Universitat Oberta de Catalunya (UOC)

Master's degree in data science

FINAL MASTER'S THESIS

Area: Medicine

Automated Identification of Initial and Progressing Multiple Sclerosis Indicators through the multiclass detection of Baseline and New Lesions

---

Author: Ander Elkoroaristizabal Peleteiro

Tutor: Eloy Martínez de las Heras

Professor: Ferran Prados Carrasco

Barcelona, June 21, 2023

Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	Automated Identification of Initial and Progressing Multiple Sclerosis Indicators through the multiclass detection of Baseline and New Lesions
Nombre del autor:	Ander Elkoroaristizabal Peleteiro
Nombre del colaborador/a docente:	Eloy Martínez de las Heras
Nombre del PRA:	Ferran Prados Carrasco
Fecha de entrega (mm/aaaa):	06/2023
Titulación o programa:	Máster Universitario en Ciencia de Datos
Área del Trabajo Final:	Medicina
Idioma del trabajo:	Inglés
Palabras clave:	*Multiple Sclerosis, MRI, Deep Learning*
Repositorio de código:	https://github.com/ander-elkoroaristizabal/nnunet-ms-segmentation

All models are wrong, but some are useful.

– George Box

Acknowledgments

Nire aita, ama eta arrebari, beste mila gauzaz gain, hona ekarri nauen bidean beti nire ondoan izateagatik.

Nire familia eta lagunei, ematen didazuen guztiagatik. A *miña familia* y amigos, por todo lo que me aportais.

A les meves companyes i companys de Basetis, per la seva comprensió, el seu suport i l'après aquests tres últims anys, que m'han ajudat a ser la persona i el professional que ara soc.

A mi tutor Eloy Martinez de las Heras por su disposición, ayuda y ánimos a lo largo del desarrollo de este trabajo.

Al Instituto de Investigaciones Biomédicas August Pi i Sunyer (IDIBAPS), al grupo de investigación ImaginEM y a las personas con Esclerosis Múltiple que con su labor y generosidad han hecho este trabajo posible.

A todas las personas con Esclerosis Múltiple y los y las sanitarias que les atienden, con la esperanza de que este trabajo sea un granito de arena mas.

Abstract

Multiple Sclerosis (MS) is an inflammatory and neurodegenerative disease that affects the Central Nervous System, and one of the most common causes of physical and cognitive disability in young adults. It is characterized by demyelination lesions that can frequently be seen on magnetic resonance imaging (MRI). The presence of these lesions is a biomarker of the disease and can be used to diagnose and monitor its progression. The manual quality assessment of MRI images for detecting new or evolving lesions in Multiple Sclerosis however, is tedious and prone to subjective bias, and takes a lot of time and effort. In contrast, automated lesion detection methods can provide a more objective and efficient alternative to this manual process. Nevertheless, the quality assessment performed by physicians remains a critical aspect of the standard diagnosis and monitoring of the disease.

The development of new machine learning tools - using both traditional and deep learning techniques - has been a huge advance on the automatic detection and classification of such lesions. Convolutional neural networks, in particular, have been widely used in the last years, with remarkable performance.

The purpose of this thesis is to develop a deep learning based method capable of segmenting longitudinal MS images - images from the same subject acquired at two or more different timepoints - into already present, non-active lesions and new or evolving lesions, as the resulting segmentation may be an effective biomarker of disease progression and treatment response. The main objective of this research is to develop an automatic lesion segmentation pipeline that can quickly and accurately detect the presence of new or evolving and pre-existing MS lesions. Additionally, a Docker container will be created to facilitate the usage of the developed method.

Keywords: Multiple Sclerosis, Magnetic Resonance Imaging (MRI), Deep Learning

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter we introduce the topic of this project and its motivation, its goals, the methodology we have used and our initial planification.

1.1 Description and motivation of the project

Multiple Sclerosis (MS) is an inflammatory and neurodegenerative chronic illness that affects the Central Nervous System (CNS), and one of the most important causes of both physical and cognitive disability in young adults. The causes of the disease are unknown, but various environmental and genetic factors seem to play a big role in the epidemiology and pathogeny of the illness. One of the most common characteristics of the MS is the presence of lesions in the CNS. Furthermore, one of the main components explaining the presence of cerebral atrophy is the damage produced by these lesions, which is also mildly associated to the degree of physical and cognitive disability.

People with MS (PwMS) present visible lesions in magnetic resonance imaging (MRI) in different localizations of the CNS. T2-w (T2-weighted) and FLAIR (Fluid-Attenuated Inversion Recovery) images are particularly useful for visualizing MS lesions due to their sensitivity in detecting areas of increased water content, such as demyelinating lesions. These MRI sequences provide high contrast between normal and pathological tissue. The FLAIR sequence is particularly helpful for detecting lesions located near the brain's ventricles, as it suppresses the signal from cerebrospinal fluid (CSF), allowing better visualization of periventricular lesions. Therefore, integrating T2-w or FLAIR images into the evaluation process can significantly improve lesion detection in MS. The detection of these lesions in MS is a usual part of the diagnosis and monitoring of the illness, and requires a lot of time and effort from the specialist. Moreover, this identification is more prone to subjective bias than an automated assessment.

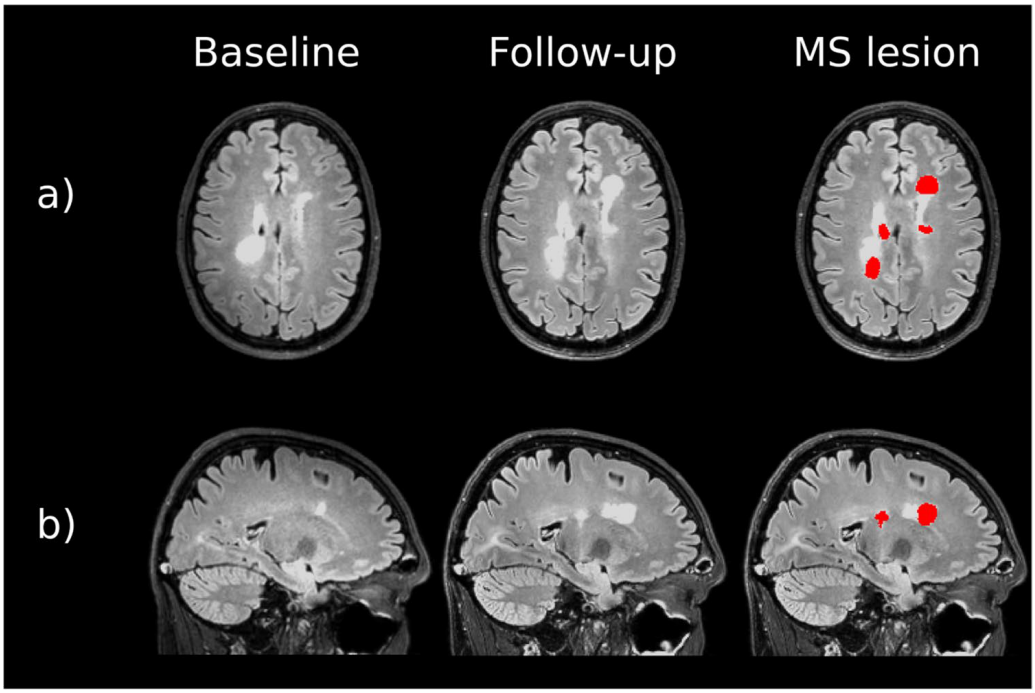

Figure 1.1: Example of new and enlarging lesions in MS, shown in red in the last column [1]. The first row (a) shows the axial view of the FLAIR MRI images, while the second row (b) shows the sagittal view.

The development of new machine learning techniques has been an important advance in the detection and classification of MS lesions. Convolutional neural networks (CNNs), specifically, are a very useful class of models that have already been successfully applied to the segmentation of longitudinal MRI images (images of the same subject acquired at two or more different time-points that characterize lesions over time, like the one shown in the figure 1.1). In 2020 CNNs became, as still are, the dominant technique [1]. Concretely, the U-Net architecture [13, 14] and different variations of it (such as nnU-Nets [15, 16] and 3D U-Nets [17, 18]) have been widely used. The relevance of the problem is such that three related challenges have been organized in the last decade: the ISBI 2015 longitudinal lesion segmentation challenge [19] and the MSSEG [20] and MSSEG-2 [21] challenges.

The objective of this thesis is to extend the previous work in the detection of MS new lesions in Longitudinal FLAIR images to the segmentation in multiple classes, in order to distinguish between healthy zones, lesions already present in the first image and lesions that are new or have evolved from the first image to the second. This will make the detection of new MS lesions and the growth or remission of existing lesions easier, faster and possibly more accurate, and therefore improve the monitoring of the disease progression and the treatment response.

Personally, this project is an opportunity to make a positive impact on the world and in the life of people with MS. Health is of paramount importance in life, and I embrace the opportunity to make an improvement in the medical treatment of people with MS, as small as it can be.

1.2 Goals

The aim of this project is to build on previous research and expand the segmentation method to classify images into multiple categories, allowing for differentiation between healthy areas, pre-existing lesions present in the first image, and new or evolving lesions that appear between the first and second images.

Because of its own nature this goal implies the existence of the following secondary goals:

- The study of the current State of the Art.
- The usage, if possible, of the already proposed architectures with a suitable output, set of hyperparameters and training procedure, among others.
- The analysis of the data in order to understand it and have an intuition of which pre-processing, data augmentation or post-processing techniques can be useful for our task.
- The determination of the most useful evaluation metrics for the purpose of our model.
- The development of a deployment architecture for making the resulting model available and useful, for example by containerizing our model into a ready-to-use Docker container.

In addition to these goals we have also tried to address the detection of slowly expanding lesions (SEL). This detection process is slightly more complex, since it is based on a criterion over the change of volume of lesions from the first time point to the second, but also specially interesting, as there are currently no automated methods for this task. Unfortunately, we have not been able to complete this additional task due to unexpected difficulties in the training of the models and time constraints.

1.3 Methodology

We have tackled this project using an iterative and agile approach, due to the complex nature of the problem we intended to solve. This approach has allowed us to refine previous steps whenever necessary, hence making it easier to improve both each of the steps and the performance of the entire solution.

The dataset we have used is composed by longitudinal FLAIR MRI images of 117 people treated at the Hospital Clínic de Barcelona (HCB), together with the target masks validated by

professionals from the ImaginEM research team from the HCB. The other resources we have used are the bibliography, the Python programming language, libraries such as Scikit-Learn and PyTorch, Git repositories with open-source implementations, Git itself as a version control system, and computing power (both CPU and GPU), used to train the model and analyze the data and the results. In addition, we have given careful attention to the task of detecting new lesions in the MSSEG-2 challenge, as it is critical for developing a robust and reliable methodology.

The consideration of the ethical dimension of the project has been a transversal part of the methodology, and hence has been considered in all phases of the project. This means that we have acted honestly, ethically, responsibly and respectfully with respect to human rights and diversity. Since our project is of medical nature, this commitment is specially important, and has been thoroughly cared for in the whole project, e.g. by making sure that the data is anonymous and the data privacy is preserved.

1.4 Planification

In this section we list the main tasks we have done during the project and give a broad planification of when each task was scheduled and has been tackled. The Gantt diagram illustrating this planning is shown in the figure 1.2. The conclusion of each of these main tasks is a milestone in the development of the project. It is important to notice that as stated in the methodology, the approach has been iterative, rather than sequential, so the planification reflects more the dates in which each task needed to be completed rather than when it has been first completed. In the design and implementation task, for example, an analysis, design, implementation and evaluation round has been done within the first weeks, and with the obtained knowledge a new refinement iteration has been done in order to improve the overall results.

The main tasks are the following:

1. **Definition and planification.** March 1st - 12th.

   The objective of this first phase has been to clearly define the topic, scope, methodology and main objectives of the project, justify its interest and relevance, and give a temporal planning.

2. **Study of the state of the art.** March 13th - 26th.

   The objective of this phase has been to study the related published research and developed solutions using the paper [1] as main source.

3. **Design and implementation.** March 27th - May 28th.

   The objective of this phase has been to design, develop and implement the project. This includes the analysis of the data and model design, implementation, train and evaluation tasks, among others.

4. **Documentation.** May 29th - June 25th.

   This phase has covered the thorough documentation of the project, putting together all the information gathered and written in the previous phases in a more structured and easy to understand way, together with the results and conclusions of the project.

5. **Presentation and defense.** June 26th - July 12th.

   In the final phase the project will be presented and defended in public to an evaluation tribunal.

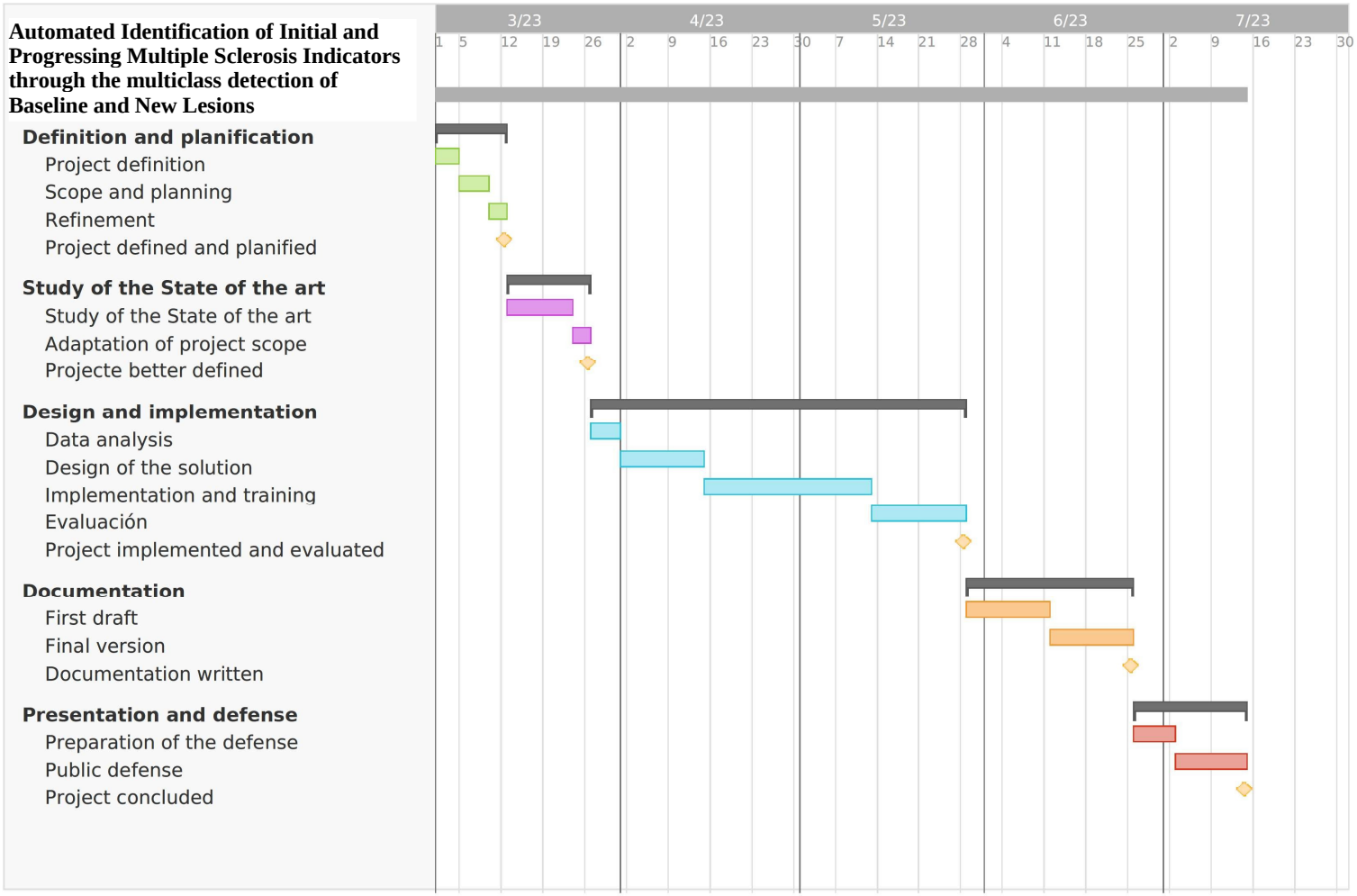

Figure 1.2: Gantt diagram of the planification.

