Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

doi:10.3233/FAIA230679

Longitudinal Segmentation of Multiple Sclerosis Lesions Using nnU-Net Architecture

Eloy MARTÍNEZ-HERAS[a,1,*], Adrian VICENTE-GOMEZ[b,*], Francesc VIVÓ[a], Marcos DIAZ-HURTADO[b], Baris KANBER[c], Jordi CASAS-ROMA[b], Sara LLUFRIU[a] and Ferran PRADOS[b,c,d,e,1]

[a] *Neuroimmunology and Multiple Sclerosis Unit, Laboratory of Advanced Imaging in Neuroimmunological Diseases, Hospital Clinic Barcelona, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS) and Universitat de Barcelona, Barcelona, Spain*

[b] *e-Health Center, Universitat Oberta de Catalunya, Barcelona, Spain*

[c] *Centre for Medical Image Computing, University College London, London, United Kingdom*

[d] *Queen Square MS Centre, Department of Neuroinflammation, UCL Institute of Neurology, Faculty of Brain Sciences, University College London, London, United Kingdom*

[e] *National Institute for Health Research Biomedical Research Centre at UCL and UCLH, London, United Kingdom*

Eloy Martínez-Heras - https://orcid.org/0000-0001-9937-3162
Adrián Vicente-Gomez - https://orcid.org/0009-0001-0421-559X
Francesc Vivó - https://orcid.org/0000-0002-6409-1197
Marcos Diaz-Hurtado - https://orcid.org/0000-0003-1528-5873
Baris Kanber: https://orcid.org/0000-0003-2443-8800
Jordi Casas-Roma - https://orcid.org/0000-0002-0617-3303
Sara Llufriu - https://orcid.org/0000-0003-4273-9121
Ferran Prados - https://orcid.org/0000-0002-7872-0142

Abstract. Multiple sclerosis is a chronic autoimmune disease that affects the central nervous systems. The detection of new lesions through conventional magnetic resonance imaging is particularly important in the management of people with multiple sclerosis. The advancements in machine learning technology in recent years have significantly transformed the analysis of medical images for multiple sclerosis, particularly in identifying and segmenting lesions., improving the accuracy and efficiency of this process. In this context, the objective of this work is to develop a system for the detection of new lesions in people with multiple sclerosis using consecutive magnetic resonance scans. The proposed system uses only pre-processed FLAIR images to train a nnU-Net, a type of deep learning architecture that has been proven very successful for image segmentation tasks. The resulting model is able to generate masks that highlight changes between baseline and follow-up images, identifying new lesions that may have appeared in the meantime. The model achieved an average Dice score coefficient

[1] Corresponding Authors: Eloy Martínez-Heras, EMARTIND@recerca.clinic.cat and Ferran Prados, fpradosc@uoc.edu.
* joint first authorship

of 0.58 on the evaluation set. Overall, this work demonstrates the potential of machine learning tools for improving the detection and monitoring of multiple sclerosis lesions in clinical practice, particularly in the context of longitudinal studies.

Keywords. Longitudinal, segmentation, lesions, multiple sclerosis, MRI

1. Introduction

Multiple sclerosis (MS) is a complex neurological disease that affects millions of people worldwide. Assessing the presence of new MS lesions and evidence of new disease activity is crucial for evaluating the efficiency of disease-modifying therapies [1,2]. Conventional magnetic resonance imaging (MRI) serves as an essential hallmark for MS progression, and an indispensable tool for diagnosing, monitoring disease activity, and gauging treatment response [3,4]. Despite its importance, MRI techniques have several limitations, including low specificity and sensitivity for assessing focal tissue damage, noise, and the presence of artifacts [5]. Despite this, longitudinal lesion segmentation is a research field with increasing interest. Recent advances in machine learning and the organization of several challenges in the field over the last decade, have eased the appearance of deep learning based solutions for detecting new lesions [6].

