Artificial Intelligence Research and Development
T. Alsinet et al. (Eds.)

doi:10.3233/FAIA240420

Multiclass Lesion Detection Using Longitudinal MRI in Multiple Sclerosis

Ander ELKOROARISTIZABAL [a,1], Francesc VIVÓ [b], Albert CALVI [b], Elisabeth SOLANA [b], Elisabet LOPEZ-SOLEY [b], Salut ALBA-ARBALAT [b], Marcos DIAZ-HURTADO [a], Baris KANBER [c], Jordi CASAS-ROMA [a], Sara LLUFRIU [b], Ferran PRADOS [a,c,d,e,2] and Eloy MARTÍNEZ-HERAS [b,1,2].

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

Ander Elkoroaristizabal - https://orcid.org/0009-0000-8576-7494,
Francesc Vivó - https://orcid.org/0000-0002-6409-1197,
Albert Calvi - https://orcid.org/0000-0002-1953-2803,
Elisabeth Solana - https://orcid.org/0000-0001-7973-2439,
Elisabet Lopez-Soley - https://orcid.org/0000-0002-2183-8094,
Salut Alba-Arbalat - https://orcid.org/0000-0003-1662-2317,
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,
Eloy Martínez-Heras - https://orcid.org/0000-0001-9937-3162

[1] These authors contributed equally.
[2] Corresponding Authors: Ferran Prados, fpradosc@uoc.edu and Eloy Martínez-Heras, EMARTIND@recerca.clinic.cat.

Abstract. Accurate detection of white matter (WM) lesions is essential for diagnosing and monitoring Multiple Sclerosis (MS), but manual lesion identification is challenging and time-consuming. This study employs the "no new U-Net" (nnU-Net) version 2 architecture to enhance the lesion segmentation process. We trained our model with a fine-tuned version of the default nnU-Net configuration incorporating extreme oversampling and a smaller learning rate to improve new or evolving lesion detection. Results showed that our nnU-Net v2 achieved a F1 score of 0.73 for baseline lesions and 0.75 for new or evolving lesions, demonstrating notable performance in identifying both types of lesions, and that the model generalized well to the MSSEG-2 dataset. This study highlights the capabilities of the nnU-Net v2 architecture for robust WM lesion detection in longitudinal cohorts. The final phase involved packaging our top-performing ensemble of models into a Docker container for easy usage, enabling the automatic distinction between baseline and new or evolving lesions.

Keywords. Multiple Sclerosis, MRI, Deep Learning

1. Introduction

Accurate identification and monitoring of white matter (WM) lesions is crucial for diagnosing and tracking People with Multiple Sclerosis (PwMS), as they serve as indicators of disease progression and evaluate treatment efficacy. Historically, the emphasis was on cross-sectional images, which capture data at individual timepoints. However, the updated McDonald criteria [1] transformed the diagnostic approach for PwMS. This criterion emphasized the importance of spatially and temporally assessing disease progression, highlighting the significance of techniques focused on longitudinal imaging. Recent advances have significantly evolved the applicability of the diagnostic criteria in clinical settings, with a palpable shift from traditional machine learning approaches towards more advanced deep learning techniques, including the automation of these tools to improve diagnostic accuracy [2].

In the MS research field, harmonizing data for consistent longitudinal analysis remains a significant challenge. Recently, the Open MS Data dataset [3] and the MSSEG-2 challenge [4] dataset have aimed to address this limitation. The MSSEG-2 dataset includes MRI data from 100 individuals with PwMS, featuring initial and follow-up 3D FLAIR scans taken one to three years apart, sourced from multiple centers using 15 distinct MRI scanners. In this challenge, 28 of 30 proposed methods employed Convolutional Neural Networks (CNNs), with U-net and its variants being popular choices [5,6,7,8,9,10]. However, variability among labelers, highlighted by discrepancies in Dice and F1 scores, raises concerns about the reliability of MRI interpretations and underscores the need for standardized protocols and consistent labeling methodologies. This inconsistency can impact the efficacy of models, as they might inadvertently learn from the discrepancies rather than authentic features. Furthermore, this variability complicates the benchmarking of segmentation algorithms, potentially increases costs and timelines due to the need for re-evaluations, and might even influence clinical decisions in real-world settings. This underlines the pressing need for more standardized protocols and consistent labeling methodologies.