Chapter 2

State of the Art

There is not much literature published about the task we are tackling, the detection of both basal and new or evolving MS lesions in longitudinal MRI images. Therefore, the most sensible thing to do is to study the State of the Art in the most similar well-researched task, which in our case is the detection of new MS lesions. Hence, in this chapter we revise the State of the Art in the task of detecting new MS lesions in longitudinal MRI images.

The main differences between our task and the one we study in this chapter are that the latter does not involve basal lesions, so it is not multi-class, nor evolving lesions, so the single class of lesions being detected is more specific: new lesions, exclusively. It is also important to mention that we will focus on the research not already included in [1], that reviews the work published between the previous review [22] and January 2022. Instead, we will focus on the research done after January 2022, and specially in that published within the MICCAI21 MS lesion segmentation challenge (MSSEG-2) [21].

We first shortly introduce - in section 2.1 - the history of MS lesion detection, explain the rising relevance of the new lesion detection task, and summarize the main challenges of MS lesion detection.

We then move on - in section 2.2 - to studying the methods published within the MSSEG-2 challenge and the results obtained by the best proposals.

For completeness, in section 2.3 we review the few peer-reviewed articles published outside the MSSEG-2 challenge after January 2022.

And finally, in section 2.4 we use the best methods published both within and outside the MSSEG-2 challenge to draw some conclusions about the design of our solution, such as which neural network architectures are commonly used, etc.

2.1 Introduction

Lesion segmentation in MRI images of PwMS is essential for the diagnosis and monitoring of the disease. At first cross-sectional images - images for a single timepoint - were used. This changed in 2017, when the latest version of the McDonald criterion [23], one of the main criteria used in the diagnosis of Multiple Sclerosis, was released. The update emphasized the relevance of studying the development of the disease in both the spatial and temporal dimensions, and therefore increased the importance of techniques based on longitudinal images [1]. Similarly, until 2020 the main techniques used in the segmentation of longitudinal images were mathematical and classical machine learning methods, but in the last years deep learning methods have got more and more pre-eminent [1]. This tendency is clearly observable in the methods submitted to the MSSEG-2 challenge, as we will see in the following section.

In is also important to notice that before the ISBI 2015 challenge [19] there was no public dataset that could be used as benchmark, so most of the previously published research was not comparable. The Open MS Data dataset [24] was made public in 2016, and last, the MSSEG-2 challenge has also made accessible its dataset. Nonetheless, these datasets are limited in size; the first two consist of a combined total of only 20 and 21 for training and testing, respectively, while the third dataset contains 100 studies. Furthermore, these datasets were labeled by only a handful of experts, leading to a potentially biased evaluation that heavily relies on the judgment of a small number of individuals in the field, whose opinions may diverge.

Both the task we are addressing and the one we are studying at the moment have multiple challenges in common. Let us take a moment to reflect on them before going into the details of the methods in the literature. The main challenges are the following:

- The scarcity of data: most medical datasets are much smaller than those that can be found on other sectors, and longitudinal MS datasets are no exception.
- The huge class-imbalance: in MS lesion datasets there are much more negative cases than positive ones, and this may affect the performance of the models.
- The need for the solutions to work on both ill and healthy people, i.e., to detect lesions when present but do not wrongly state there are in a healthy person.
- The small size and high variability of MS lesions. One of the clearest evidences is the high inter-labeler variability, visible for example in table 3.5.
- The variety of possible machines and images used in the clinical practice, and that therefore the model needs to be able to process with similar performance.

2.2 MSSEG-2 Challenge

The MSSEG-2 challenge is a Multiple Sclerosis new lesion segmentation challenge organized by OFSEP (the French observatory on MS) and FLI (the France Live Imaging research project) as part of the MICCAI 2021 conference. Its objective was to promote the development of methods for the automatic segmentation of MRI scans of MS patients in order to help clinicians in their daily practice.

The dataset of the MSSEG-2 challenge is composed by the MRI data of 100 MS patients. Each patient has a baseline 3D FLAIR scan and another 3D FLAIR follow-up scan obtained one to three years later. The dataset has been gathered from several centers using a total of 15 different MRI scanners.

Although the results of the challenge are already known (and are studied in section 2.2.2), all proposals presented to the MSSEG-2 challenge are studied in the Methods subsection, since all of them add some novelty and even proposals with (comparatively) bad results have ideas that are valuable for us.

2.2.1 Methods

From the 30 methods (from 27 teams) submitted to the MSSEG-2 challenge, 28 train some kind of CNN, one uses an already trained CNN, and another uses a classical algorithm, so the tendency to use Deep Learning methods is clearly visible in this challenge. Concretely, the U-net architecture and variations of it are widely used, while requiring a minimum volume for a lesion to be considered is a common postprocessing. The preprocessing, data augmentation and intermediate feature extraction steps are, as a matter of fact, where most methods differ.

Prados and Kanber [25] use a boundary-shift integral pipeline together with a Gradient Boosting Machine (LightGBM, concretely), and apply N4 bias correction, denoising and multiple bias correction as preprocessing steps.

Masson et al. [15] use a 2-channel nnU-net [11] for the segmentation task, whose result they post-process carefully choosing the threshold for a pixel to be classified as lesion (threshold tuning or moving) and requiring a minimum volume to the lesions, so that very small lesions are discarded. One channel is used for the basal image, while the other is used for the follow-up. Reorientation, skull stripping, registration, cropping, N4 bias correction, Nyul standardization and voxel intensity correction are applied as preprocessing steps. It is important to mention that the authors use another dataset of their own - compromised of 153 examples - together with an extensive data augmentation strategy to address the small data challenge. Concretely,

they use isotropic rescaling, 3D rotation, mirroring and intensity enhancements to augment the dataset.

Kang et al. [26] use a more convoluted approach. For the voxel segmentation they first use a 3D U-net for extracting the brain and spinal cord, then use another 3D U-net on concatenated patches from the two timepoints for segmentation, and third they use a 3D ResNet to reduce the False Positives of the previous segmentation. In parallel they use another 3D U-net for the segmentation of whole lesions using patches from each timepoint. Finally, they compare the lesions detected in each timepoint and combine the result with the voxel-wise segmentation to give the final segmentation. N4 bias correction and cropping are the two preprocessing steps they use.

Fenneteau et al. [2] use a siamese U-net like architecture formed by two MPU-nets [27] (figure 2.1) for segmenting each image and then merge the results with another refinement CNN block, which outputs the final segmentation. They apply co-registration, N4 bias correction, image resampling, histogram standardization and image normalization as preprocessing steps.

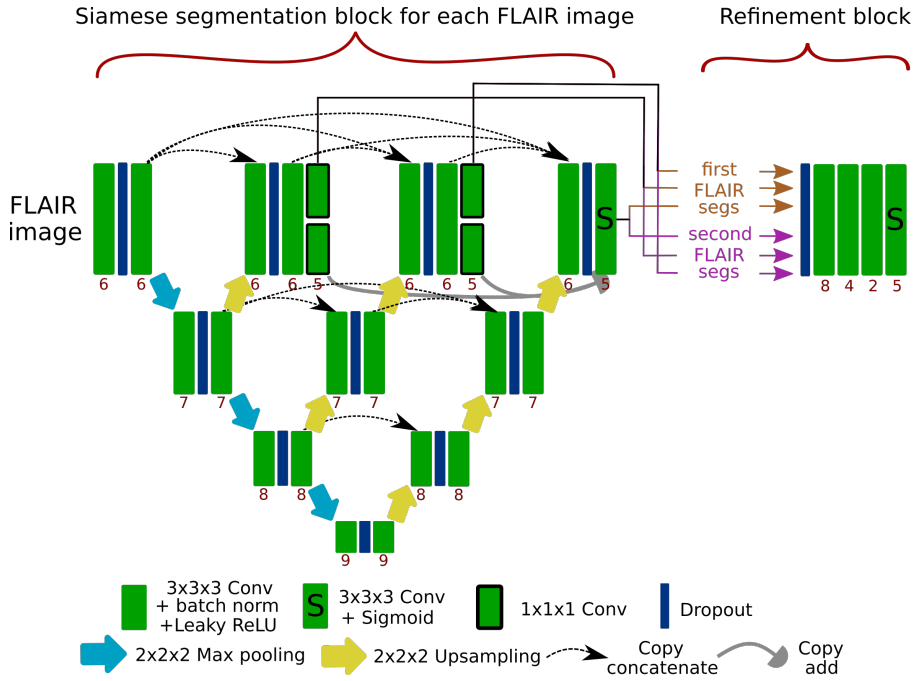

Figure 2.1: Architecture with the Siamese improved MPU-net in segmentation block and the convolutional refinement block used by Fenneteau et al. [2].

Gibicar et al. [28] use five 2D SC U-nets [29] together with an attention map from the subtraction of images for segmenting (2D) patches of the images, whose probabilities they then aggregate using the probability map proposed by Li et al. [30]. For the preprocessing they use the Anima script [31] proposed by the organizers of the challenge together with Gaussian inten-

sity normalization to obtain images with only the brain and spinal cord, bias corrected and with intensity normalized volumes. They use random affine transformations for data augmentation.

Löhr et al. [16] use a nnU-Net for segmenting all lesions in each timepoint (using a dataset of their own) and subtract the segmentations to get the new lesions. They then postprocess the result by running a connected components analysis and applying a threshold on the volume of the detected lesions. They train the network using their own dataset with 489 patients and fine-tune it with the challenge dataset. Furthermore, they apply random flipping, random rotation along all three axes and shearing for data augmentation. They preprocess the images by aligning them, applying skull stripping, normalizing the intensity distributions, standardizing their values and co-registering them. Finally, it is also worth mentioning that they combine the Dice loss - frequently used in the challenge - with the cross-entropy loss in the training because they claim that the former provides an edge when dealing with class-imbalance and the latter smoothens the training convergence.

Sarica and Seker [32] use an Attention Gate U-net [33] on the concatenation of the 2D slices from the axial, sagittal and coronal views of the images - which are stacked together - and then reconstruct the 3D segmentation by the majority vote of the views. Their motivation is to use the contextual information from all directions. Skull stripping and early-fusion are the two preprocessing steps they perform.

Kamraoui et al. presented three methods [34, 35, 36] to the challenge. The three proposals use the same preprocessing, which relies on the Anima script [31] for the correction of bias field, the denoising and the skull stripping of the images. They also use the same data augmentation technique, Image Quality Data Augmentation (IQDA) [37], which randomly applies blur, edge enhancement or axial subsampling distortion to training patches. However, although they use the same network for the segmentation task, a 3D U-net on the concatenation of the (3D) patches of the two timepoints, its training has a different twist in two of the proposals, [36] being the base. In [34] they additionally use a custom CNN for estimating the lesion-wise F1 loss, and this estimation is used (together with the voxel-wise Dice loss) in the training of the segmentation model. In [35], instead, they create new synthetic images by keeping or removing some of the new lesions and generating others from each real training example, using models for generating new lesions or simulating healthy tissue when removing a lesion.

Preloznik and Špiclin [3] use a double pathway custom 3D CNN combining both high and low resolutions for the image segmentation (figure 2.2). The preprocessing steps they apply are image re-orientation, Atlas registration, image normalization and image cropping.

Ashtari et al. [38] use a 3D U-net with Pre-activation [39] for the segmentation, which they train using intensive data augmentation and deep supervision [40] to bypass the small

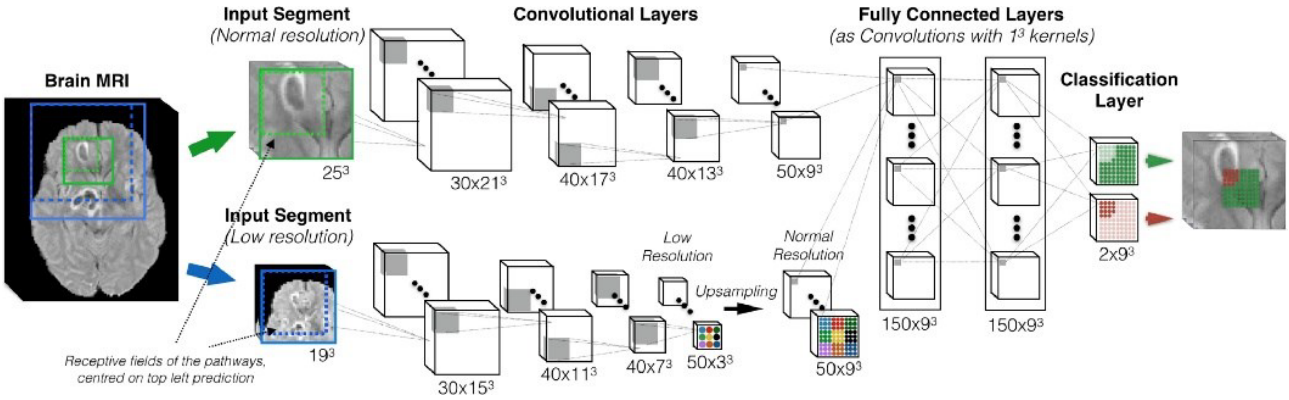

Figure 2.2: Double pathway CNN architecture used by Preloznik and Špiclin [3].

data problem. Concretely, they apply a combination of spatial transformations (random affine and random flips along all axes) and intensity transforms (random Gaussian noise, random Gaussian smoothing, random intensity scaling and shifting, random bias field and random contrast adjusting) using the MONAI [41] library. As preprocessing steps they concatenate both images forming a 2-channel 3D image, normalize the intensities of each image, and resample to the same voxel space. During training they also crop the images by filtering out the zero regions, and then extract random 128^3 patches from the images and oversample the patches with lesions in order to obtain a 50–50 distribution of patches with and without lesions. They compute the loss at each different resolution combining the Dice loss and the Focal loss and then take a weighted average. Finally, they also use weight decay and cosine annealing of the learning rate.

Efird et al. [42] use an ensemble of five cross-validation trained 3D U-nets with anti-aliasing steps that perform the segmentation on 2-channel patches of size 96^3. They use the Anima scripts [31] for bias correction, denoising and skull stripping, and combine it with a resampling and a cropping to the boundary box of the brain. At this point they apply several data augmentation strategies: random volume reorientation, a random elastic or affine transformation and intensity augmentations such as procedurally generated bias fields, modifying gamma by random power raising of intensities, random blur kernels and random high-frequency noise. Then they apply a last preprocessing step, which consists on clipping the image values below the 5^{th} percentile and above the 95^{th} percentile and then normalizing them to the $[-1, 1]$ range. During training, they extract random patches from the images and use early-stopping together with a custom loss function. This loss function is the combination of a distribution-based loss and a modified Dice loss that takes extreme values when the training example has no lesions.

Dalbis et al. [4] use a 2D U-net that takes one-by-one all different 2-channel slices from the coronal, axial and sagittal views and produces a 2D segmentation, which they then aggregate

using majority voting to obtain the 3D result (figure 2.3). To be precise, an ensemble of five cross-validated nets is used. The preprocessing steps are affine registration using ANTs [43], cropping the FoV (Field of View) to the brain, resampling to 256^3 voxels and pixel normalization. The data augmentation techniques they perform are contrast augmentation, rotations, flipping across the three orthogonal planes, elastic deformations and bias field augmentation on the 3D images. They use a combination of TopK cross-entropy loss [44] and Dice loss for training. In addition, they submitted two versions of the proposal to the challenge, one using just the given dataset, and another one using additional datasets with 46 cases.

