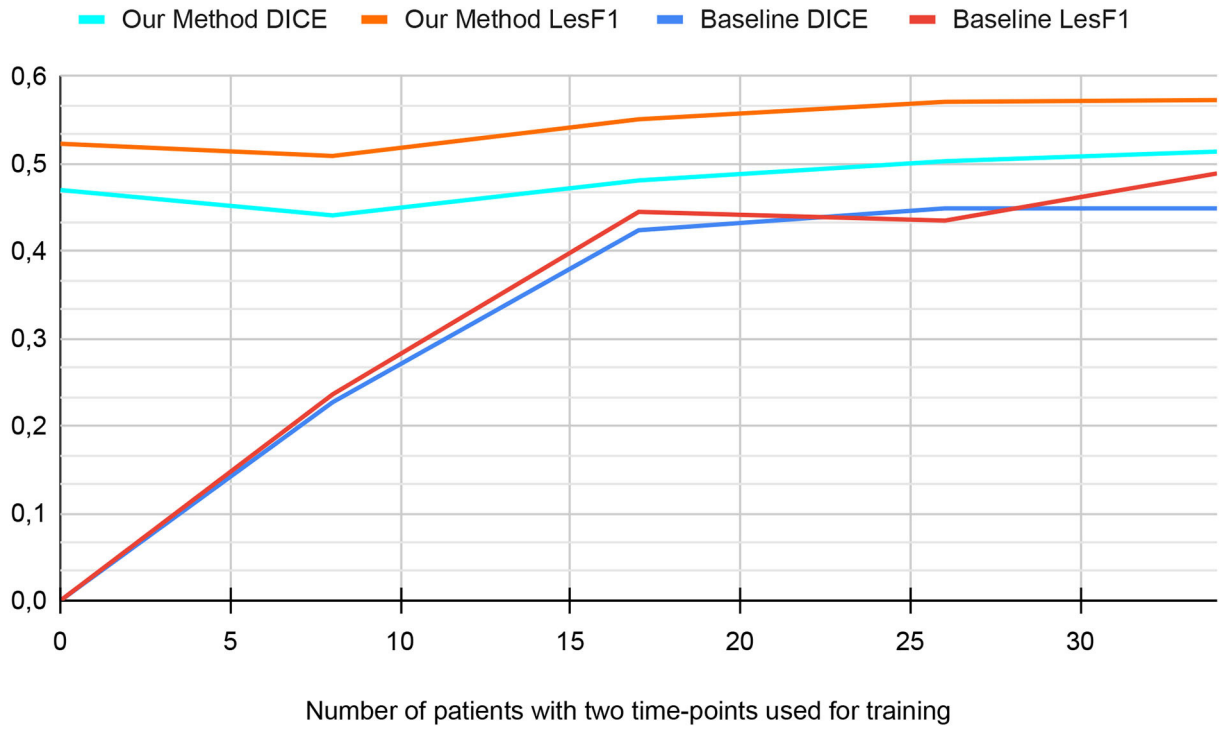

FIGURE 6
The performance in the internal validation of our method and the baseline based on the number of patients used for training (from MSSEG2 Training-set).

scores. Moreover, these both scores are significantly better than all the listed state-of-the-art methods. The DICE score obtained by MedICL was not significantly better than the one obtained by our method. Third, all but one (Empenn) leader-board automatic method obtained better DICE than raters segmentation. Our proposed method, LaBRI-IQDA, and MedICL even surpassed all raters in Average Scores.

Figure 7 shows the segmentation of new lesions by our proposed method. As a ground-truth reference, we compare the segmentation with the consensus segmentation of raters. We also compare each rater segmentation against their consensus. From the five segmentation, we see that our segmentation is the most accurate with the consensus. Each of the human experts Rater 2, Rater 3, and Rater 4 missed one or multiple lesions when segmenting this sample. Although Rater 1 did not miss any lesions, we see that our segmentation is the closest to the consensus.

Overall, our method obtained the best result in the MSSEG2 challenge evaluation (during the challenge and after). Moreover, the result of the experiments showed that our segmentation is objective and can produce more accurate segmentations than human raters.

4. Discussion

The transfer-learning from a single time-point MS lesion segmentation task is an effective method to train the model for the task of two time-points new MS lesion segmentation even with a small dataset. Indeed, it enables us to exploit the large

TABLE 3 Results of MSSEG2-challenge (MICCAI, 2021) evaluation.