The "no new U-Net" (nnU-Net) architecture is a deep learning-based framework specifically oriented for imaging segmentation tasks. Its primary goal is to standardize and automate the majority of the design choices related to parameter configurations [11].

The nnU-Net version 2 introduces significant enhancements, particularly in its support for medical imaging, improved preprocessing steps, and more robust model training procedures for efficient segmentation. This feature allows for a nuanced approach to training, letting focus on specific target regions in the images. Instead of training on broader, generalized labels, it can now fine-tune its models to identify and differentiate between intricate structures and regions within datasets. In this article, we aim to apply the nnU-Net v2 architecture to enhance results in the MSSEG-2 challenge, with a specific focus on different label segmentation. Our primary objectives are to explore the existing limitations of the challenge concerning the segmentation of baseline and follow-up lesions and to demonstrate how the inherent capability of nnU-Net for multiclass label segmentation offers a promising solution to recognize baseline as well as new or evolving lesions.

2. Methods

Our dataset consists of 117 collections from PwMS, each containing two 3D-FLAIR images (initial and subsequent assessment). These images were sourced from the ImaginEM research team from Hospital Clínic with the gold standard annotations provided by specialists from our center. The Ethics Committee at the Hospital Clinic of Barcelona approved the study, and all participants gave informed written consent for research and publication.

2.1. Preprocessing

We initially performed skull stripping from FLAIR images. This procedure, conducted using the HD-BET algorithm [12], ensures that subsequent modeling focuses solely on the brain's white and gray matter tissue, eliminating potential distractions from non-brain tissues. Following this, the images were aligned with the Montreal Neurological Institute (MNI) coordinate system through a 6-degree-of-freedom (6 DOF) rigid registration transformation. This step is critical for ensuring consistency and optimizing the training process. Finally, we addressed the intensity inhomogeneities in the FLAIR images, which can arise from factors such as coil uniformities or field strength variations, using the N4 algorithm [13].

2.2. Training and Validation Sets

Between subjects, both baseline and new or evolving lesions demonstrate notable differences in their count and respective volumes, see Table 1. The most noticeable thing is that 75 of the cases in our dataset do not have new or evolving lesions, which constitutes almost 65% of the total cases. Although having cases without new or evolving lesions is common and indeed part of the challenge of the task at hand, having such a high percentage of cases without lesions increases the difficulty of detecting these lesions.

Table 1. Summary statistics of the dataset.

	Baseline lesions	New or evolving lesions
Number of lesions [n]	64 ± 41	2 ± 4
Total Volume [mm3]	12478 ± 12518	154 ± 589
Cases without lesions [n]	0	75

In light of these variations, we adopted a stratified division method for forming the train-test division and the five-fold cross-validation splits. Additionally, another fold was reserved for testing. This stratification was crafted to consistently incorporate a certain number of both baseline and new or evolving lesions, see Table 2. This challenge is even greater considering the low number of new lesions in general, with just 2 lesions per case on average.

Table 2. Summary statistics of different splits.

Fold	Number of baseline lesions	Number of new lesions
0	58.6	2.2
1	59.4	1.2
2	62.1	1.4
3	83.4	3.2
4	49.6	1.9
Test	66.2	1.3

2.3. Implementation Details and System Configuration of nnU-Net

For our analysis, we specifically employed the 3D full-resolution configuration of nnU-Net in line with the three-dimensional attributes of our dataset. The entirety of the model training and evaluation was executed on a computing system equipped with a Linux operating system, 32GB of RAM, and powered by a 12GB NVIDIA GeForce RTX 4070 Ti GPU. Prior to using the nnU-Net architecture, intensity normalization was achieved in the FLAIR images using Z-normalization. Specifically, each image was individually normalized by deducting its mean and subsequently dividing it by its standard deviation.