Although the task of detecting new lesions may seem straightforward, assessing the results can involve different considerations. Unlike other segmentation tasks like tissue segmentation or parcellation, the areas to detect new lesions can vary in size, with the majority being small or very small, and can appear in different locations of the brain tissue, predominantly in the white matter but also in the gray matter. Detecting the absence of new lesions is crucial from both a methodological and clinical perspective. While the appearance of new lesions is the primary sign of disease progression in MS, a stable MS patient may not have any new lesions. Therefore, if an algorithm correctly identifies the absence of new lesions in people with MS who have not developed new lesions, it should be considered a good segmentation, even if most segmentation metrics evaluate it as 0. As a result, the effectiveness of a new lesion segmentation technique should be evaluated based on its ability to detect new lesions and its ability to prevent over-segmentation in the absence of new lesions.

Manual delineation, while considered the gold standard, provides higher reliability in detecting enlarging lesions and new lesions in regions of accumulated damage, such as periventricular areas. However, it is time-consuming and subject to inter-observer variability, which can significantly impact the accuracy of the lesion assessment [7,8]. In response to these challenges, automatic lesion segmentation has emerged as a promising alternative, It offers more consistent segmentations, reproducible results and improved processing speed. The MS research community would greatly benefit from a common dataset for accurate estimation and comparison of proposed methods. Publicly available datasets, such as those released by ISBI 2015 [9] and MSSEG-2 challenges, provide a limited but valuable resource [10]. The recent MSSEG-2 challenge, which focuses solely on new lesions, has been another vital step for benchmarking the current methods on the field towards the potential adoption of these methods in clinical practice.

In recent years, techniques for longitudinal MS lesion segmentation have evolved significantly [6,9,10]. From image subtraction and thresholding to Bayesian

generative models and deep neural network-based methods, such as convolutional neural networks (CNN), substantial progress has been made in this field. The nnU-Net is a powerful and flexible 3D CNN that has demonstrated remarkable performance in various medical imaging tasks [11]. Its benefits include robustness against data class imbalance, which is a common issue in MS segmentation, and the ability to adapt to different datasets with minimal modifications. We hypothesize that the utilization of nnU-Net will significantly enhance the detection of new MS lesions in longitudinal MRI studies.

2. Method

In this section we introduce the used dataset, the data preprocessing strategy and the implemented method for segmenting longitudinal MS lesions.

2.1. Dataset

The dataset used for the training model has been obtained by combining two different open datasets: MSSEG-2 (https://portal.fli-iam.irisa.fr/msseg-2/data/) and Open MS Data (https://github.com/muschellij2/open_ms_data). Respectively, each of the datasets contributes with 40 and 20 distinct patients, each with their baseline and follow-up FLAIR images (ranging from 80 to 1000 days apart between scans), as well as the mask for new lesions. Therefore, there will be a total of 60 patients. It is important to note that 11 of them belonging to the MSSEG-2 dataset did not have any new lesions, so their mask for new lesions is blank. This aspect should be considered when elaborating the model of training with this dataset.

2.2. Preprocessing

MRIs are taken under different acquisition protocols, and although efforts are made to replicate the same parameters of the initial acquisition (baseline) when obtaining follow-up images, it is necessary to standardize certain characteristics among the different images of the dataset involved. The following preprocessing steps ensure that the model being trained is based on more homogeneous MRI characteristics across timepoints, resulting in a more reliable final model: orientation to MNI, skull stripping and intensity inhomogeneity correction (see Figure 1).