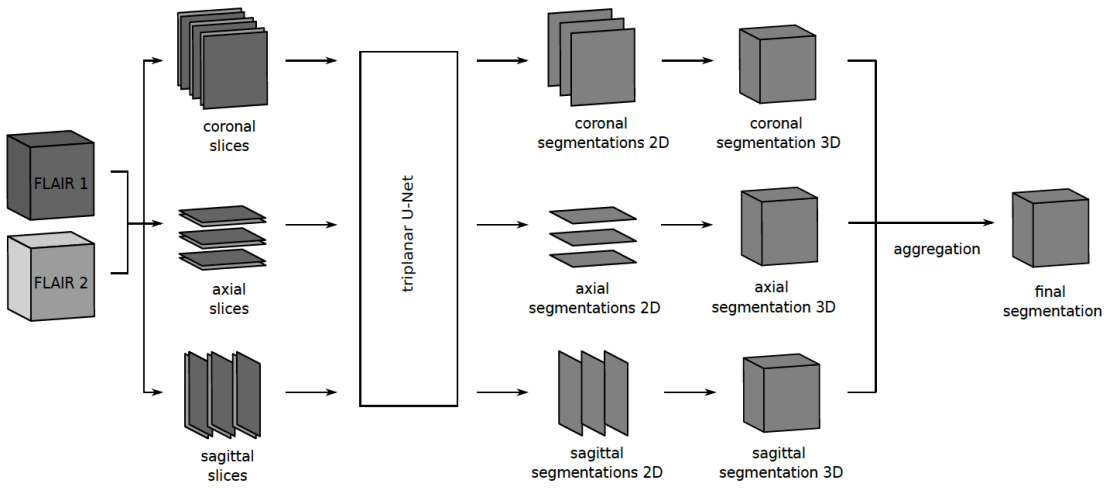

Figure 2.3: Segmentation process performed by Dalbis et al. [4].

Zhang et al. [45] use a slightly modified Tiramisu [46] encoder-decoder network that takes a concatenation of 2.5D stacked slices (one from each timepoint) as input and outputs the 3D segmentation. To be precise, they use an ensemble of five cross-validated nets. The authors claim that the use of 2.5D slices reduces the memory and computation requirements while maintaining both local and global information. The output layer of the network is Tanh activated, and they use the L2 loss for training the network. They use the Anima scripts [31] to perform the preprocessing. The data augmentation strategies they apply are image quality augmentations (blur, down-sample, sharpen and addition of Gaussian noise), image intensity distribution augmentations (Contrast Limited Adaptive Equalization (CLAHE), brightness and contrast adjusts and radial, horizontal and vertical gradient brightness application) and spatial augmentations (aspect ratio change, rotation, distortion, scaling, cropping, flipping and transposing).

La Rosa et al. [5] use several custom mathematical processing of the images together with an already trained longitudinal SAMSEG [47] model to perform the segmentation (figure 2.4). The two only preprocessing steps they perform are skull stripping and image normalization.

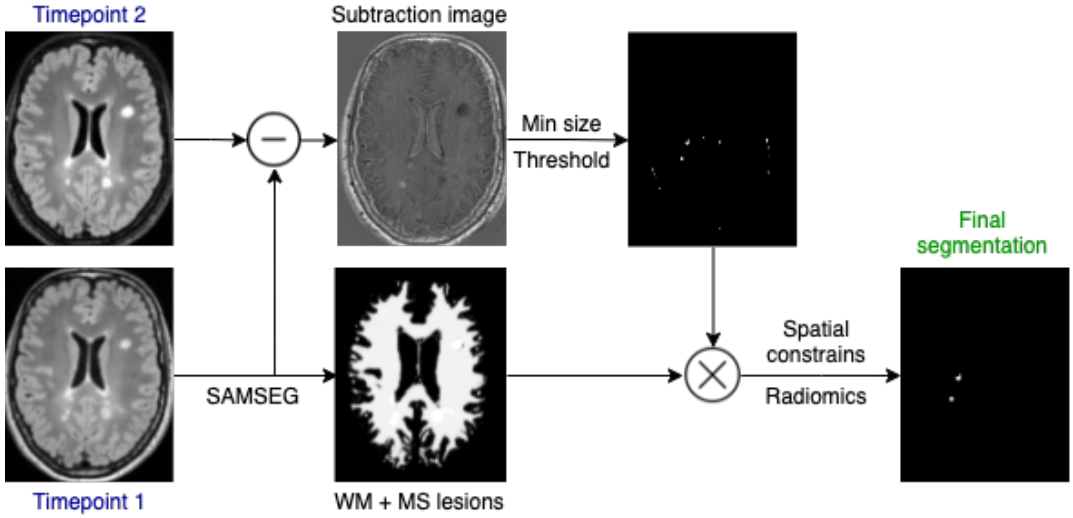

Figure 2.4: Scheme of the segmentation process performed by La Rosa et al. [5].

Macar et al. [48] presented three methods to the challenge, all three sharing the same preprocessing, data augmentation and base segmentation network architecture, the 3D U-net. The network from their first proposal follows the structure of the 3D U-net, but modifies its components, such as the activation layers and the batch normalization steps; the third uses an Attention Gate (AG) 3D U-net [33] together with Monte-Carlo dropout; and the second uses an ensemble of four nets, three modified 3D U-nets and one AG 3D U-net. The motivation for the modifications to the 3D U-net is to better adapt it for small batch sizes, the motivation for the attention gate is that it can help the network learn lesion structures better and more efficiently, and the motivation for using MC dropout is to reduce the variance of the output. Their preprocessing consists of several steps: they first resample the images, then extract the brain and spinal cord, perform a co-registration (using `sct_register_multimodal` from SCT [49] and `antsRegistration` from ANTs [43]), N4 bias correction using the corresponding Anima script [31] and finally crop the images around the volume of interest. The data augmentations they perform are random lateral flipping, random affine transformations, random elastic transformations and random MRI bias artifacts. They also oversample positive patches during training in order to get a more balanced dataset, use early-stopping and use soft Dice score as loss function.

Basaran et al. [14] use a cascade of two networks: the first 3D U-net detects all lesions in a single image - trained with ISBI 2015 and MICCAI 2016 data - and a second 3D U-net detects the new lesions from the subtraction of the segmentations and the follow-up image. The second network is trained using the MSSEG-2 dataset after removing the subjects with no lesions. Their preprocessing consists of a co-registration, a free form deformation using MIRTK [50], a brain extraction, a resampling and an intensity normalization.

Siddiquee and Myronenko [51] use an ensemble of five shallow and five deep versions of the SegResNet [52] network. The choice of the design is influenced by the small size of MS lesions. The only preprocessing step they use is the normalization of the images, and the data augmentation techniques they use are random flip on each axis, random contrast adjustment and random brightness.

Nichyporuk et al. [6] take advantage of the fact that the dataset includes the ground truth of all four raters and propose a "consensus" strategy: they use four nnU-Nets to mimic each experts' segmentation and then feed their predictions together with the original images to another nnU-net that brings them together to form the final "consensus" segmentation (figure 2.5). The motivation behind this approach is that although using only the consensus ground truth reduces the variability of the target, it also excludes the information that can be in the labels of the individual raters and left out of the consensus, and makes the model overconfident [53, 54]. In addition, they modify the nnU-nets in order to include lesion size reweighting [55] and feature-wise linear modulation (FiLM) [56] to try to improve the segmentation of small lesions and account for differences in image acquisition, respectively. The preprocessing of the images consists of N4 bias correction, Nyul normalization and symmetric diffeomorphic image registration [57].

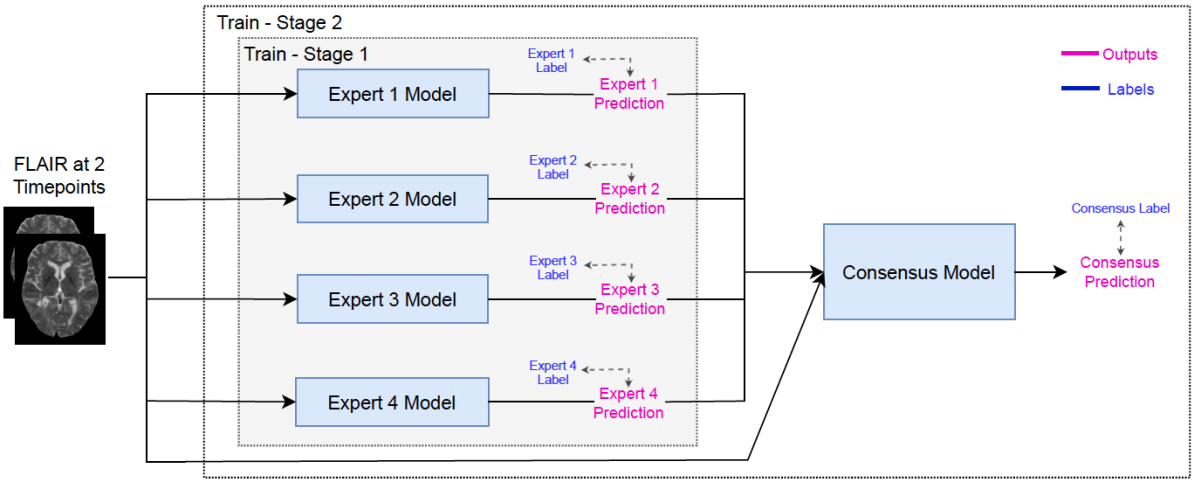

Figure 2.5: Consensus architecture used by Nichyporuk et al. [6].

Schmidt-Mengin et al. [18] use online hard example mining (OHEM) [58] for training a patch-wise 3D U-net in order to deal with the huge class-imbalance between positive and negative voxels. However, their own results show that OHEM is not effective in this setting, and they hypothesize that it may be due to the usage of the Dice loss. The only preprocessing steps are a resampling and a z-normalization of the images, which they then concatenate forming a double-channel image.

Hamzaoui et al. [7] pre-select the Regions of Interest (ROIs) using a sensitive classical image processing procedure - depicted together with the preprocessing in figure 2.6 - and then feed the interesting patches to a custom CNN to refine the segmentation. They augment the patches during training by applying random noise and random bias field.

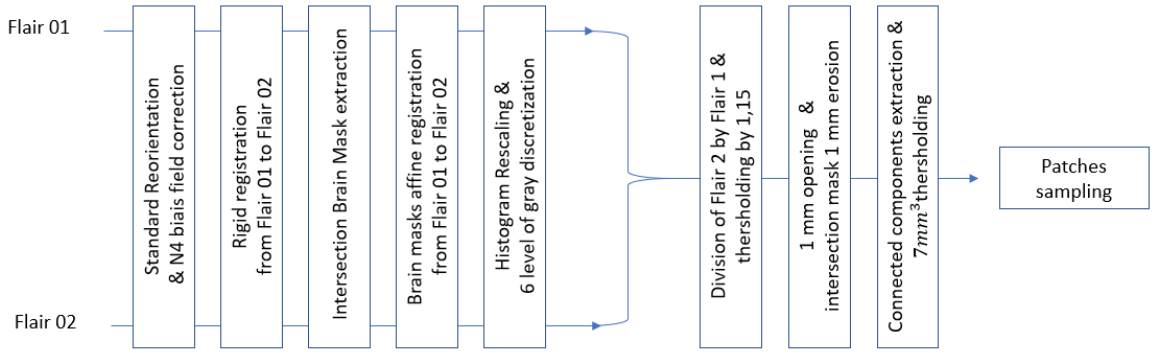

Figure 2.6: Preprocessing and image processing procedure used by Hamzaoui et al. [7].

McKinley et al. [59] use two networks in a cascade: they use a fine-tuned DeepScan [60] model to detect lesions in each timepoint, and then feed its output (together with the original images) to another similar network to find the new lesions. On inference their pipeline needs to be applied in each direction (i.e. six times) and the outputs aggregated. They preprocess by applying skull-stripping (using the HD-BET algorithm [61]), rigid co-registration and standardization. They use the focal loss and cosine annealing for training the second network.

Cabezas et al. [8] use a dual path U-net approach combined with self-supervised pre-training and attention gates. They use a dual path encoder (shown in figure 2.7) with shared weights between paths and attention gates which they first pre-train using a private dataset and then train with the challenge data. The preprocessing steps they use are Otsu thresholding [62], some morphological operations and the N4 bias correction algorithm [10] from the SimpleITK python library. They use five cross-validated networks as an ensemble for inference, each of them working on patches of the input images. They oversample patches with activity during training, and remove from the output lesions in the brain boundary or with a very small volume.

Andresen et al. [9] use an unsupervisedly trained NCR-Net [63] for co-registration and a 3D U-net for segmentation (figure 2.8). Due to memory constraints the only use three successive horizontal slices as input to the network. In order to cope with the class imbalance they oversample positive samples (that are in addition horizontally flipped when re-used) and undersample slices without lesions. As postprocessing they eliminate voxel segmentations outside the brain.

Salem et al. [64] use a double U-net architecture that combines in an end-to-end fashion intensity-based and deformation-based features. A first U-net block automatically learns the

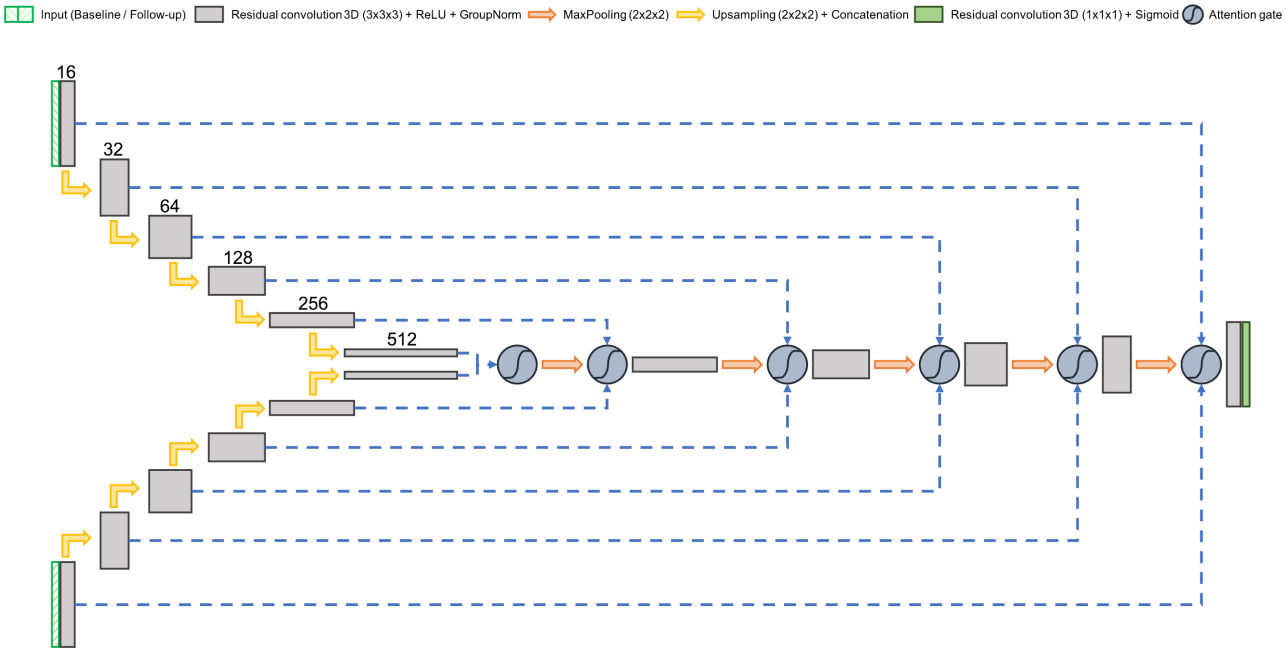

Figure 2.7: Scheme of the dual path U-net used by Cabezas et al. [8].