	Experiment	Avg. Score	DICE	$LesF_1$
	Raters 1 vs. 2	0.466	0.426	0.507
	Raters 1 vs. 3	0.499	0.434	0.564
	Raters 1 vs. 4	0.434	0.382	0.486
Challenge-day	LaBRI-IQDA (Kamraoui et al., 2021)	0.507	0.498	0.515
Challenge-day	MedICL (Zhang et al., 2021)	0.503	**0.506**	0.5
Challenge-day	SNAC (Cabezas et al., 2021)	0.496	0.484	0.513
Challenge-day	Mediaire-B (Dalbis et al., 2021)	0.489	0.436	0.541
Challenge-day	Empenn (Masson et al., 2021)	0.478	0.423	0.532
	The Proposed Method	**0.523***	0.495	**0.550***

From top to bottom, the table shows the challenge raters' agreement on the segmentation compared to each other, the leader-board results of the challenge-day top methods, and the result of the method described in this article (obtained after challenge-day). For automatic methods, bold values indicate the best result for a metric, and * indicates that the advantage is statistically significant (Wilcoxon test).

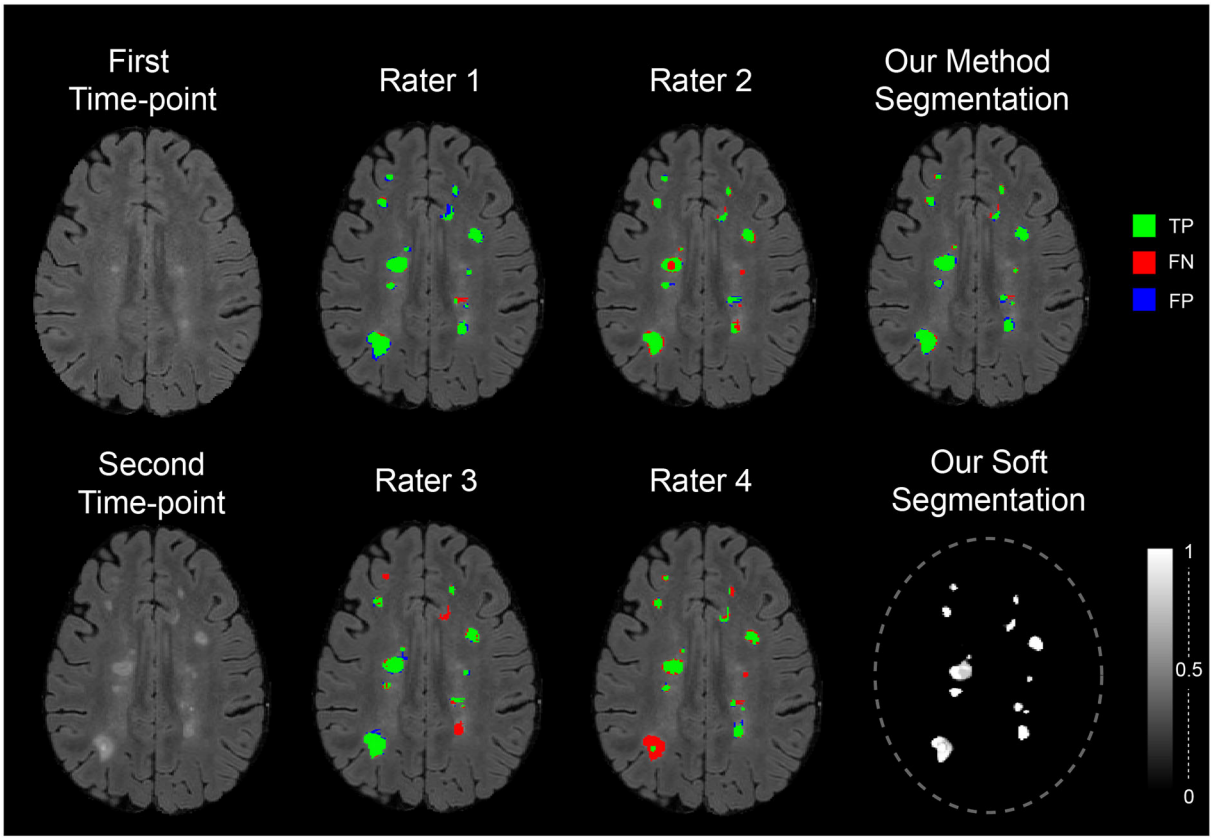

FIGURE 7
The segmentation of the proposed method and the expert rater on a sample image from MICCAI 2021—longitudinal multiple sclerosis lesion segmentation testing dataset. The segmentations are compared against the consensus of the four raters using the colors: green, red, and blue to symbolize TP, FN, and FP regions of new lesions.