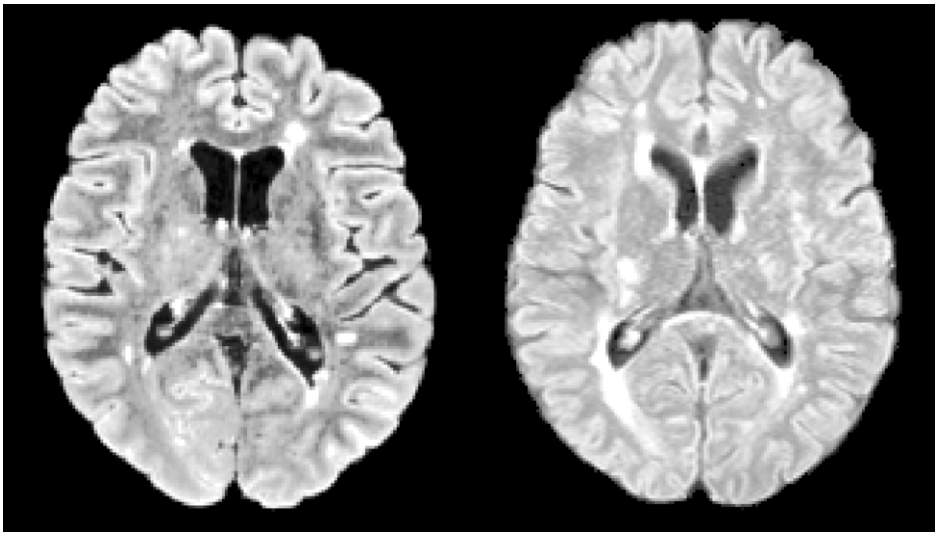

Figure 1. Example of the MSSEG-2 (left) and the Open MS Data (right) images.

A consistent orientation has been established for all images in the final dataset. Specifically, the Montreal Neurological Institute (MNI) coordinate system ("Neurological") is used. This registration and reorientation phase entails aligning the FLAIR images with the MNI template. This procedure uses a 6-degree-of-freedom (6 DOF) rigid registration transformation to ensure consistency across the dataset and improve the learning process. After orienting the image, we proceeded with the skull removal from the FLAIR images. This was carried out using the HD-BET algorithm [12]. The result is crucial for accurate subsequent analyses and model training avoiding the appearance of spurious results outside the white matter and gray matter brain tissues. Finally, we correct the image intensities for inhomogeneities due to coil uniformities, field strength or other biases [13]. For bias inhomogeneity correction, we utilized the N4 algorithm [14] to achieve a more uniform intensity of the whole FLAIR images of the dataset. Before starting the training process, images were cropped to include only the brain area. A z-score intensity normalization with mean 0 and standard deviation 1 was applied to each subject to regularize the intensity values.

2.3. nnU-Net architecture

The nnU-Net (no-new-Net) is a framework proposed by Insesee et al. [11] that proposes an architecture based on the U-Net (see Figure 2), which is able to adapt to the proposed dataset since up to now each case of use forced to specialize the architecture of the network. Baseline and a follow-up FLAIR scan were used as model inputs. The model output is a binary 3D prediction mask of new lesions in the follow-up scan. The nnU-Net architecture includes a contracting and an expanding path. Each resolution level in the contracting path contains two convolutional layers with a $3 \times 3 \times 3$ convolution kernel, followed by instance normalization and a leaky ReLU (rectified linear unit) activation, similar to ReLU activations used in the original architecture but with negative slope with value 0.01 which allows a more stable network training and improved model performance.

At each level of the contracting and expanding paths a combination of convolutional layer, stride and activations were used (see Figure 2). By incorporating skip connections, features extracted from the contracting path are concatenated with features obtained from the expanding path at each respective resolution level.