2.4. Training Details

The networks are trained for 1000 epochs, where each epoch consists of 250 randomly selected mini-batches instead of iterating over the full dataset. The optimizer used is Stochastic Gradient Descent (SGD) with Nesterov momentum ($\mu = 0.99$) and an initial learning rate of 0.01. The learning rate decays following the polyLR learning policy, which decreases almost linearly to zero. The loss function combines cross-entropy and smooth Dice loss to handle class imbalance and improve model confidence. For validation, 50 randomly selected mini-batches are used to compute the Exponential Moving Average (EMA) of the Dice score. Data augmentation techniques such as rotations, scaling, Gaussian noise, and others are applied stochastically. To handle class imbalance, oversampling ensures each mini-batch includes at least one patch with a foreground class. This approach ensures robust detection of lesions, particularly baseline and new or evolving lesions, by choosing the class uniformly if multiple foreground classes are present.

2.5. Test

The final models were tested in the test split, first, and in the MSSEG-2 dataset and the Open MS Longitudinal Data dataset, comprising MRI data from 60 PwMS, second. For

the evaluation, we used the same metrics as in the MSSEG-2 challenge: (1) The number of wrongly detected lesion voxels in cases without lesions (2) The number of wrongly detected lesions in cases without lesions (3) The voxel-level Dice score in cases with lesions (4) The lesion-level F1 score in cases with lesions.

3. Results

3.1. Multiclass Lesion Detection with Early Stopping

We initially ran the pipeline using the default configuration of nnU-Net with a slight modification: the inclusion of early stopping. This was necessary for two main reasons. First, each epoch (training and validation) takes at least 225 seconds with our GPU, so training for 1000 epochs would require over two and a half days, making iterative improvements impractical. Second, the performance of baseline models drops significantly after fewer than 100 epochs, as shown in Figure 1, due to the models ceasing to predict new or evolving lesions, resulting in a Dice score of zero for this class. Using the models from the best epoch of each fold as an ensemble, the test results were decent, with a Dice score of 0.72 for baseline lesions and 0.47 for new or evolving lesions on the validation split. However, the high average number of false negatives (21 per case, compared to 32 true positives and 11 false positives) when detecting new or evolving lesions was considered too high.

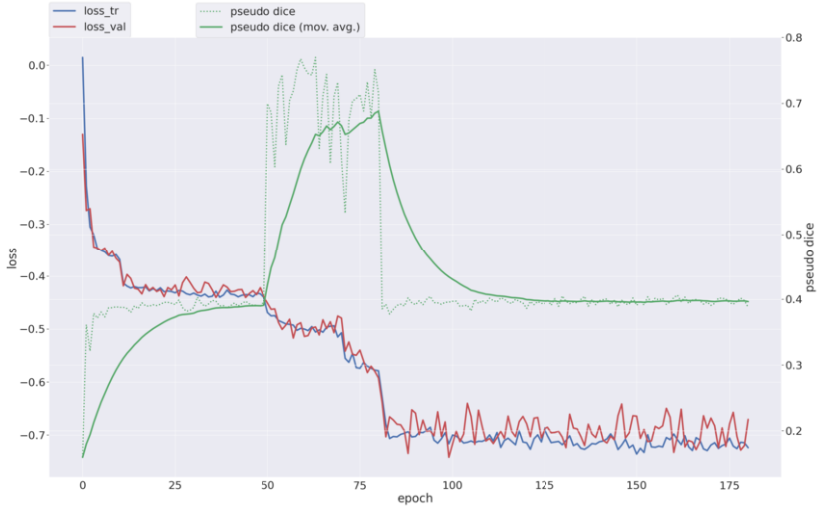

Figure 1. Training progress of fold 1 for the multiclass lesion detection model.