available MS cross-sectional datasets compared to longitudinal datasets. In our case, the encoder for the first task was compatible with the siamese-encoder of the second task and thus was used to extract MS-relevant features from the two time-points. Additionally, we used a learnable aggregation module for time-points feature combination. Besides, by freezing the encoder weights after the transfer-learning from the first to the second task, we ensure that the extracted features in the second task

are dataset-independent from the second task dataset (smaller dataset). This independence ensures that the high performance of the proposed method is stable and generalizing.

Longitudinal time-points synthesis is an original approach on how to augment data diversity. It can be extended to other change detection tasks where longitudinal data are hard to acquire. According to the results of our experiments, this strategy turns out to be very effective when used as pretraining. Indeed, when the model is first pretrained with time-point synthesis, it is subject to a wider range of diversity, which aims to constrain the model to extract more generalizing features.

The proposed data augmentation method is an effective technique to make our learning process less dependent on MRI quality and acquisition artifacts. It simulates different acquisition conditions to enhance generalization and helps to better over-sample the available new lesions examples. Our data-augmentation comparison (refer to Table 2) showed the proposed augmentation method contributes to segmentation accuracy in both internal validation and challenge evaluation (i.e., MRI from scanners not seen during training).

The ablation study performed using the internal validation process showed that each contribution, taken separately, enhanced the segmentation accuracy. It also showed that when combining all contributions, we achieved the best results. Similarly, the challenge evaluation showed that the proposed method achieved better results than the best-performing methods of the challenge.

Our experiment in Section 3.1.2 has shown interesting behavior of our method when trained on only 8 patients (minor performance decrease compared to using synthetic data only). The fine-tuning and optimization by selecting the best weights combination based on a very limited validation set has foreseeably led to overfitting. Thus, it is advised that the number of samples and their quality (containing enough new MS lesions) are sufficient so the fine-tuning step could enhance the performance. If the labeled dataset is not sufficient, combining both synthetic and real data could also be explored.

Our study explored the possibility of using a similar task such as MS lesion segmentation to better train new MS lesion segmentation models. Transfer learning has led to satisfactory results. However, other methods for instance multi-task learning and consistency regularization should be explored likewise. Other of our experiments (that have not been covered in our paper) investigated such strategies on both single time-point MS and new MS lesion segmentation. Unfortunately, it is difficult to deal with the different class imbalances and complexities of both tasks which makes optimizing jointly over single time-point MS and new MS lesion segmentation harder. We believe that a training-set containing both the segmentation of new lesions and the segmentation of other lesions contained in both time points could lead the community to propose better segmentation/detection models.

Although it is sometimes difficult for experts to agree upon whether a lesion is new or not, their consistency in the segmentation of new lesions is even more difficult. This inconsistency, despite being mitigated by the consensus of several experts, will have repercussions on the quality of the segmentation accuracy. Thus, we believe that if there is interest in the quantification of new lesion volume, the output of models trained only on one modality (FLAIR) and for the task of new lesion segmentation should be taken with precaution. Combining the outputs of this model with another one trained on a single time-point with several modalities (T1w and FLAIR) could lead to better and more accurate segmentation.

Besides the detection of new lesions, another interesting biomarker for MS clinicians is the measurement of disappearing lesions. Our proposed method could potentially be used for this task by inverting the time-point order. However, it has not been validated in our study and requires the appropriate expert annotations.

5. Conclusion

In this article, we propose a training pipeline to deal with the lack of data for new MS lesion segmentation from two time points. The pipeline encompasses transfer learning from single time-point MS lesion segmentation, pretraining with time-point synthesis, and data-augmentation adapted for MR images. Our ablation study showed that each of our contributions enhances the accuracy of the segmentation. Overall, our pipeline was very effective for new MS lesions segmentation (Best score in MSSEG2-challenge; MICCAI, 2021) and can be extended to other tasks that suffer from longitudinal data scarcity.

Data availability statement

The data analyzed in this study was obtained from France Life Imaging (FLI)—Information Analysis and Management (IAM) node, the following licenses/restrictions apply: Users must subscribe to the challenge to get data access via Shanoir-NG (next generation). Requests to access these datasets should be directed to Shanoir-NG, https://shanoir.irisa.fr/shanoir-ng/challenge-request.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the patients/participants or patients/participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

RK: method design and implementation, experiment, coding, writing, and editing. BM: coding, writing, and editing. JM: writing and editing. PC: method design, experiment, writing, and editing. All authors contributed to the article and approved the submitted version.