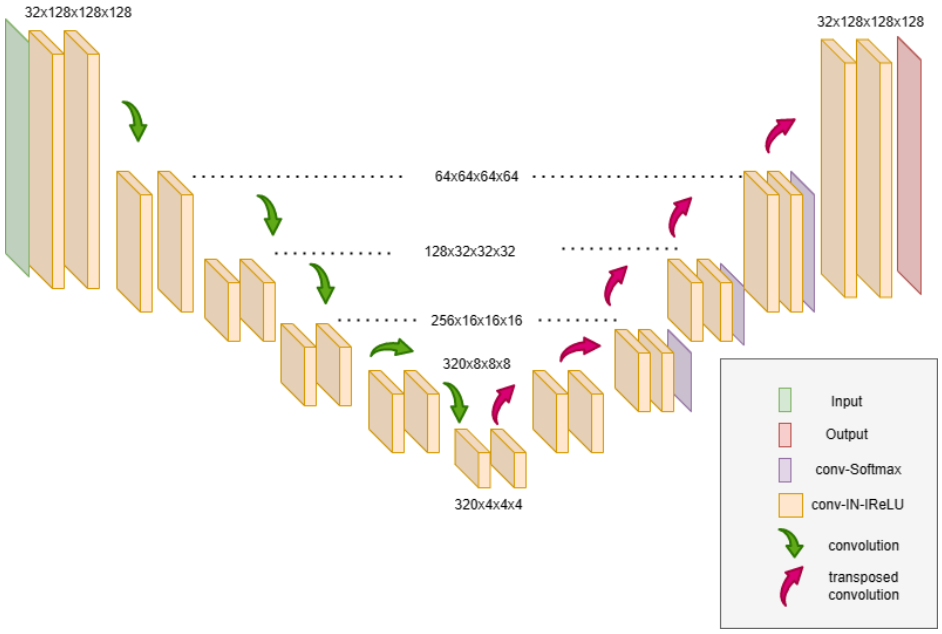

Figure 2 Schema of our nnU-Net architecture.

2.4. *Hyperparameters*

We used the Adam optimizer with an initial learning rate of 0.01, defining an epoch as iteration over 250 training batches, if the value of training and validation losses is not reduced by at least 0.005 on passing 30 epochs, the value of the learning rate is reduced by a factor of 5. On the other hand, no early stopping criterion was used, but a maximum threshold of 300 epochs was established, carrying out a checkpoint to be able to resume training in case of equipment failure every 25 epochs.

2.5. *Training*

The Dice score was selected as the evaluation metric for training progress. Training was carried out using a 5-fold cross-validation technique, which offers a dependable assessment of the model's capabilities while maintaining manageable computational requirements. In this process, for each fold, a random selection of 80% (48) was allocated for model training, and the remaining 20% (12) were designated for validation. Furthermore, an inherent feature of nnU-Net's default architecture is its use of data augmentation to enrich our training data and enhance the model's ability to generalize.

2.6. Evaluation metrics

Two classical evaluation metrics have been used for assessing the goodness of the proposed solution: Dice score coefficient and recall. Mean and standard deviation have been computed for each metric and dataset. If no lesions were present in the ground truth, and no lesions are detected by the method, it has been assumed the best result, for example 1 in Dice score coefficient instead of 0.

3. Results

The nnU-Net model achieved a good performance in terms of Dice score coefficient in the training process, with the measure ranging between 0.3 and 0.6. The average value for the final model of each k-fold is presented in Table 1.

Table 1. Average Dice score coefficients obtained for the final model of each k-fold.

k-fold	Dice score
0	0.48
1	0.37
2	0.52
3	0.50
4	0.49

The testing set consisted of 12 MR FLAIR images, out of which 3 did not have new lesions. The evaluation metrics obtained for the test set are shown in Table 2.

Table 2. Individual results, mean and standard deviation of the evaluation metrics for the test set.

Image	Dice score	Recall
MS_00001	0.37	0.75
MS_00006	1.00	1.00
MS_00007	0.00	0.00
MS_00010	0.81	0.73
MS_00014	0.00	0.00
MS_00017	1.00	1.00
MS_00019	0.50	0.39
MS_00021	0.71	0.63
MS_00028	0.57	0.46
MS_00042	0.65	0.74
MS_00047	0.38	0.28
MS_00056	1.00	1.00
All	**0.58±0.35**	**0.62±0.36**