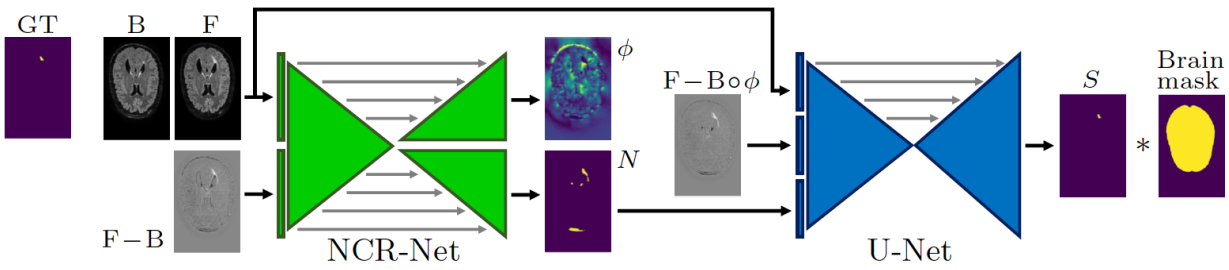

Figure 2.8: New lesion segmentation framework used by Andresen et al. [9].

deformation field corresponding to two patches and a second two-branch U-net [13] takes the original image patches and the deformation field and produces the segmentation. Since both networks are trained together the loss function they use is the summation of the unsupervised registration loss [65] and the cross-entropy loss. The preprocessing they use consists on a brain mask extraction with ROBEX [66], N4 bias correction from the ITK library and image normalization using histogram matching based on [67].

2.2.2 Results

Table 2.1 summarizes the results obtained by the four experts and the best methods presented to the MSSEG-2 challenge. We consider that a proposal is among the best if it is among the top nine in any metric (including experts) and is not in the six worst in any other. The organizers of the challenge decided that the evaluation of the proposed methods had to be done separately for cases with and without lesions, referenced as "With lesions" and "No lesions" in the table. For cases with lesions the segmentation-wise Dice (S-Dice) metric and the lesion-wise F1 (L-F1) metric - bigger is better - were used. The precise computation of the lesion-wise F1 metric is explained in [20]. Since these two metrics do not make sense when there are no lesions, the average number of (wrongly) detected lesions (Nlesions) and the average volume of the lesions (Vol-lesions) - smaller is better - were used as metrics in the no-lesions subset. All experts are among the nine best in all categories. For the submitted methods, those metrics among the 9 best are written in green, while metrics not in the 18 best are written in orange.

Table 2.1: Best MSSEG-2 results

Expert / Method	**No lesions**		**With lesions**	
	Nlesions	**Vol-lesions**	**S-Dice**	**L-F1**
Masson et al. [15]	0.286	4.258	0.432	**0.532**
Kamraoui et al. [36]	**1.143**	**38.486**	**0.500**	**0.517**
Ashtari et al. [38]	**0.036**	**0.470**	**0.409**	0.446
Dalbis et al. [4]-A	0.429	15.908	0.437	**0.525**
Dalbis et al. [4]-B	0.536	**29.235**	0.443	**0.541**
Zhang et al. [45]	0.536	12.713	**0.507**	0.500
Macar et al. [48]-2	**0.107**	**0.498**	**0.409**	**0.412**
Nichyporuk et al. [6]	**0.107**	**1.031**	0.423	0.453
McKinley et al. [59]	**0.071**	5.373	**0.403**	**0.431**
Cabezas et al. [8]	0.321	5.726	**0.485**	0.514
Expert 1	0.036	1.453	0.631	0.712
Expert 2	0.107	3.981	0.536	0.607
Expert 3	0.000	0.000	0.598	0.636
Expert 4	0.036	0.623	0.461	0.524

The immediate pattern we observe is also the one we could most easily suspect: solutions with good metrics in the subset with lesions do not perform as well in the subset without lesions, and vice versa. This is related to the specificity-sensitivity trade-off: methods that rarely predict a positive will get better results in the without lesions subset, but probably worse in the with lesions subset, and the other way around.

Taking this into account it is specially surprising to notice that only one of the proposed methods studied their results (or at least spoke about them in their submission) divided by cases with and without lesions.

It is also interesting to notice the huge variability between labelers, as shown by the low Dice and F1 scores obtained when compared to the consensus ground truth.

Furthermore, we see that, although some methods seem to be slightly better than others, there is no clear best method, and hence a more thorough discussion is necessary in order to consider possible approaches.

2.3 Methods published outside MSSEG-2

Only a handful of articles have been published in peer-reviewed journals after [1]. Most of them [68, 69, 70, 71] are closely related to the methods presented to the MSSEG-2 challenge, while others include some novelties. We now focus on the latter.

Basaran et al. [72] present a pipeline based on the nnU-Net architecture and lesion-aware data augmentation for new MS lesion segmentation. Concretely, they use axial subsampling (inspired by [36]) and the lesion-aware CarveMix [73] method for data augmentation together with other usual techniques. Their method outperforms all those presented to the challenge by achieving top results in all four evaluated metrics: an average Dice score of 0.510 and F1 score of 0.552 on cases with new lesions, and an average of 0.036 (false positive) lesions and volume of 0.192 on cases with no new lesions.

Valencia et al. [74] present a method for generating synthetic T1-weighted images from FLAIR ones, which they then use for improving the results of the method they presented [64] to the MSSEG-2 challenge. They achieve an F1 score of 0.582 on cases with new lesions while only wrongly detecting new lesions in 3 cases out of the 28 without new lesions.

Hashemi et al. [75] combine T2-w and FLAIR images with modified U-Net and Attention U-Nets to perform the segmentation. Since their work requires T2-w images they have not been able to benchmark their approach against the MSSEG-2 dataset.

2.4 Discussion

Taking into consideration the methods with best results (shown in table 2.1) together with the methods that did not perform as well and the methods presented outside the MSSEG-2 challenge we can see the following patterns:

- Preprocessing is very important: all high-performance methods use several preprocessing steps, most of them an extension to the steps included in the recommended Anima scripts. The usage of the Anima scripts [31] themselves is not as common as one could think considering that they were recommended by the organizers of the challenge, but is more logical taking into account that they are not as widely used in the neuroimaging field as other software tools such as FSL [76] and ANTs [43].
- Too much processing, on the contrary, does not seem to help: none of the MSSEG-2 pipelines relying more heavily on custom mathematical or computer-vision based techniques obtained a remarkable result.
- Data augmentation is also very important: 8 out of the 10 best methods in the MSSEG-2 challenge use intensive data augmentation strategies, and another one uses the data from the 4 experts. Task specific augmentations such as the ones used in [72, 74] seem to give relevant boosts in performance.
- On the other hand, having more data to use sometimes helps, but is no warranty: several methods use their own data at some point of their pipeline, but only two of them are among the top 10 methods.
- Surprisingly, class-imbalance does not seem to have such a big impact on the performance: only 5 of the 10 best methods in the MSSEG-2 challenge tackle it directly (and explicitly).
- U-Nets are the clearly dominating architecture, with a total of 8 out of the 10 best solutions: 3D U-Nets are used in 4 solutions, 2D U-Nets in 2 and nnU-Nets in another 2. nnU-Nets seem specially promising, since most methods that use them [15, 6, 72] get very good results.
- Using ensembles is usually a good option, specially when used together with cross-validation.

In addition, none of the submitted solutions explicitly speaks about taking into account a similar distribution of people with and without lesions (and the amount of lesions) when doing the train-validation-test split, which seems like a must for the task at hand.

It is also worth noting that none of the reviewed methods uses foundational models, which can at first be surprising. But the 3D nature of our images and the fact that we need to use images from two timepoints at the same time makes them much less useful in this context that in the general computer vision setting.

With the review we did of the literature - and that we have explained in this chapter - we got a better understanding of the current State of the Art in the MS new lesion detection task. We gathered a list of possible preprocessing tasks and data augmentation strategies, identified a promising architecture, and drew some conclusions from the analysis. With this new knowledge we were able to design our baseline semantic segmentation pipeline for the basal and new or evolving lesion segmentation task, which we explain in chapter 3.

Chapter 3

Implementation

In this chapter we explain how we have implemented our solution to the task of detecting both basal and new or evolving lesions in longitudinal Multiple Sclerosis MRI images. All the results shown in this chapter have been obtained using the code in the repository of the project, https://github.com/ander-elkoroaristizabal/nnunet-ms-segmentation.

Let us start by explaining our initial design. The essential task-specific preprocessing steps - co-registration, skull-stripping and bias inhomogeneity correction - have already been applied to the dataset, so we do not need to include them in our pipeline. As core and baseline we have decided to use the nnU-Net [11], that automatically sets up a complete segmentation pipeline, including a network architecture adapted to the dataset and preprocessing, oversampling and data augmentation techniques that work well in our setting, as shown by the MSSEG-2 results. Furthermore, we train five models through cross-validation and use them as an ensemble for testing and inference.

In the first section - section 3.1 - we detail and analyze the dataset we have used and explain how the extracted knowledge has influenced the choice of train-validation-test split strategy.

Then in section 3.2 we explain our baseline semantic segmentation pipeline, the nnU-Net: the design choices is based on, the implications of these choices for our case, its architecture and training process, etc.

In section 3.3 we explain the concrete experiments we have done and the motivation behind each experiment.

In section 3.4 we explain the results we have obtained using the models from the best experiment on the test split, the MSSEG-2 dataset and the Open MS Longitudinal Data dataset.

To conclude, in section 3.5 we explain how we have deployed the model using Docker in order to give an accessible and easy-to-use way of exploiting it.

3.1 Dataset

The dataset we have used is formed by 117 cases, each composed of two FLAIR images (the basal image and the follow-up) and the ground truth, all three in zipped NIFTI format (.nii.gz). The images were collected at the *Hospital Clìnic de Barcelona*, and their gold standard ground truth has been labeled by experts in the ImaginEM research group.

All images have been completely anonymized, meaning that identifying the person with MS a concrete image comes from is impossible. This anonymization process is essential for guaranteeing the data privacy and protection. Furthermore, our access to this data is granted by an agreement between the UOC (*Universitat Oberta de Catalunya*) and IDIBAPS (*Instituto de Investigaciones Biomédicas August Pi i Sunyer*) and limited to the scope of this project.

The NIFTI format was introduced by the Neuroimaging Informatics Technology Initiative to help store brain data obtained through magnetic resonance imaging techniques. The NIFTI format stores one three-dimensional volume of the brain scanned using MRI, and can be easily read in python using the NiBabel or SimpleITK libraries. Both the two images and the labels have the same shape, $182 \times 218 \times 182$, and each voxel represents one cubic millimeter (mm^3).

3.1.1 Dataset preprocessing

Some preprocessing steps have already been applied by the ImaginEM team to the dataset [12]. MRIs are taken under different acquisition protocols, and although efforts are made to replicate the same parameters of the initial (baseline) acquisition when obtaining follow-up images, it is necessary to standardize certain characteristics among the different images of the dataset involved. The preprocessing steps that they have used - orientation to MNI, skull stripping and intensity inhomogeneity correction - ensure that the model being trained is based on more homogeneous MRI characteristics across timepoints, resulting in a more reliable final model.

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, they proceed with the skull removal from the FLAIR images. This is carried out using the HD-BET algorithm [61]. 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, they correct the image intensities for inhomogeneities due to coil uniformities, field strength or other biases [77]. For bias inhomogeneity correction, they utilize the N4 algorithm [10] to achieve a more uniform intensity of the whole FLAIR images of the dataset.

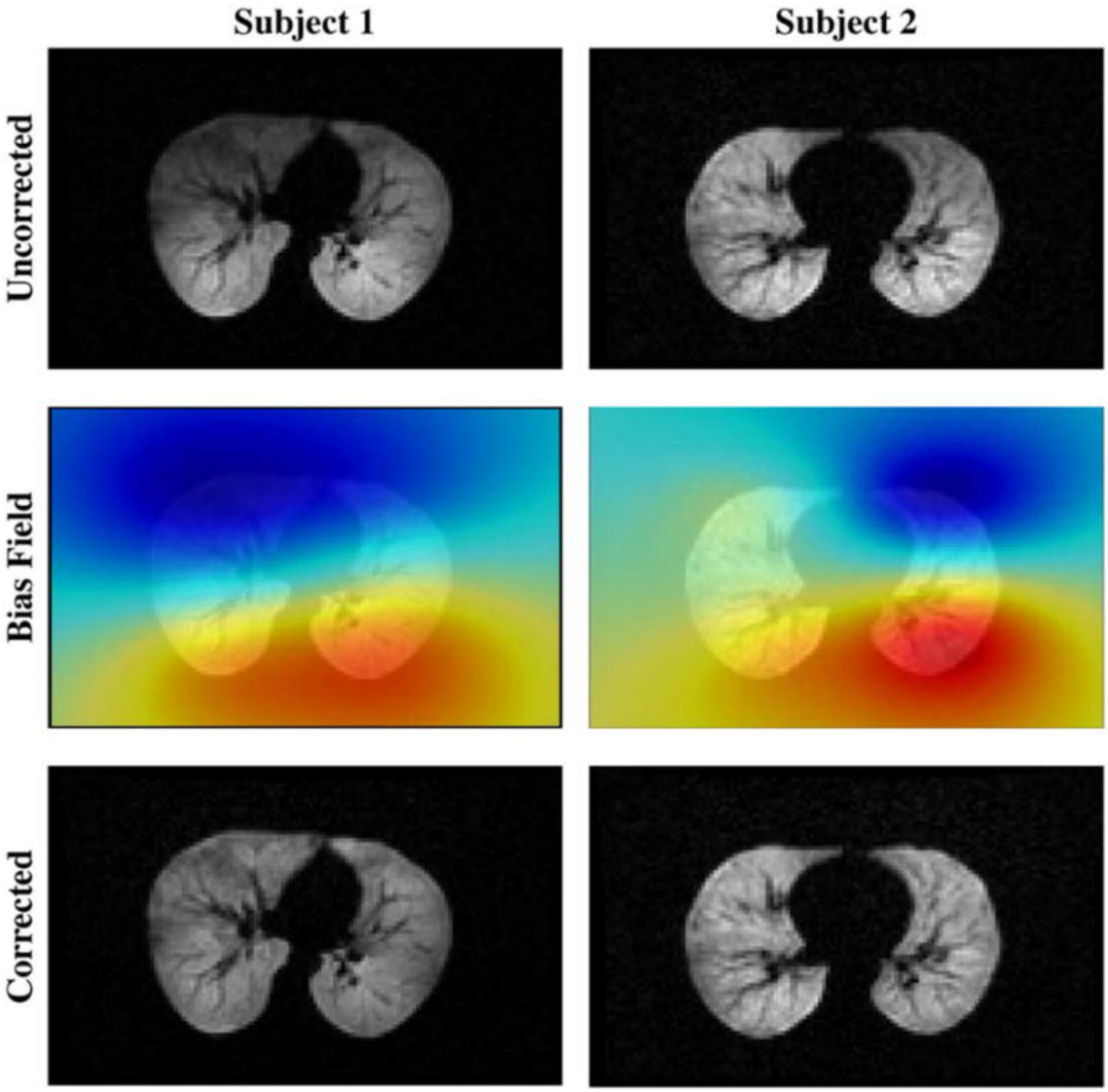

Figure 3.1: Illustration of N4 Bias Correction on lung MRI images [10]. Top row: Axial ^{3}He lung MRI from two subjects evidencing bias field artifacts. Middle row: Calculated bias field. Bottom row: Corrected images.

3.1.2 Exploration

One of the first steps when wanting to tackle a task, such as the detection of basal and new or evolving lesions, on a concrete dataset, is to study the dataset to be used. Since our dataset is composed by images, we have done so both quantitatively and qualitatively, that is, through a statistical analysis and a visual one. This section summarizes our findings.

3.1.2.1 Quantitative analysis

We first have studied the distribution of the lesions, divided by basal and new or evolving, and their size, throughout the dataset. Table 3.1 shows the mean and standard deviation of the number and volume of lesions of each type together with the number of cases without lesions.

Table 3.1: Summary statistics of the dataset

	Basal lesions	New or evolving lesions
Number of lesions	64 ± 41	2 ± 4
Total volume	$12478 \pm 12518\,\text{mm}^3$	$154 \pm 589\,\text{mm}^3$
Cases without lesions	0	75

The most noticeable thing is that 75 of the cases in our dataset do not have new or evolving lesions, i.e., almost 65% of cases. Although having cases without new or evolving lesions is usual, and in fact one of the challenges of the task we are addressing, having a too high percent of such cases increases the difficulty of detecting these lesions. This challenge is even bigger considering the low amount of new or evolving lesions in general, with just two lesions per case on average, and their mean total volume, of just $154\,\text{mm}^3$ or 0.002% of the voxels[1].