3.2. New or Evolving Lesion Detection with Extreme Oversampling and Lower Learning Rate

We focused on an extreme oversampling strategy to better handle the class imbalance between baseline and new or evolving lesions. Specifically, we ensured all training cases

had a baseline lesion class and always selected new or evolving lesions when present. This aimed to enhance the model's ability to learn the minority class while maintaining performance on baseline lesions. However, the model's performance was similar to the previous model (3.1) using the default configuration in nnU-Net. It was identified that this issue stemmed from a saddle point of the loss function used. The logical solution was to avoid this saddle point. After considering several options, we tested a simple one: decreasing the learning rate. Specifically, we reduced the learning rate to 0.001, which significantly and consistently improved the convergence issue, as exemplified by Figure 2. This, combined with the extreme oversampling strategy, considerably improved validation results (Dice score of 0.72 for baseline lesions and 0.55 for new or evolving lesions).

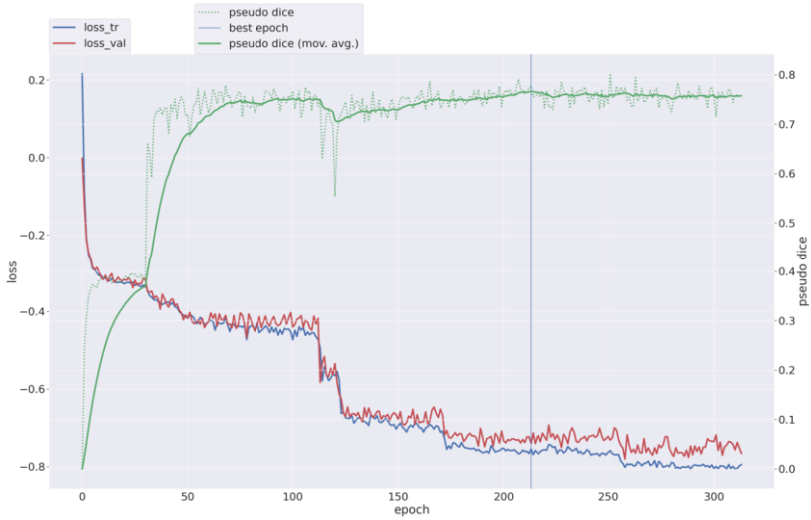

Figure 2. Training progress of fold 1 for the multiclass lesion detection model with the smaller learning rate.

3.3. Test Cohorts

Table 3 shows the evaluation metrics on the test split. As we can see, the results are very good for both baseline and new or evolving lesions detection, demonstrating a very high capacity to identify and delineate lesions while generating few false positives.

Table 3. Results on the test split.

	Cases without lesions		Cases with lesions	
	Lesion volume [mm3]	Lesion number [n]	DICE	F1-score
Baseline lesions	-	-	0.72	0.73
New or evolving lesions	0.48	0.04	0.64	0.75

Table 4 shows the evaluation metrics on the MSSEG-2 and Open MS Longitudinal Data datasets for new or evolving lesions (which are the only labeled lesions). As we can

see, our model generalizes very well to the MSSEG-2 dataset, ranking among the best in all four metrics[3], but not as well to the Open MS Longitudinal Data dataset.

Table 4. New or evolving lesion detection results on the external test datasets.

	Cases without lesions		Cases with lesions	
	Lesion volume [mm3]	Lesion number [n]	DICE	F1-score
MSSEG-2	0.00 - **1st**	0.00 - **1st**	0.46 - 6th	0.58 - **1st**
Open MS Longitudinal Data	-	-	0.27	0.24

Figure 3 shows examples of correct predictions and errors made by the model when predicting baseline lesions. The model rarely misses a lesion, and when it does, it is usually in the most difficult cases—where the lesion is not clearly visible or lies very close to the border of the brain, making it difficult to distinguish the border from lesions. Regarding segmentation, as expected, the model is generally able to detect and correctly label the central voxels of lesions, but as it moves toward the borders, accuracy decreases and mistakes (both false positives and false negatives) occur.