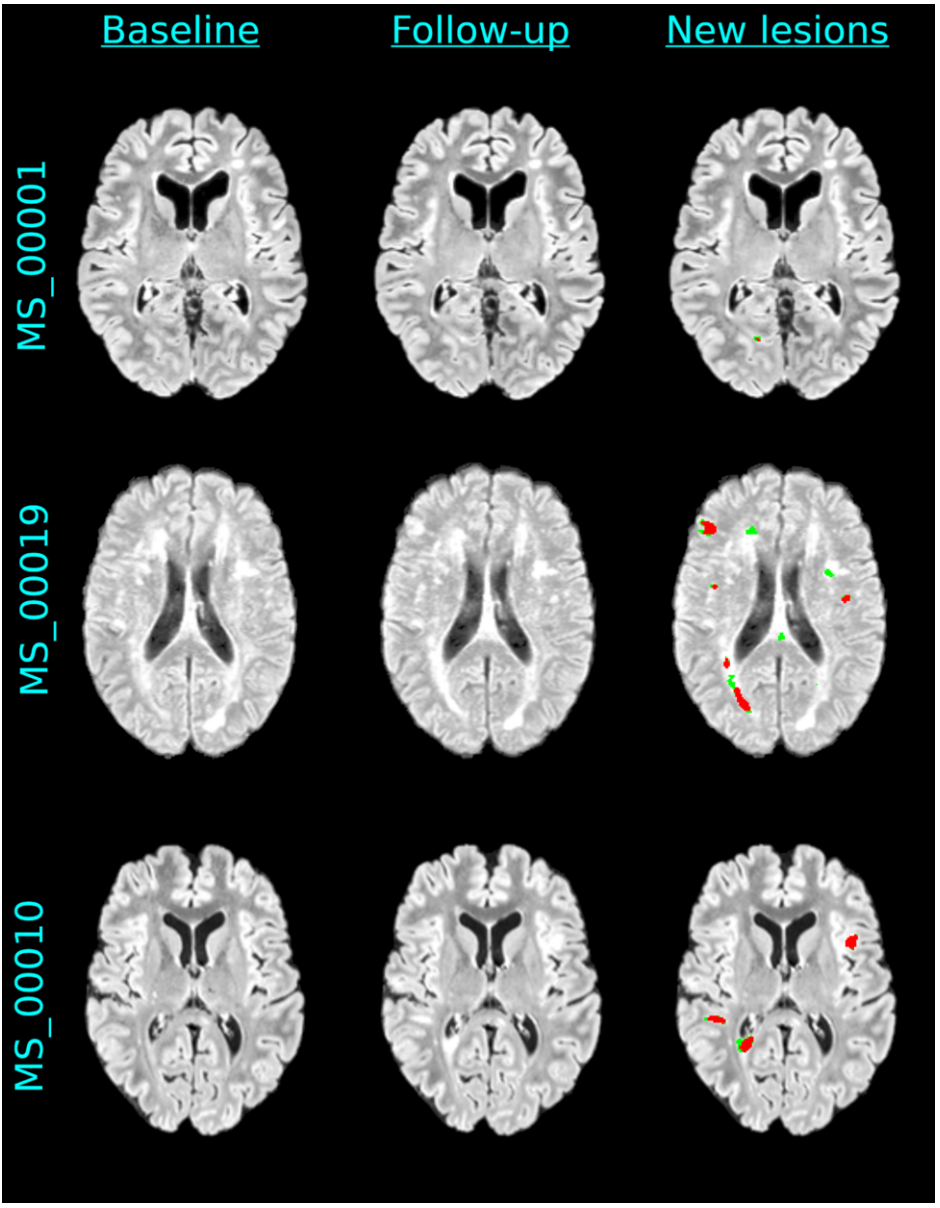

Figure 3. The mask produced by the model is shown in green, while the expert-segmented area is represented in red. The figure displays three distinct examples of people withMS, with Dice score coefficients of 0.37, 0.50, and 0.81, respectively.

4. Discussion

The primary aim of this work was to develop a model that can accurately detect new MS lesions in MRI scans taken at different timepoints. By comparing our

results with those obtained in MSSEG-2 Challenge, we observed that the model generated achieved an average Dice score coefficient of 0.58, which reflected a good performance compared to the majority of models proposed in MSSEG-2. Our work presents a self-customisation of a publicly available nnU-Net architecture [11]. The presented method uses the same architecture as Basaran et al. [15] but there are important implementation differences that need to be highlighted. Unlike them, we changed the preprocessing steps order by first registering to the MNI space, followed by skull stripping, and finally applying bias field correction [15]. The reason for first performing registration followed by skull stripping, is because we wanted to take advantage of the skull as an immutable part of the head for getting an accurate alignment between the two timepoints. We also normalize the intensity of the images using a z-score to minimize differences between images coming from different scanners. Moreover, during the training process, we used Dice score coefficient only for assessing the goodness of the results and the hyperparameter optimization follows two different strategies, Adam vs Nesterov. We also used different model hyperparameters to adapt the training to our more limited hardware.