A related and also noticeable thing is precisely how the number and total volume of basal lesions is much larger than that of new or evolving lesions. All cases have basal lesions, with an average of 64 per case, and their volume sums up on average to $12478\,\text{mm}^3$ or 17% of the voxels.

In addition, the 42 cases with new or evolving lesions have from 1 to 29 lesions, with total volume ranging from $1\,\text{mm}^3$ to $5342\,\text{mm}^3$, so there is also a big variability within cases with new or evolving lesions. The variability on the number and volume of basal lesions is also big, since the number of these lesions varies from 5 to 252, and the total volume varies from just $170\,\text{mm}^3$ to $56941\,\text{mm}^3$ (80% of the mask!). This is specially relevant when dividing our dataset into train, validation and test sets.

3.1.2.2 Qualitative analysis

Since the detection of lesions is a strongly visual task, in this section we visualize some lesions to get an idea of how they look and how easy they are to spot and delimiter for the untrained eye. This way we can get an understanding of the difficulty of our particular task and dataset. We use the orange color to display basal lesions, and red to display new or evolving lesions.

[1] Recall that the shape of our images and labels is $182 \times 218 \times 182$.

Since the images are 3-dimensional, we always work with slices (not projections!) from the original images. One consequence is that the only voxel shared by the three slices is the central one, which can generate situations that may at first seem strange, such as a non-central lesion that only appears in one of the slices. The slices we use are in fact very concrete slices, since they are taken from the sagittal, coronal and axial planes of the brain, shown in Figure 3.2. Most images have a zoom applied, as many lesions are too small to be seen well when looking at the whole picture.

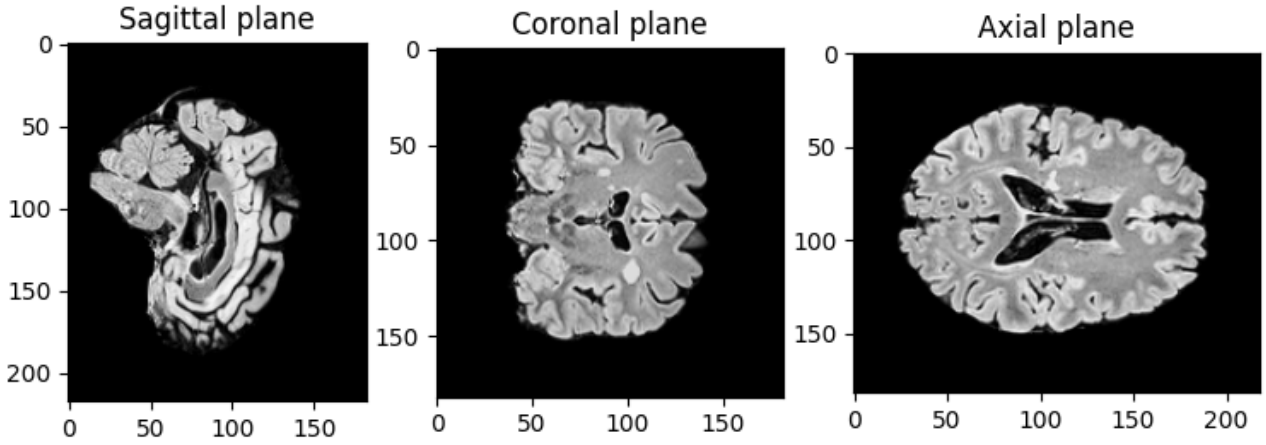

Figure 3.2: Sagittal, coronal and axial planes.

Figure 3.3 shows one basal lesion and its surroundings, both in the basal image, in the follow-up image, and with the basal lesions' segmentation superposed to the basal image. This is the easy to spot case: when the same white area appears in both the basal and the follow-up images. Notice, however, that in this case delimiting the lesion is not easy. It is also worth noticing that this pattern does not always indicate a basal lesion, as shown by Figure 3.4, where similar areas can be found in all three planes, but that are not, however, labeled as basal lesions. So in order to achieve good results a model needs to learn how to differentiate between these cases. And once again, the segmentation is even more difficult to see clearly, and has the particularity of being discontinuous in the coronal plane slice.

Let us now focus on new or evolving lesions. Figure 3.5 shows one new lesion and its surroundings, both in the basal image, in the follow-up image, and with the new or evolving lesions' segmentation superposed to the follow-up image. An increase in the brightness of an area from the basal image to the follow-up one usually denotes a new or enlarging lesion, as exemplified by the figure. But once again, this may not always be the case, as shown by the coronal slice in figure 3.6. And as can clearly be seen in both figures, segmentation of new or enlarging lesions is also difficult.

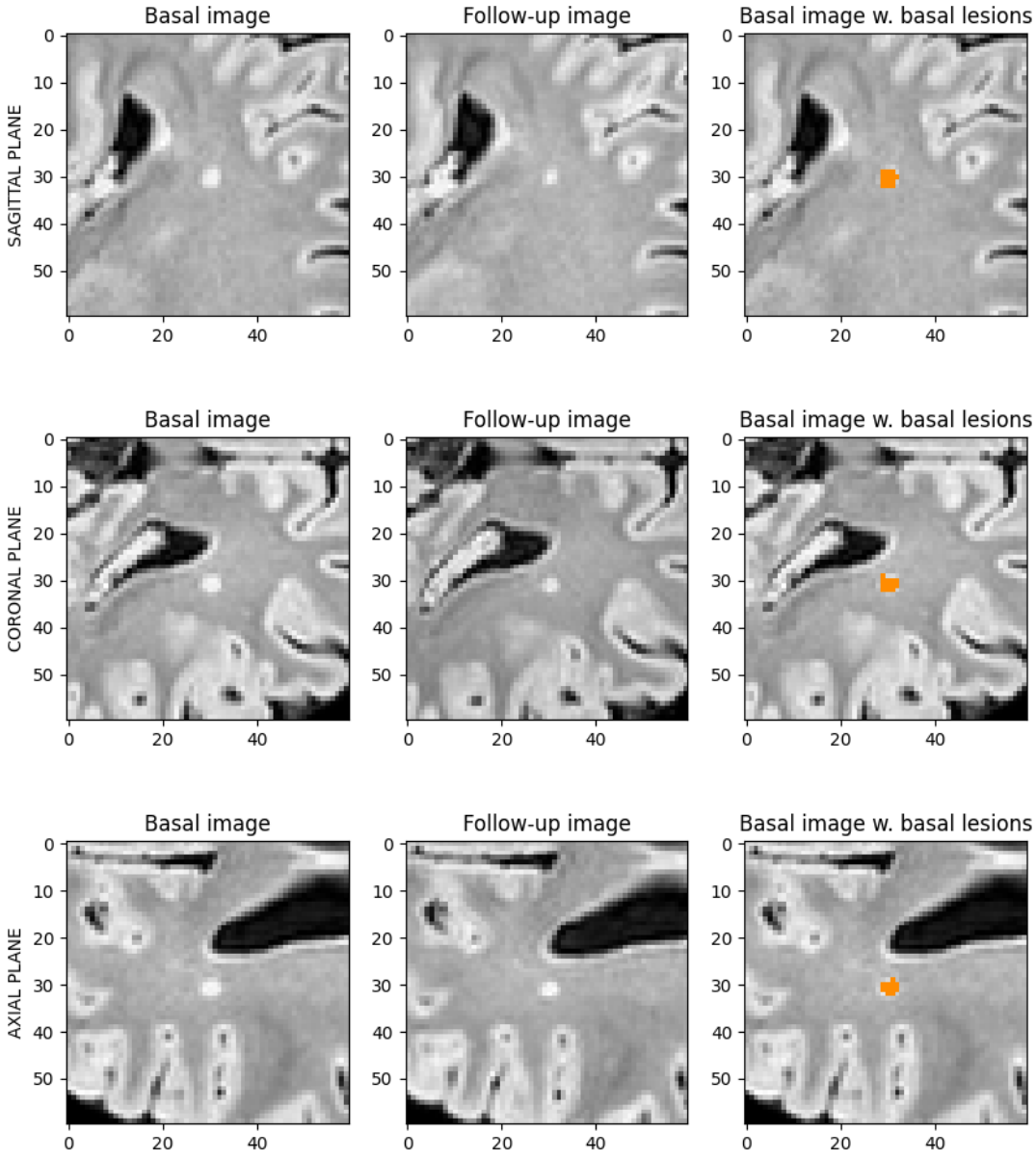

Figure 3.3: Example of an easy-to-spot basal lesion.

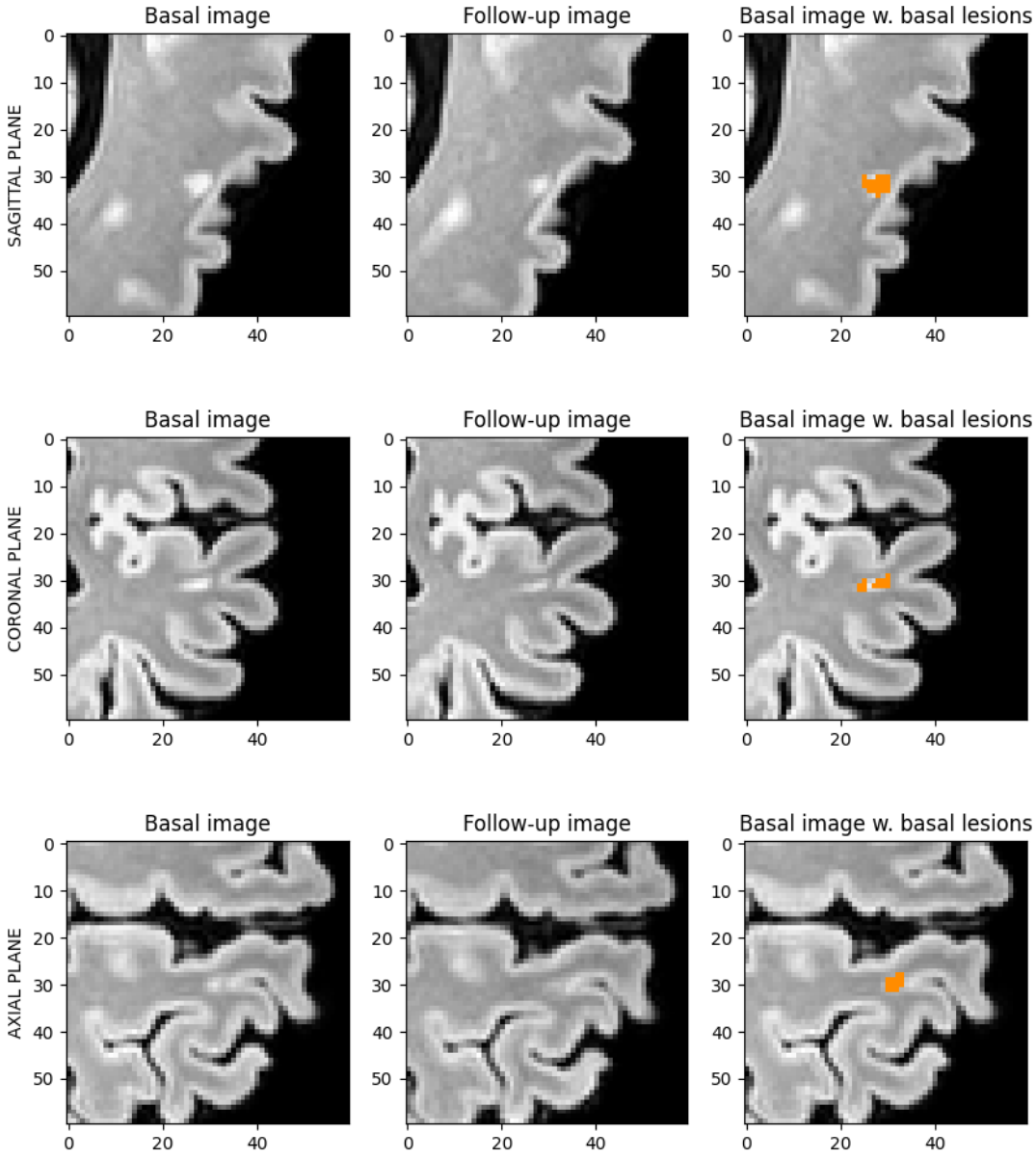

Figure 3.4: Example of anti-intuitive patterns in basal lesion detection.

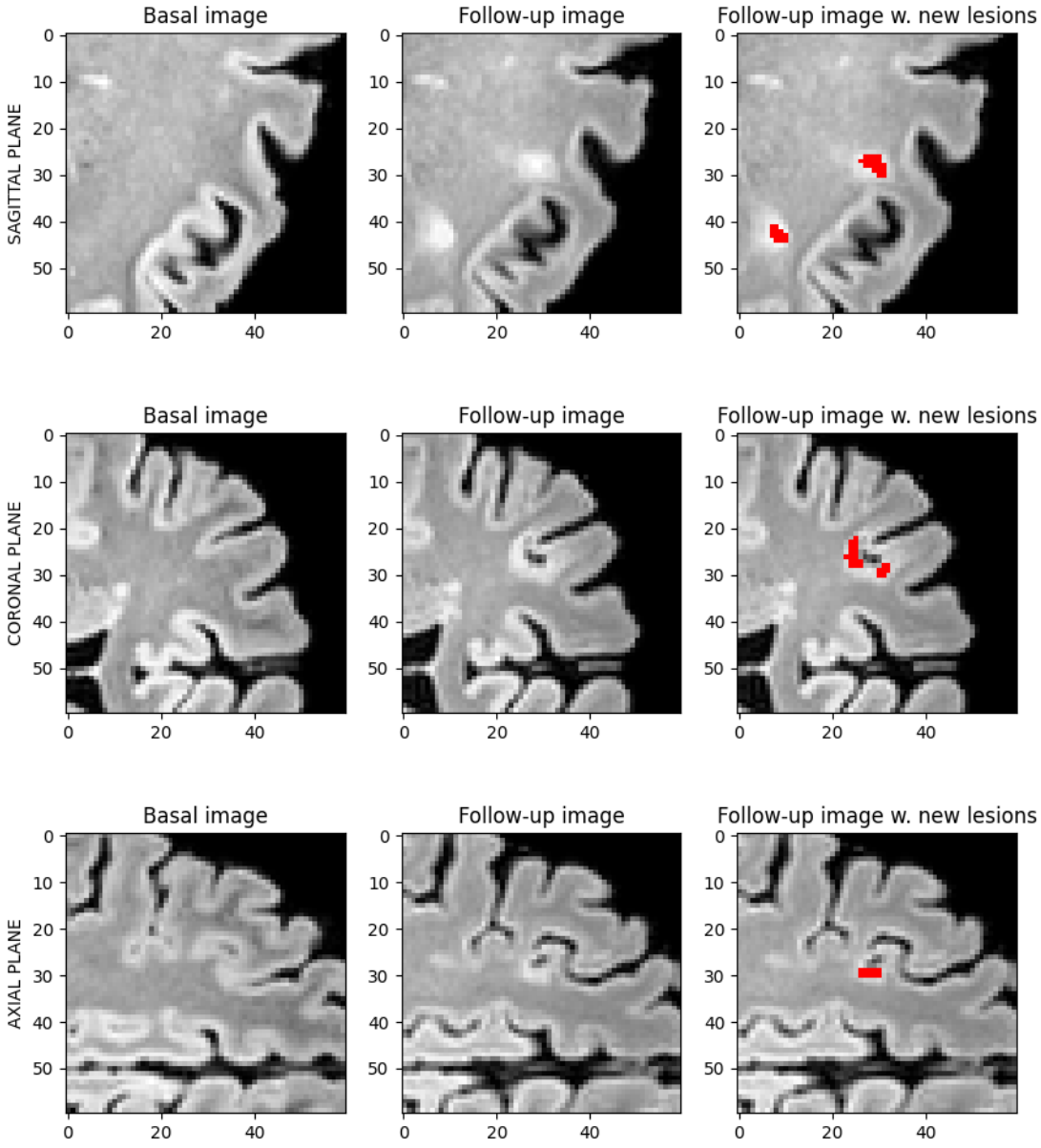

Figure 3.5: Example of an easy-to-spot new lesion.