Funding

This study 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 and RRI IMPACT), the French Ministry of Education and Research, and the CNRS for the DeepMultiBrain project. This study has also been supported by the PID2020-118608RB-I00 grants from the Spanish Ministerio de Economia, Industria Competitividad.

Acknowledgments

The authors gratefully acknowledge the support of NVIDIA Corporation with their donation of the TITAN Xp GPU used in this research.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bosc, M., Heitz, F., Armspach, J.-P., Namer, I., Gounot, D., and Rumbach, L. (2003). Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. *NeuroImage* 20, 643–656. doi: 10.1016/S1053-8119(03)00406-3

Bron, E. E., Klein, S., Papma, J. M., Jiskoot, L. C., Venkatraghavan, V., Linders, J., et al. (2021). Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. *NeuroImage* 31:102712. doi: 10.1016/j.nicl.2021.102712

Cabezas, M., Luo, Y., Kyle, K., Ly, L., Wang, C., and Barnett, M. (2021). "Estimating lesion activity through feature similarity: a dual path Unet approach for the MSSEG2 MICCAI challenge," in *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, 107.

Carass, A., Roy, S., Jog, A., Cuzzocreo, J. L., Magrath, E., Gherman, A., et al. (2017). Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage* 148, 77–102. doi: 10.1016/j.dib.2017.04.004

Cheng, M., Galimzianova, A., Lesjak, Ž., Špiclin, Ž., Lock, C. B., and Rubin, D. L. (2018). "A multi-scale multiple sclerosis lesion change detection in a multi-sequence MRI," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Springer), 353–360. doi: 10.1007/978-3-030-00889-5_40

Commowick, O., Cervenansky, F., and Ameli, R. (2016). "MSSEG challenge proceedings: multiple sclerosis lesions segmentation challenge using a data management and processing infrastructure," in *MICCAI*. Available online at: https://scholar.google.fr/scholar?hl=fr&as_sdt=0%2C5&q=MSSEG+challenge+proceedings&btnG=#d=gs_cit&t=1659870818068&u=%2Fscholar%3Fq%3Dinfo%3AZnRobGdVz6gJ%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Dfr

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., et al. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci. Rep.* 8, 1–17. doi: 10.1038/s41598-018-31911-7

Coupé, P., Tourdias, T., Linck, P., Romero, J. E., and Manjón, J. V. (2018). "Lesionbrain: an online tool for white matter lesion segmentation," in *International Workshop on Patch-based Techniques in Medical Imaging* (Springer), 95–103. doi: 10.1007/978-3-030-00500-9_11

Dalbis, T., Fritz, T., Grilo, J., Hitziger, S., and Ling, W. X. (2021). "Triplanar U-net with orientation aggregation for new lesions segmentation," in *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, 57.

Elliott, C., Arnold, D. L., Collins, D. L., and Arbel, T. (2013). Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. *IEEE Trans. Med. imaging* 32, 1490–1503. doi: 10.1109/TMI.2013.2258403

Ganiler, O., Oliver, A., Diez, Y., Freixenet, J., Vilanova, J. C., Beltran, B., et al. (2014). A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. *Neuroradiology* 56, 363–374. doi: 10.1007/s00234-014-1343-1

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). NNU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18, 203–211. doi: 10.1038/s41592-020-01008-z

Johnson, J. M., and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *J. Big Data* 6, 1–54. doi: 10.1186/s40537-019-0192-5

Kamraoui, R. A., Ta, V.-T., Manjon, J. V., and Coupé, P. (2021). "Image quality data augmentation for new MS lesion segmentation," in *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, 37.

Kamraoui, R. A., Ta, V.-T., Tourdias, T., Mansencal, B., Manjon, J. V., and Coupé, P. (2022). DeepLesionBrain: towards a broader deep-learning generalization for multiple sclerosis lesion segmentation. *Med. Image Anal.* 76:102312. doi: 10.1016/j.media.2021.102312

Khan, A. R., Khan, S., Harouni, M., Abbasi, R., Iqbal, S., and Mehmood, Z. (2021). Brain tumor segmentation using K-means clustering and deep learning with synthetic data augmentation for classification. *Microsc. Res. Techn.* 84, 1389–1399. doi: 10.1002/jemt.23694