While the model exhibited proficiency in identifying new lesions, it faced difficulties in accurately identifying the smallest lesions as exemplified by the MS_00001 case in Figure 3. This case features small new lesions and a suboptimal Dice score coefficient of 0.37 in comparison to other instances. The training process emphasized optimizing the Dice score coefficient value, which led to better detection of larger lesions at the expense of smaller ones. This is because larger lesions have a more substantial impact on this similarity coefficient. Despite this drawback, the selected model exhibited a relatively high success rate in identifying larger lesions, as demonstrated in Figure 3. In the accuracy values of Table 2 reveal solid performance, with a 62% correct identification rate for voxels with lesions. While the model's performance may seem moderate, it's crucial to remember that it acts as a diagnostic support tool, requiring expert validation before making final decisions. By effectively directing physicians to potential new lesions, the model substantially reduces the time needed to analyze the entire brain volume.

This work has certain limitations that must be acknowledged. Firstly, the method was developed using a reduced dataset and will require further testing with real clinical data and larger datasets to assess its overall effectiveness more thoroughly. We used the Dice score coefficient alone for assessing the training progress, it would be good to explore other metrics.

5. Conclusions

This work presents a nnU-Net model capable of detecting new lesions between MRIs performed on the same subject followed up for MS, where both images were taken months or years apart. The presented solution achieves good performance using publicly available data. Future work should explore the performance of the proposed method using a larger dataset from a clinical setting where scans can be obtained from different scanner manufacturers and with variable acquisition parameters.

Code availability

The code derived from this publication is publicly available at: https://github.com/ADaS-Lab/nnUnet-LongMS/. In order to ease the accessibility of nnU-Net, we have developed a Docker container that simplifies the execution of the models on new data. This container includes all the required dependencies, libraries, and scripts for running the pre-trained models. By using this container, users can easily deploy the models using their own data without having to worry about installing additional software or configuring complex environments.

Data availability

The data described in this manuscript are in the public domain at MSSEG2 (https://portal.fli-iam.irisa.fr/msseg-2/data/) and Open MS Data (https://github.com/muschellij2/open_ms_data) websites.

Acknowledgments

This work was partially funded by the Instituto Carlos III (ISCIII), co-funded by the European Union through the Plan Estatal de Investigación Científica y Técnica y de Innovación 2015-2024 (PI15/00587, PI18/01030 and PI21/01189 to SL) and National Institute for Health Research Biomedical Research Centre at UCL and UCLH.

References

[1] S.E. Hughes, and G. Macaron, Fast Facts: Multiple Sclerosis: 5th edition. A new era of disease modification and treatment, Karger Medical and Scientific Publishers, 2021.

[2] H.L. Weiner, and J.M. Stankiewicz, Multiple Sclerosis: Diagnosis and Therapy, John Wiley & Sons, 2012.

[3] M.P. Wattjes, O. Ciccarelli, D.S. Reich, B. Banwell, N. de Stefano, C. Enzinger, F. Fazekas, M. Filippi, J. Frederiksen, C. Gasperini, Y. Hacohen, L. Kappos, D.K.B. Li, K. Mankad, X. Montalban, S.D. Newsome, J. Oh, J. Palace, M.A. Rocca, J. Sastre-Garriga, M. Tintoré, A. Traboulsee, H. Vrenken, T. Yousry, F. Barkhof, À. Rovira, Magnetic Resonance Imaging in Multiple Sclerosis study group, Consortium of Multiple Sclerosis Centres, and North American Imaging in Multiple Sclerosis Cooperative MRI guidelines working group, 2021 MAGNIMS-CMSC-NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis, *Lancet Neurol.* **20** (2021) 653–670.