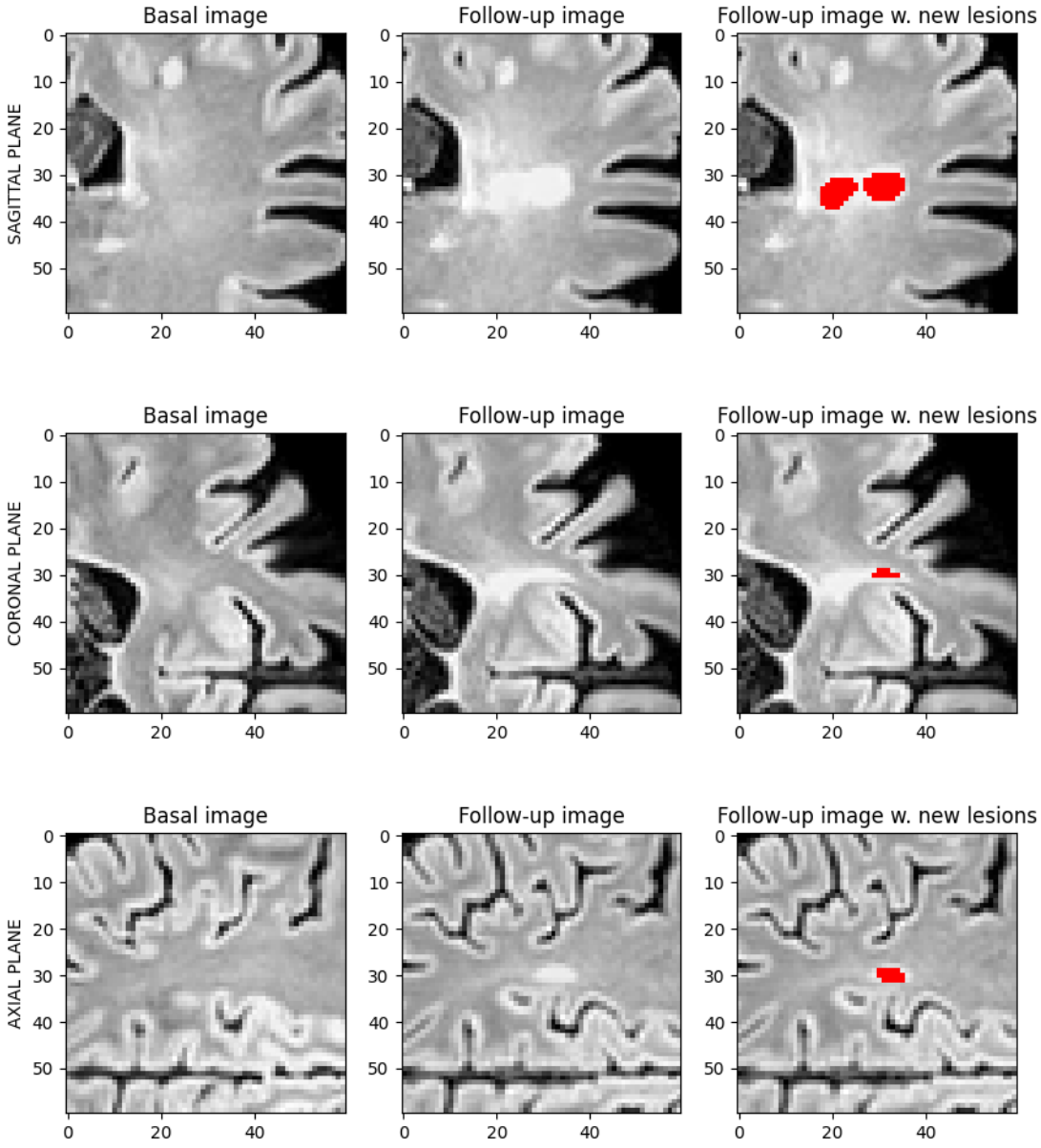

Figure 3.6: Example of an anti-intuitive pattern in new or evolving lesion detection.

Furthermore, our task also includes the detection of other - not enlarging - evolving lesions, such as diminishing or disappearing lesions. These lesions are usually denoted by a bright area that can be seen in the basal image and that decreases or disappears in the follow-up one. These lesions are visually the complete opposite of new or enlarging lesions, and in addition less common, so it is probably harder for a model to learn that they belong to the new or evolving lesions class. One direct consequence is that the class of new or evolving lesions - as a whole - is harder to detect than the class of basal lesions, since the variability is much bigger.

Finally, let us look at both lesion classes together. Figure 3.7 shows an example with both basal and new or evolving lesions marked. As we can see, when lesions are separated one can try to detect them and guess their type, but when lesions get closer the differences among them get less clear, and boundaries much harder to determine.

3.1.3 Train-validation-test splits

As we have seen in the previous sections, there is a lot of variability in both the amount of lesion voxels and number of lesions for both basal and new or evolving lesions. Due to these differences, a stratified split strategy has been used for generating both the train and test splits and the five-fold Cross Validation splits. Concretely, the splitting is done on the combination of the number of basal lesions and the number of new or evolving lesions. Figure 3.8 shows a scatter plot of the number of basal and new or evolving lesions in each case of the dataset. The stratification we have done roughly corresponds to dividing the plot into a 3×3 grid, and assigning all cases within each grid cell to a split randomly.

One can clearly see that even within each of these cells the differences can be very big, which combined with the few data points we have is the reason why even with stratified sampling the statistics of the resulting split are not as similar as one could hope for. Some average statistics of each of the training folds and the test set are displayed in table 3.2. Nevertheless, the sampling strategy used is good enough for our task, as we will show in the following sections.

Table 3.2: Summary statistics of different splits

Fold	Number of lesions	Number of basal lesions	N. of new or evolving lesions	Mean basal lesion vol.	Total basal lesion vol.	Mean new or evolving lesion vol.	Total new or evolving lesion vol.
0	60.8	58.6	2.2	170 mm^3	10005 mm^3	112 mm^3	110 mm^3
1	60.7	59.4	1.2	144 mm^3	8533 mm^3	34 mm^3	67 mm^3
2	63.5	62.1	1.4	241 mm^3	14953 mm^3	78 mm^3	412 mm^3
3	86.6	83.4	3.2	199 mm^3	16595 mm^3	45 mm^3	196 mm^3
4	51.4	49.6	1.9	151 mm^3	7476 mm^3	50 mm^3	215 mm^3
Test	67.5	66.2	1.3	222 mm^3	4694 mm^3	41 mm^3	54 mm^3

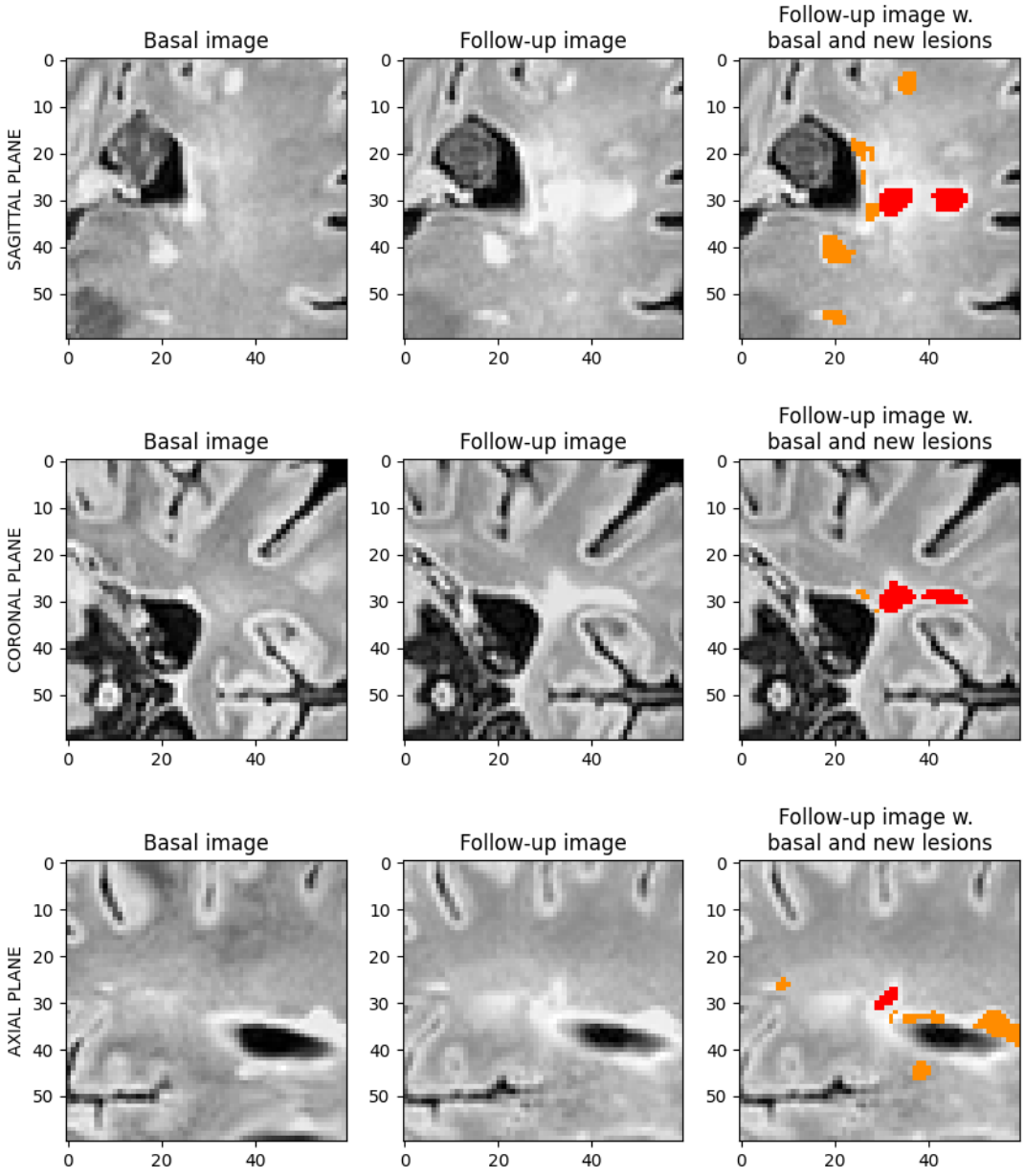

Figure 3.7: Example of basal (in orange) and new or evolving lesions (in red) together.

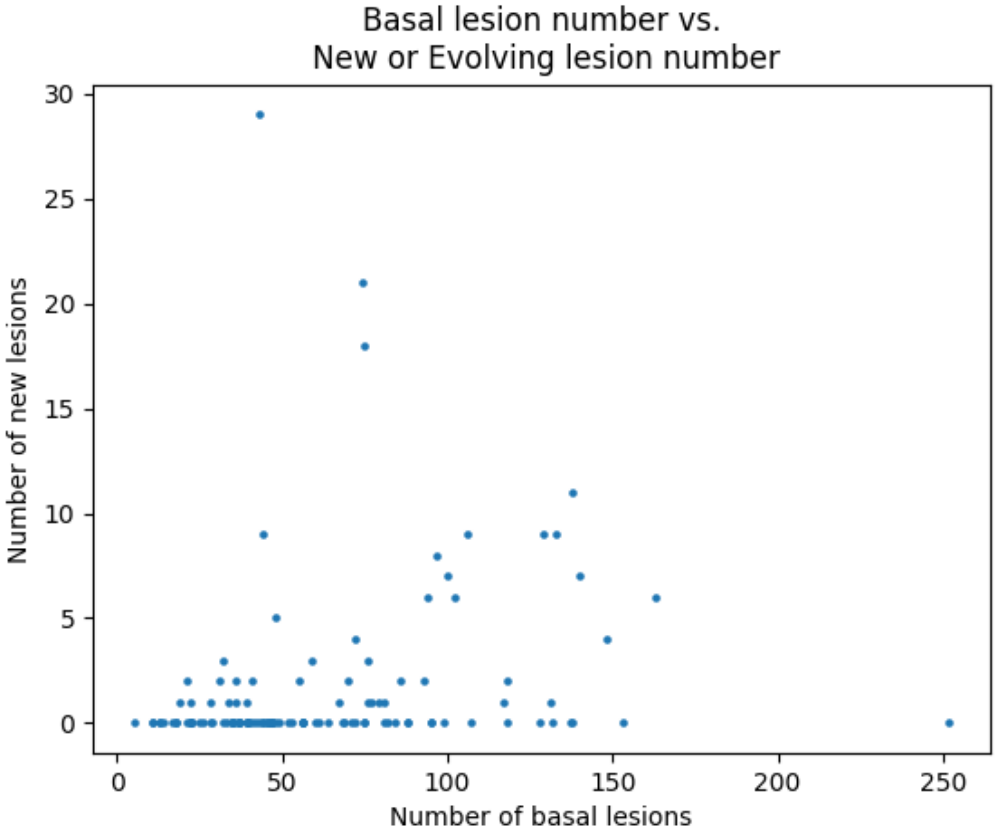

Figure 3.8: Scatter plot of the number of basal and new or evolving lesions of each case.

3.2 Segmentation pipeline: the nnU-Net

The "no new U-Net", or nnU-Net [11] is an out-of-the-box tool that automatically configures entire state-of-the-art semantic segmentation pipelines for arbitrary biomedical datasets without requiring expert knowledge or extensive computing resources to run. Figure 3.9 shows how nnU-Net systematically addresses the configuration of entire segmentation pipelines, providing a detailed visualization of the most relevant design choices. The automatic configuration of a great part of the pipeline drastically reduces the need for empirical design choices, hereby reducing the development time. Concretely, we have used version 2 of nnU-Net, which is a complete overhaul of the code of nnU-Net that makes it easier to modify the code and fine-tune the configuration to new datasets, further reducing development time.

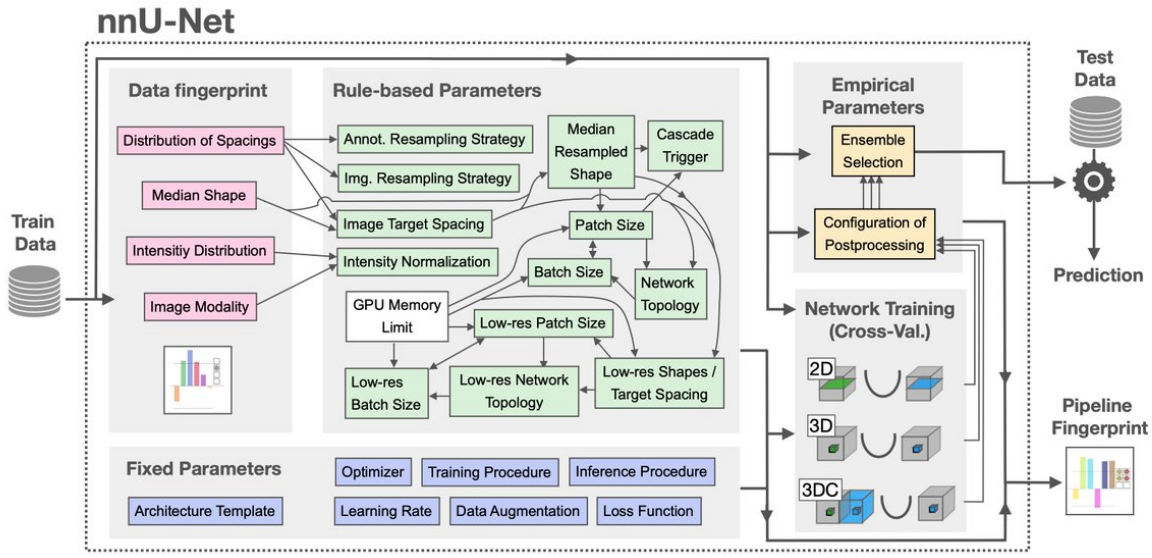

Figure 3.9: nnU-Net automatic method configuration [11].