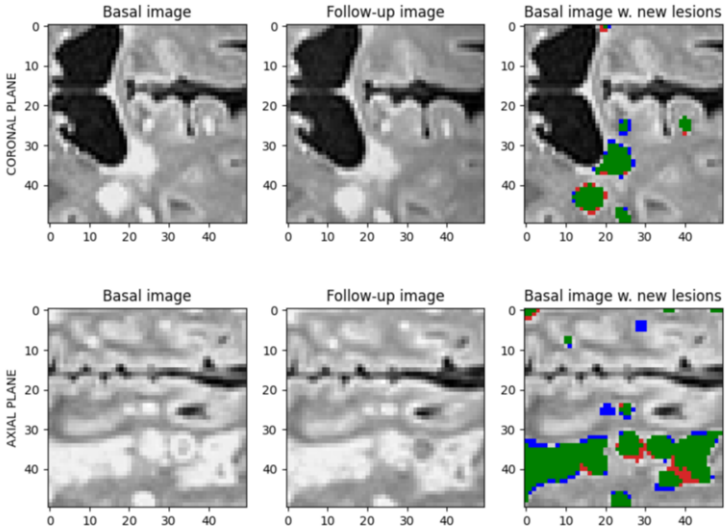

Figure 3. Example of baseline lesion detection where green indicates correct predictions, blue indicates false negatives, and red indicates false positives.

When predicting new or evolving lesions, clear lesions are usually detected, but the model struggles with diminishing, unclear, or border cases. As shown in Figure 4 the model detects almost all new lesions, but the segmentation is not as accurate. In this specific example, the model does not delineate the borders well, which reduces the Dice score, as shown in Table 3.

[3]We use the challenge's train set as test set, since its test set is not publicly available.

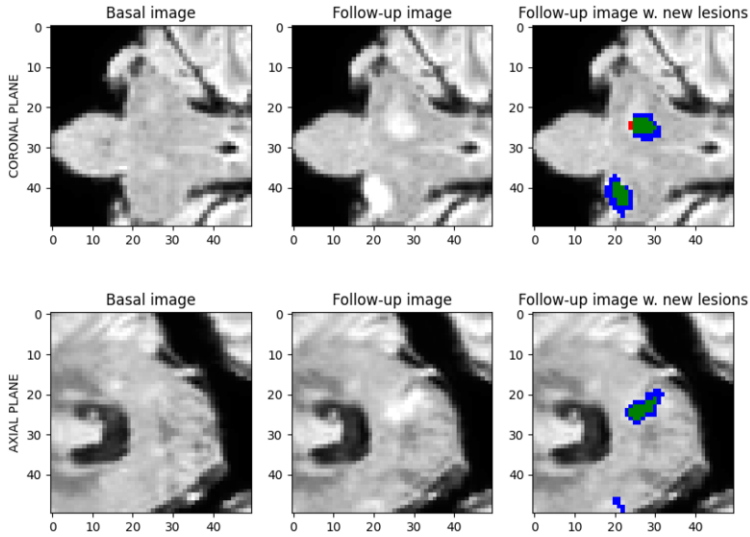

Figure 4. New or evolving lesion detection on the MSSEG-2.

3.4. Deployment

The final phase is the deployment process, which involves packaging our top-performing ensemble of models into a Docker container. To leverage the model's predictions on a dataset, users simply need to rename the images in a specific pattern following the nnU-Net format, ensuring the model can distinguish between baseline and follow-up images. Then, a single terminal command initiates the process. The generated model can be accessed upon a justified request.

4. Discussion

In this study, we aimed to generalize the previous work on detecting new or evolving Multiple Sclerosis lesions on longitudinal MRI images [14] to the multiclass detection of both baseline and new or evolving lesions. We relied on the nnU-Net segmentation pipeline as a baseline, incorporating preprocessing, data augmentation, and oversampling strategies. By fine-tuning the oversampling strategy and training process, we achieved excellent results in detecting both lesion types. The new or evolving lesion detection generalized well to different datasets, particularly the MSSEG-2 dataset, where our model ranked among the best in all four metrics. However, performance dropped when tested on the MS Longitudinal Data dataset, likely due to its lower image resolution. The performance of the model is quite sensitive to preprocessing steps such as skull stripping, alignment, and intensity correction. Any inaccuracies or inconsistencies in these steps, along with the presence of anisotropic resolution (where the slice thickness is larger than the in-plane resolution), can impact the model's ability to accurately detect lesion types, suggesting an area for further im-