[4] A.J. Thompson, B.L. Banwell, F. Barkhof, W.M. Carroll, T. Coetzee, G. Comi, J. Correale, F. Fazekas, M. Filippi, M.S. Freedman, K. Fujihara, S.L. Galetta, H.P. Hartung, L. Kappos, F.D. Lublin, R.A. Marrie, A.E. Miller, D.H. Miller, X. Montalban, E.M. Mowry, P.S. Sorensen, M. Tintoré, A.L. Traboulsee, M. Trojano, B.M.J. Uitdehaag, S. Vukusic, E. Waubant, B.G. Weinshenker, S.C. Reingold, and J.A. Cohen, Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria, *Lancet Neurol.* **17** (2018) 162–173.

[5] R. Bakshi, A.J. Thompson, M.A. Rocca, D. Pelletier, V. Dousset, F. Barkhof, M. Inglese, C.R.G. Guttmann, M.A. Horsfield, and M. Filippi, MRI in multiple sclerosis: current status and future prospects, *Lancet Neurol.* **7** (2008) 615–625.

[6] M. Diaz-Hurtado, E. Martínez-Heras, E. Solana, J. Casas-Roma, S. Llufriu, B. Kanber, and F. Prados, Recent advances in the longitudinal segmentation of multiple sclerosis lesions on magnetic resonance imaging: a review, *Neuroradiology*. **64** (2022) 2103–2117.

[7] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S.C. Pop, P. Girard, R. Améli, J.-C. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, T. Glatard, J. Beaumont, S. Doyle, F. Forbes, J. Knight,

A. Khademi, A. Mahbod, C. Wang, R. McKinley, F. Wagner, J. Muschelli, E. Sweeney, E. Roura, X. Lladó, M.M. Santos, W.P. Santos, A.G. Silva-Filho, X. Tomas-Fernandez, H. Urien, I. Bloch, S. Valverde, M. Cabezas, F.J. Vera-Olmos, N. Malpica, C. Guttmann, S. Vukusic, G. Edan, M. Dojat, M. Styner, S.K. Warfield, F. Cotton, and C. Barillot, Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure, *Sci. Rep.* **8** (2018) 13650.

[8] O. Commowick, M. Kain, R. Casey, R. Ameli, J.-C. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, S. Camarasu-Pop, T. Glatard, S. Vukusic, G. Edan, C. Barillot, M. Dojat, and F. Cotton, Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset, *Neuroimage*. **244** (2021) 118589.

[9] A. Carass, S. Roy, A. Jog, J.L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C.H. Sudre, M. Jorge Cardoso, N. Cawley, O. Ciccarelli, C.A.M. Wheeler-Kingshott, S. Ourselin, L. Catanese, H. Deshpande, P. Maurel, O. Commowick, C. Barillot, X. Tomas-Fernandez, S.K. Warfield, S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, A. Jesson, T. Arbel, O. Maier, H. Handels, L.O. Iheme, D. Unay, S. Jain, D.M. Sima, D. Smeets, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, P.-L. Bazin, P.A. Calabresi, C.M. Crainiceanu, L.M. Ellingsen, D.S. Reich, J.L. Prince, and D.L. Pham, Longitudinal multiple sclerosis lesion segmentation: Resource and challenge, *Neuroimage*. **148** (2017) 77–102.

[10] O. Commowick, B. Combès, F. Cervenansky, and M. Dojat, Automatic methods for multiple sclerosis new lesions detection and segmentation, Frontiers Media SA, 2023.

[11] F. Isensee, P.F. Jaeger, S.A.A. Kohl, J. Petersen, and K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, *Nat. Methods*. **18** (2021) 203–211.

[12] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick, M. Bendszus, K.H. Maier-Hein, and P. Kickingereder, Automated brain extraction of multisequence MRI using artificial neural networks, *Hum. Brain Mapp.* **40** (2019) 4952–4964.

[13] M. Ganzetti, N. Wenderoth, and D. Mantini, Intensity Inhomogeneity Correction of Structural MR Images: A Data-Driven Approach to Define Input Algorithm Parameters, *Front. Neuroinform.* **10** (2016) 10.

[14] 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 Trans. Med. Imaging*. **29** (2010) 1310–1320.

[15] B.D. Basaran, P.M. Matthews, and W. Bai, New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation, *Front. Neurosci.* **16** (2022) 1007453.