Given a new dataset, nnU-Net analyzes the provided training cases, creates a "data fingerprint", and then creates several candidate U-Net configurations using a combination of the fixed and rule-based parameters shown in Figure 3.9:

- `2d`: a 2D U-Net. Can be used for 2D and 3D datasets.
- `3d_fullres`: a 3D U-Net that operates on a high image resolution (for 3D datasets only).
- `3d_cascade_fullres`: a cascade of 3D U-Nets where first a 3D U-Net operates on low resolution images (`3d_lowres`) and then a second high-resolution 3D U-Net refines the predictions of the former (for 3D datasets with too large[2] image sizes only).

[2]Relative to the memory of the GPU used.

Which exact configurations are created depends on the dataset. In datasets with "small" image sizes, the `3d_cascade_fullres` and `3d_lowres` configurations are omitted, since the patch size of the full resolution U-Net (`3d_fullres`) is big enough to cover a large part of the input images.

nnU-Net configures its segmentation pipelines using a three-step recipe:

- Fixed parameters, that do not change between datasets. During the development of nnU-Net the authors identified a robust configuration (i.e., certain architecture and training properties) that can be used almost always, and that therefore they fix. Examples include the loss function, most of the data augmentation strategy and the learning rate.

- Rule-based parameters, which use the data fingerprint to adjust concrete segmentation pipeline properties by following hard-coded heuristic rules. The network topology (pooling behavior and depth of the network architecture) is adapted to the patch size, the patch size, network topology and batch size are optimized jointly given some GPU memory constraint, etc.

- Empirical parameters, that are basically trial-and-error. The selection of the best U-net configuration for a given dataset (2D, 3D full resolution, 3D low resolution, 3D cascade) and the optimization of the postprocessing strategy, for example.

We now summarize the main characteristics of the nnU-Net when applied to our dataset. We also restrict ourselves to the `3d_fullres` configuration, since it is the one we have used, due to the 3D nature of our dataset and our sufficient GPU memory. To be precise, the training and evaluation of the models has been done using a computer with Linux operating system, 32GB of RAM and a 12Gb NVIDIA GeForce RTX 4070 Ti GPU.

3.2.1 Patch and batch sizes

The patch and batch sizes are specially relevant in 3D semantic segmentation due to the memory requirements of this kind of images and the impact on both the network topology and the training of the models. nnU-Net prioritizes that the batch size is at least 2, since this makes the training significantly more robust, and then maximizes the size of the patches, so that the model has as much contextual information as possible, which typically increases model performance. Since in our case the GPU used has a memory of 12Gb, the nnU-Net design choices output a batch size of 2 and patches of size $128 \times 128 \times 128$. Taking into account the size of the lesions we have seen in the qualitative analysis of the images this patch size seems enough for most lesions.

3.2.2 Architecture

The `3d_fullres` configuration of the nnU-Net uses a template that closely follows the original 3D U-net [78]. Due to the usually small batch size, batch normalization (often used to speed up or stabilize training) is replaced by instance normalization. Moreover, ReLUs are replaced with Leaky ReLUs (with negative slope of 0.01). In order to help the training of all layers in the network, deep-supervision is used: additional auxiliary losses are computed in the decoder and are included into the training loss with weight inversely-proportional to the depth they correspond to. The U-net typical configuration of two blocks per resolution step in both the encoder and decoder is used, with each block consisting of a convolution, followed by instance normalization and a leaky ReLU nonlinearity. Strided convolutions are used for downsampling and convolution transposed for upsampling. The initial number of features maps is set to 32 (considered a trade-off value between memory requirements and performance) and doubled (halved) with each downsampling (upsampling). Furthermore, the number of feature maps is capped at 320 to limit the model size.

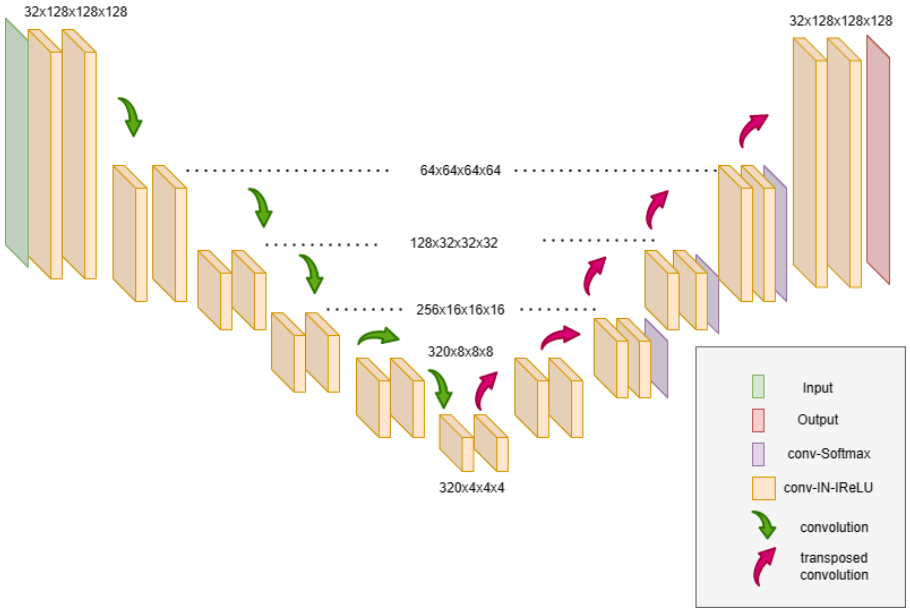

Figure 3.10: Architecture of the U-Net used [12].

3.2.3 Preprocessing

Z-normalization is applied as intensity normalization technique, i.e., each image is normalized independently by first subtracting its mean and then dividing by its standard deviation.

3.2.4 Training

Networks are trained for 1000 "epochs", with epochs defined as an iteration over randomly selected 250 mini-batches (as opposed to over the full dataset). SGD (Stochastic Gradient Descent) with Nesterov momentum ($\mu = 0.99$) and an initial learning rate of 0.01 is used as optimizer. The large momentum term is used to improve training stability with small batches. The learning rate is decayed using the 'polyLR' learning policy, which is an almost linear decrease to zero.

The sum of the cross-entropy and the (smooth) Dice loss is used as loss function. This combination of losses hence takes into account both whether the model is detecting the foreground cases with a robust to class-imbalance metric (via the Dice loss) and its security when predicting (via the cross-entropy loss) with a training-stabilizing metric. The cross-entropy loss is specially relevant when models are intended to use as an ensemble, since in this case the predicted probabilities are averaged across models before determining the final class.

Fifty randomly selected mini-batches are used for validation, and in addition to the validation loss, a validation metric is also computed: the Exponential Moving Average (EMA) of the Dice score.

Several data augmentation techniques are applied stochastically with a predefined probability: rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma correction and mirroring.

Finally, oversampling is applied to more robustly handle class-imbalances. For each of the mini-batches, one of the patches is forced to have a foreground class, while the other is chosen randomly. If a given patch is forced to have a foreground class, then the concrete class to be oversampled is chosen uniformly between the present foreground classes. Since in our case all images have basal lesions, this implies that if only basal lesions appear this class is chosen, and if both basal and new or evolving lesions appear, one of them is chosen with equal probability. Once the concrete foreground class has been chosen, one of its voxels is randomly chosen and used as center for the patch used for training.

3.2.5 Postprocessing

Motivated by organ image segmentation, the following connected component-based postprocessing strategy is considered and empirically tested: for each foreground class and their union (i.e. considering all classes one same class), all but the biggest lesion are removed. If any of these options improves the segmentation results on the validation set, it is then applied whenever inference is run. Since in our dataset we expect to have more than one lesion of each type this strategy is unlikely to help.

3.2.6 Inference

New cases are predicted using a sliding-window approach, with the patch size as window size and adjacent predictions overlapping by half the patch size. Since the accuracy of the segmentation tends to decrease when getting closer to the window border, a Gaussian importance weighting is applied, increasing the weight of the central voxels in the softmax aggregation. Furthermore, images are mirrored along all axes as a test time augmentation strategy.

3.3 Experimentation

nnU-Net has shown the ability to robustly find high quality configurations on new datasets, so it can serve as a very strong baseline, but task-specific empirical optimization is prone to improve the segmentation performance. Precisely, we have used nnU-Net as a basis for the pipeline and as baseline model, that with empirically tested dataset-specific training modifications we have been able to optimize. In this section we summarize the main experiments and modifications we have done to the baseline in order to improve our results. It is important to notice that the modifications we detail here have been applied one on top of the other, so its impact cannot be assigned to that single change, but to it together with the previous changes.

3.3.1 Baseline (with early stopping)

We first ran the pipeline with the default configuration, explained in section 3.2, with a small modification: the inclusion of early-stopping. Two were the main reasons for the inclusion of the early-stopping. First, that each epoch (both training and validation) takes at least 225 seconds with our last-generation GPU, so training for 1000 epochs would take more than two and a half days, hence needing two weeks to train all five folds uninterruptedly, and making it much harder to make iterative improvements. The second reason is that the performance of baseline models drops significantly after less than 100 epochs, as shown in Figure 3.11, so training much longer past this point makes no sense. The reason for this drop in results is that the baseline models suddenly start to never predict new or evolving lesions, so the Dice score for this class becomes zero. On the other hand, the Dice score obtained for basal lesions was good both for the best and the final models, which made us think that the drop could be related to the scarcity of new or evolving lesions in the dataset and thus in the models' training.

Using the models from the best epoch of each fold as an ensemble the test results were not bad, obtaining a Dice score of 0.72 for basal lesions and of 0.47 for new or evolving lesions on the test split. However, we considered that the average number of False negatives (21 per case, as compared to 32 True Positives and 11 False Positives) when detecting new or evolving lesions,

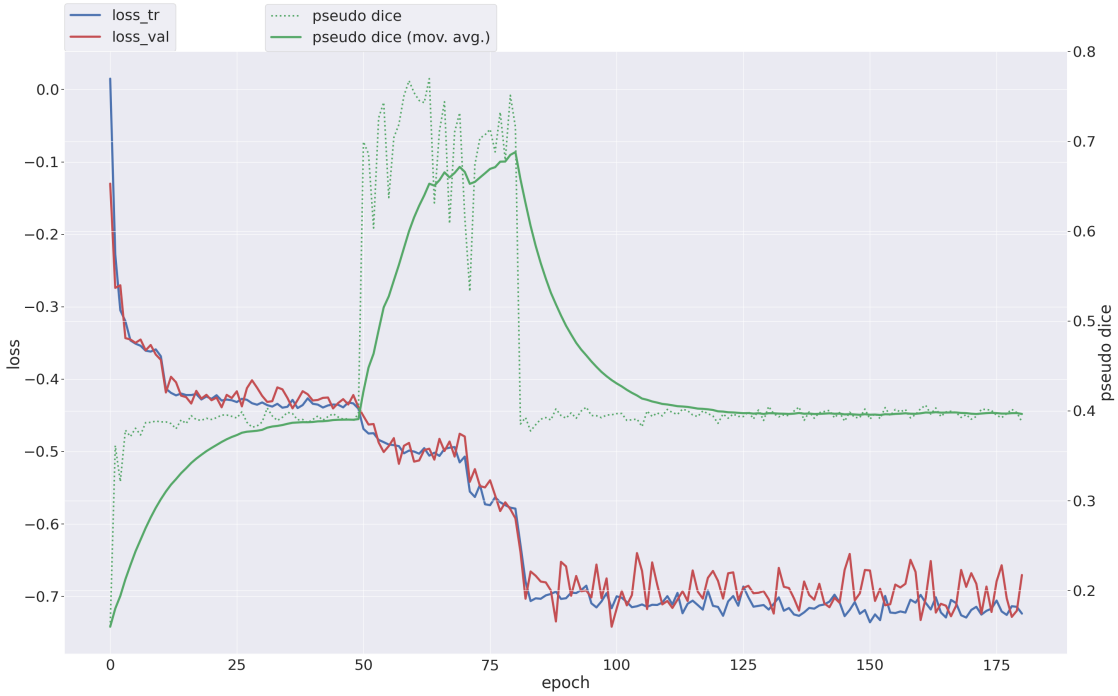

Figure 3.11: The training progress of the first fold baseline model.

which were often confused with the background by the models, was too high. We thought that the scarcity of new or evolving lesions could also be a possible cause of this tendency, and so this was our main focus for the second iteration.

3.3.2 Extreme oversampling of new or evolving lesions

In the second iteration we focused on the oversampling strategy for better handling the class-imbalance between healthy tissue and new or evolving lesions. Concretely, we applied an "extreme" strategy of oversampling new or evolving lesions: we required all training patches to have a foreground class, and whenever new or evolving lesions were present, to choose this class. With this strategy our objective was to make the model better learn the minority class, while still getting a similar performance on the other two due to most patches still having background and basal lesion voxels. It is interesting to notice that this strategy, as radical as it may seem, can be used within the standalone implementation of nnU-Net by simply changing two parameters, so it seems like it has been used by its authors at some point. Nevertheless, the effectiveness of this oversampling strategy was not uniform across all folds, and although it slightly improved cross-validation results, its impact in performance was heavily affected by the drop already seen in the baseline model training, which now occurred before in most folds, as showcased by Figure 3.12. At this point it was clear that solving this issue was essential for

improving the results. As a matter of fact, posterior analysis showed that results on the test split did not improve exclusively with this oversampling strategy change.

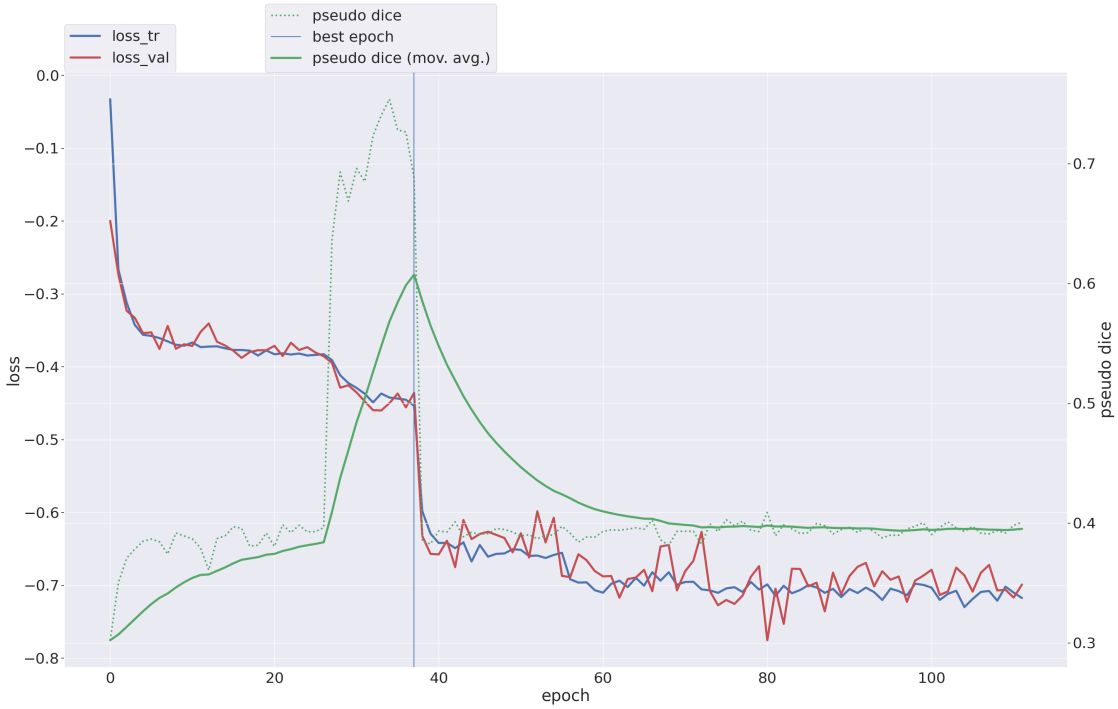

Figure 3.12: The training progress of the first fold model using extreme oversampling.