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. Availableo online at: https://arxiv.org/pdf/1412.6980.pdf

Krüger, J., Opfer, R., Gessert, N., Ostwaldt, A.-C., Manogaran, P., Kitzler, H. H., et al. (2020). Fully automated longitudinal segmentation of new or enlarged

multiple sclerosis lesions using 3D convolutional neural networks. *NeuroImage* 28:102445. doi: 10.1016/j.nicl.2020.102445

Mårtensson, G., Ferreira, D., Granberg, T., Cavallin, L., Oppedal, K., Padovani, A., et al. (2020). The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *Med. Image Anal.* 2020:101714. doi: 10.1016/j.media.2020.101714

Manjón, J. V., and Coupé, P. (2016). volBrain: an online MRI brain volumetry system. *Front. Neuroinform.* 10:30. doi: 10.3389/fninf.2016.00030

Manjón, J. V., Coupé, P., Martí-Bonmatí, L., Collins, D. L., and Robles, M. (2010). Adaptive non-local means denoising of MR images with spatially varying noise levels. *J. Magn. Reson. Imaging* 31, 192–203. doi: 10.1002/jmri.22003

Manjón, J. V., Romero, J. E., Vivo-Hernando, R., Rubio, G., Aparici, F., de la Iglesia-Vaya, M., et al. (2020). "Blind MRI brain lesion inpainting using deep learning," in *International Workshop on Simulation and Synthesis in Medical Imaging* (Springer), 41–49. doi: 10.1007/978-3-030-59520-3_5

Masson, A., Le Bon, B., Kerbrat, A., Edan, G., Galassi, F., and Combes, B. (2021). "A NNUnet implementation of new lesions segmentation from serial FLAIR images of MS patients," in *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, 5.

MICCAI (2021). *Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*. Available online at: https://portal.fli-iam.irisa.fr/msseg-2/data/

Olivas, E. S., Guerrero, J. D. M., Martinez-Sober, M., Magdalena-Benedito, J. R., and Serrano, L. (2009). *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques: Algorithms, Methods, and Techniques*. IGI Global.

Omoumi, P., Ducarouge, A., Tournier, A., Harvey, H., Kahn, C. E., Louvet-de Verchère, F., et al. (2021). To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines). *Eur. Radiol.* 31, 3786–3796. doi: 10.1007/s00330-020-07684-x

Pérez-García, F., Sparks, R., and Ourselin, S. (2021). Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput. Methods Prog. Biomed.* 2021:106236. doi: 10.1016/j.cmpb.2021.106236

Salem, M., Cabezas, M., Valverde, S., Pareto, D., Oliver, A., Salvi, J., et al. (2018). A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis. *NeuroImage* 17, 607–615. doi: 10.1016/j.nicl.2017.11.015

Schmidt, P., Pongratz, V., Küster, P., Meier, D., Wuerfel, J., Lukas, C., et al. (2019). Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. *NeuroImage* 23:101849. doi: 10.1016/j.nicl.2019.101849

Shaw, R., Sudre, C., Ourselin, S., and Cardoso, M. J. (2018). "MRI K-space motion artefact augmentation: model robustness and task-specific uncertainty," in *International Conference on Medical Imaging with Deep Learning-Full Paper Track*.

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 1–48. doi: 10.1186/s40537-019-0197-0

Sudre, C. H., Cardoso, M. J., Ourselin, S., and Alzheimer's Disease Neuroimaging Initiative (2017). Longitudinal segmentation of age-related white matter hyperintensities. *Med. Image Anal.* 38, 50–64. doi: 10.1016/j.media.2017.02.007

Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., et al. (2018). "Training deep networks with synthetic data: bridging the reality gap by domain randomization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (IEEE)*, 969–977. doi: 10.1109/CVPRW.2018.00143

Tripathi, S., Chandra, S., Agrawal, A., Tyagi, A., Rehg, J. M., and Chari, V. (2019). "Learning to generate synthetic data *via* compositing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE)*, 461–470. doi: 10.1109/CVPR.2019.00055

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *J. Big Data* 3, 1–40. doi: 10.1186/s40537-016-0043-6

Zhang, H., Li, H., and Oguz, I. (2021). "Segmentation of new MS lesions with Tiramisu and 2.5 D stacked slices," in *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, 61.

Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., et al. (2020). Generalizing deep learning for medical image segmentation to unseen domains *via* deep stacked transformation. *IEEE Trans. Med. Imaging* 39, 2531–2540. doi: 10.1109/TMI.2020.2973595