provement. To improve the research, future work should focus on expanding the dataset to a broader population of MS patients. This expansion would help ensure that the model's performance is robust and generalizable across diverse patient demographics and imaging conditions. Additionally, incorporating super-resolution imaging techniques into clinical image settings could significantly enhance detection accuracy by improving image quality and resolution. Finally, we containerized our pipeline and model using Docker, providing a user-friendly way to generate baseline and new or evolving lesion masks. Comprehensive instructions can be found in the GitHub repository: https://github.com/ander-elkoroaristizabal/nnunet-ms-segmentation.

References

[1] Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T, Comi G, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. The Lancet Neurology. 2018;17(2):162-73.

[2] Diaz-Hurtado M, Martínez-Heras E, Solana E, Casas-Roma J, Llufriu S, Kanber B, et al. Recent advances in the longitudinal segmentation of multiple sclerosis lesions on magnetic resonance imaging: a review. Neuroradiology. 2022;64(11):2103-17.

[3] Lesjak Ž, Pernuš F, Likar B, Špiclin Ž. Validation of white-matter lesion change detection methods on a novel publicly available MRI image database. https://github.com/muschellij2/open_ms_data. Neuroinformatics. 2016;14(4):403-20.

[4] Commowick O, Cervenansky F, Cotton F, Dojat M. MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure. In: MICCAI 2021-24th International Conference on Medical Image Computing and Computer Assisted Intervention. http://portal.fli-iam.irisa.fr/msseg-2; 2021. p. 126.

[5] Salem M, Valverde S, Cabezas M, Pareto D, Oliver A, Salvi J, et al. A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. NeuroImage: Clinical. 2020;25:102149. Available from: https://www.sciencedirect.com/science/article/pii/S2213158219304954.

[6] Schmidt-Mengin M, Soulier T, Hamzaoui M, Yazdan-Panah A, Bodini B, Ayache N, et al. Online hard example mining vs. fixed oversampling strategy for segmentation of new multiple sclerosis lesions from longitudinal FLAIR MRI. Frontiers in Neuroscience. 2022;16. Available from: https://www.frontiersin.org/articles/10.3389/fnins.2022.1004050.

[7] Ashtari P, Barile B, Van Huffel S, Sappey-Marinier D. Longitudinal multiple sclerosis lesion segmentation using pre-activation U-Net. In: MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure; 2021. p. 45.

[8] Basaran BD, Matthews PM, Bai W. New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation. Frontiers in Neuroscience. 2022;16. Available from: https://www.frontiersin.org/articles/10.3389/fnins.2022.1007453.

[9] Hitziger S, Ling WX, Fritz T, D'Albis T, Lemke A, Grilo J. Triplanar U-Net with lesion-wise voting for the segmentation of new lesions on longitudinal MRI studies. Frontiers in Neuroscience. 2022;16. Available from: https://www.frontiersin.org/articles/10.3389/fnins.2022.964250.

[10] Sarica B, Seker DZ. New MS Lesion Segmentation using Deep Residual Attention Gate U-Net using 2D slices of 3D MR Images. In: MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure; 2021. p. 25.

[11] Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods. 2021;18(2):203-11.

[12] Isensee F, Schell M, Pflueger I, Brugnara G, Bonekamp D, Neuberger U, et al. Automated brain extraction of multisequence MRI using artificial neural networks. Human brain mapping. 2019;40(17):4952-64.

[13] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. IEEE transactions on medical imaging. 2010;29(6):1310-20.

[14] Martínez-Heras E, Vicente-Gomez A, Vivó F, Diaz-Hurtado M, Kanber B, Casas-Roma J, et al. Longitudinal Segmentation of Multiple Sclerosis Lesions using nnU-Net architecture. In: CCIA 2023: Conference of the Catalan Association for Artificial Intelligence, Món Sant Benet, October 25-27th; 2023. p. 163-72.