3.3.3 Improved convergence

In the last iteration we have focused on avoiding the performance drop. At first we saw this drop on performance as a divergence of the training process, since the validation metric improved until a maximum and then quickly decreased. After looking more carefully at the progress plot we noticed an at first confusing pattern: both the training loss and the validation loss decreased remarkably when this drop happened, as clearly shown in Figure 3.12, meaning that the training did not, in fact, diverge, but converge to a suboptimal solution. We finally identified the problem as a saddle point of the loss function used, since although never predicting new or evolving lesions is not a minimum of the loss function, in this point gradients pointing towards better solutions, that is, correct predictions, are difficult to find. The logical solution is to avoid falling into this saddle point, and after some research we have decided for the simplest possible fix: a decrease of the learning rate. To be precise, we have reduced the learning rate to 0.001 and the converge issue has strongly and consistently improved, as exemplified by Figure 3.13, which has considerably improved the results. Furthermore, the results we have obtained in this last experiment are good enough for us to successfully conclude the experimentation phase.

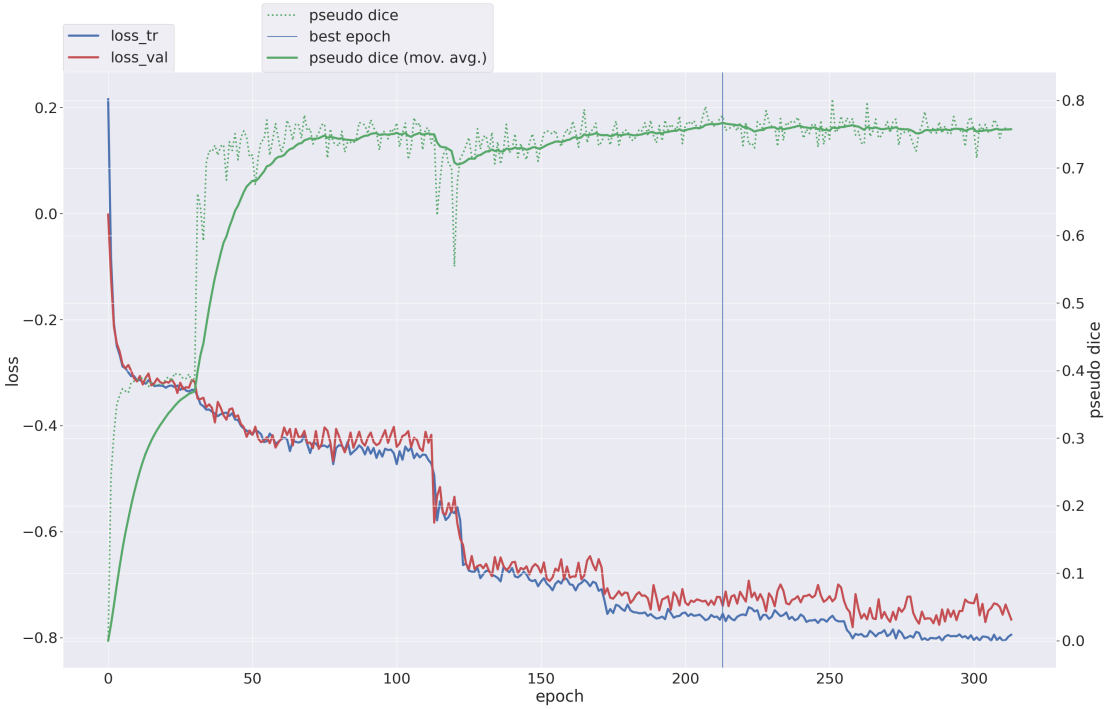

Figure 3.13: The training progress of the first fold model with smaller learning rate.

3.4 Final results

Our final model is the ensemble of the five models we have obtained in the last experiment. To better understand the performance and behavior of the model we have carefully evaluated it. We have defined more adequate metrics and tested the model against both the test split and, additionally, the MSSEG-2 and Open MS Longitudinal Data datasets. It is important to notice that images from these two other datasets look significantly different, as shown by Figure 3.14, and that these results have been obtained with no fine-tuning or adjustment whatsoever.

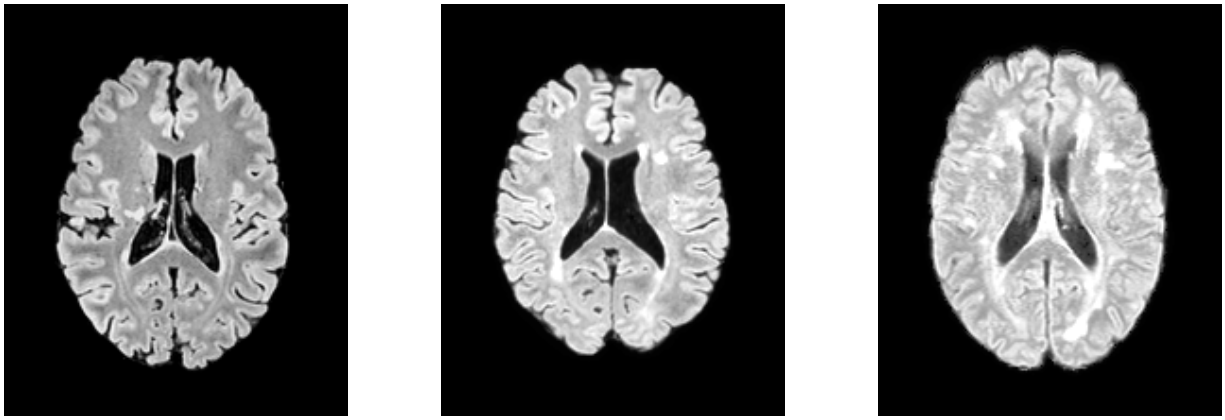

Figure 3.14: One axial view of an image in our, the MSSEG-2 and the MS Open Data datasets.

In order to complement the quantitative evaluation we also display some "confusion plots" (Figures 3.15, 3.16 and 3.17): plots similar to those shown in the qualitative exploration of the dataset that show the True Positives (shown in green), False Positives (shown in red) and False Negatives (shown in blue) of the model when detecting a specific kind of lesion.

3.4.1 Evaluation metrics

For the evaluation we use the same metrics as in the MSSEG-2 challenge, namely:

1. The number of wrongly detected lesion voxels on cases without lesions.
2. The number of wrongly detected lesions on cases without lesions.
3. The voxel level Dice score on cases with lesions.
4. The lesion level F1 score on cases with lesions.

As already explained in section 2.2.2, the main reasons for using this metrics are, on the one hand, that evaluation needs to be performed both at the voxel and at the lesion levels, and on the other hand, that the most adequate metrics for the with lesions subset do not make sense in the without lesions subset. Since both detecting the presence (when present) and the absence (when not present) of new or evolving lesions is essential from the clinical perspective, the performance needs to be evaluated in both situations.

Lesions are identified via a connected-component analysis (with a 18-connectivity kernel) of labels and predictions, and all lesions smaller than $3\,\mathrm{mm}^3$ are removed. Furthermore, a real lesion is considered to be detected (and thus a True Positive of lesion detection) whenever 10% of its voxels are overlapped by predicted lesions. Similarly, a predicted lesion is considered to be correct when ground truth lesions cover at least 10% of its voxels. Notice that this detection criteria is an approximation of the harder and more complex algorithm used in the MSSEG-2 challenge [20]. The original computation also ensures that the detection is not due to an overly large segmentation (i.e. to predicting a lot of lesions), but as our model does not have a tendency to over-predict the computation we use is a good approximation of the original. F1 score is then computed as

$$F_1 := 2 \cdot \frac{Se_L \cdot P_L}{Se_L + P_L},$$

where Se_L is the lesion sensitivity, defined as the number of detected lesions over the number of real lesions, and P_L is the lesion positive predictive value, i.e. the number of correctly predicted lesions over the number of predicted lesions [20].

We evaluate these metrics using a one-vs-all approach for each of the foreground classes, that is, basal and new or evolving lesions.

3.4.2 On the test split

Table 3.3 shows the evaluation metrics on the test split. As we can see the results are exceptional in both the basal lesion and the new or evolving lesion detection, showing an impressive capacity to find and delimitate lesions while generating very few False Positives in the no lesions subset.

Table 3.3: Results on the test split

	Cases without lesions		Cases with lesions	
	Lesion volume	Lesion number	Dice	F1 score
Basal lesions	-	-	0.72	0.73
New or evolving lesions	0.48 mm^3	0.04	0.64	0.75

Figure 3.15 shows an example of correct guesses and the mistakes the model does when predicting basal lesions. One can see that the model only rarely misses a lesion, and when it does happen, it is usually in the most difficult cases: cases where the lesion is not clearly visible or that lie very close to the border of the brain, where it is difficult to distinguish the border itself from lesions. With respect to the segmentation, what was expected happens: the model is normally able to detect and correctly label the central voxels of lesions, but when moving towards the borders the accuracy decreases and mistakes (both FPs and FNs) happen.

When predicting new or evolving lesions, clear lesions are very usually detected, but the model seems to have difficulties detecting diminishing, unclear or border cases. Segmentation performance seems to be less impacted by the issues on boundaries, and more by the poorer performance when detecting lesions: the model seems to detect better new or evolving lesion boundaries, but the fraction of undetected or wrongly detected lesions, where we have made no right guesses, reduces the Dice-score.

Furthermore, the analysis of the confusion matrix shows that our model only very rarely confuses basal and new or evolving lesions, i.e., that most mistakes are predicting background instead of lesion and vice versa.

3.4.3 On the MSSEG-2

Table 3.4 shows the evaluation metrics on the MSSEG-2 dataset, which only has labels for new lesions, together with the corresponding position in each ranking. It is important to remark that evolving lesions, such as diminishing or enlarging lesions, that our model predicts in the same class as new lesions, are not considered in this dataset. Nevertheless, as we can see we get excellent results that rank among the best results in all four categories of the ranking (whose best solutions are shown in Table 2.1), including perfect results in the no-lesions subset.

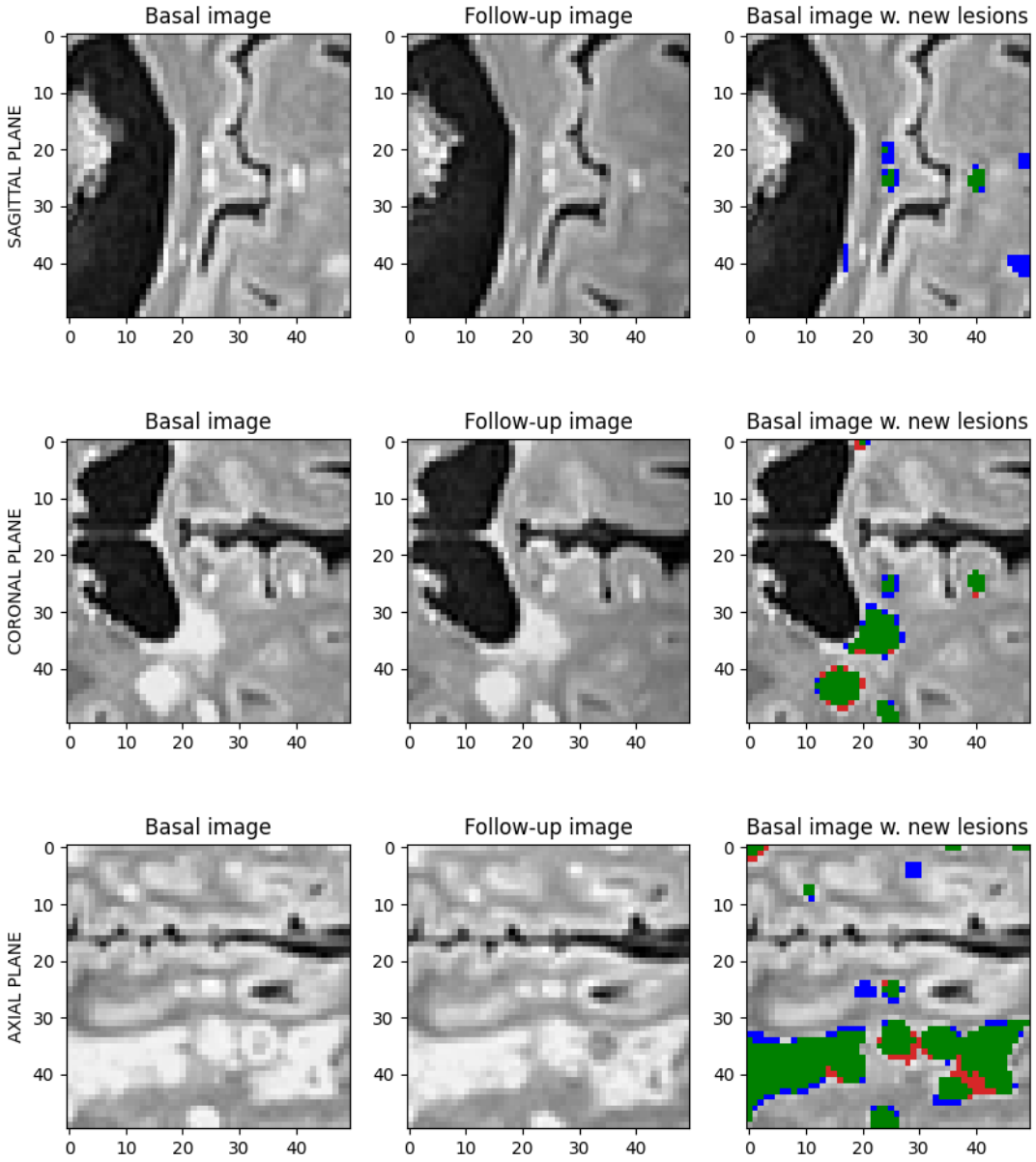

Figure 3.15: Example of the confusions the model makes when detecting basal lesions. True Positives are shown in green, False Positives in red and False Negatives in blue.

Table 3.4: Results on the MSSEG-2 dataset

	Cases without lesions		Cases with lesions	
	Lesion volume	Lesion number	Dice	F1 score
New or evolving lesions	0.00 (**1st**)	0.00 (**1st**)	0.46 (6th)	0.58 (**1st**)

Figure 3.16 shows an example of the correct guesses and mistakes the model does when predicting new or evolving lesions on the MSSEG-2 dataset. In general the difference in brightness and (the lack of) internal contrast between the images in the MSSEG-2 dataset and in ours (visible in Figure 3.14) seem to confuse a bit our model. The direct consequences are that, on the one hand, it is harder for our model to detect some lesions, and, on the other hand, that the model is worse predicting the boundaries of lesions, which worsens the segmentation score but does not affect the lesion-wise F1 score.

3.4.4 On the Open MS Longitudinal Data dataset

Table 3.5 shows the evaluation metrics on the MS Open Data dataset, which only has cases with lesions and labels for new or evolving lesions. These results are undoubtedly much worse than those obtained in the previous two datasets. After some analysis we have found some possible causes. First, images in this dataset have a lower spatial resolution, which usually affects the model's performance. After a qualitative analysis of the images we clearly see that the detection of lesions is much more difficult with this image quality. In addition, although the training images have been augmented to mimic lower resolution images, the model has not seen real low-resolution cases. And second, the average number of lesions and lesion voxels is much larger in this dataset than in our dataset: this dataset has 40 lesions and 6915 lesion voxels on average, many more than in our dataset, where we just have 2 lesions and 154 lesion voxels, as shown in Table 3.1. Therefore, the model needs to find many more lesions in more difficult conditions, which results in a worse performance.

Table 3.5: Results on the Open MS Longitudinal Data dataset

	Cases without lesions		Cases with lesions	
	Lesion volume	Lesion number	Dice	F1 score
New or evolving lesions	-	-	0.27	0.24

For this dataset visual inspection of the images clearly shows why our model has so poor results. Figure 3.17 shows an example of the correct guesses and mistakes the model does when predicting new or evolving lesions on the Open MS Longitudinal Data dataset. In addition to the issue with resolution, in this dataset the images are even brighter than in the MSSEG-2, and

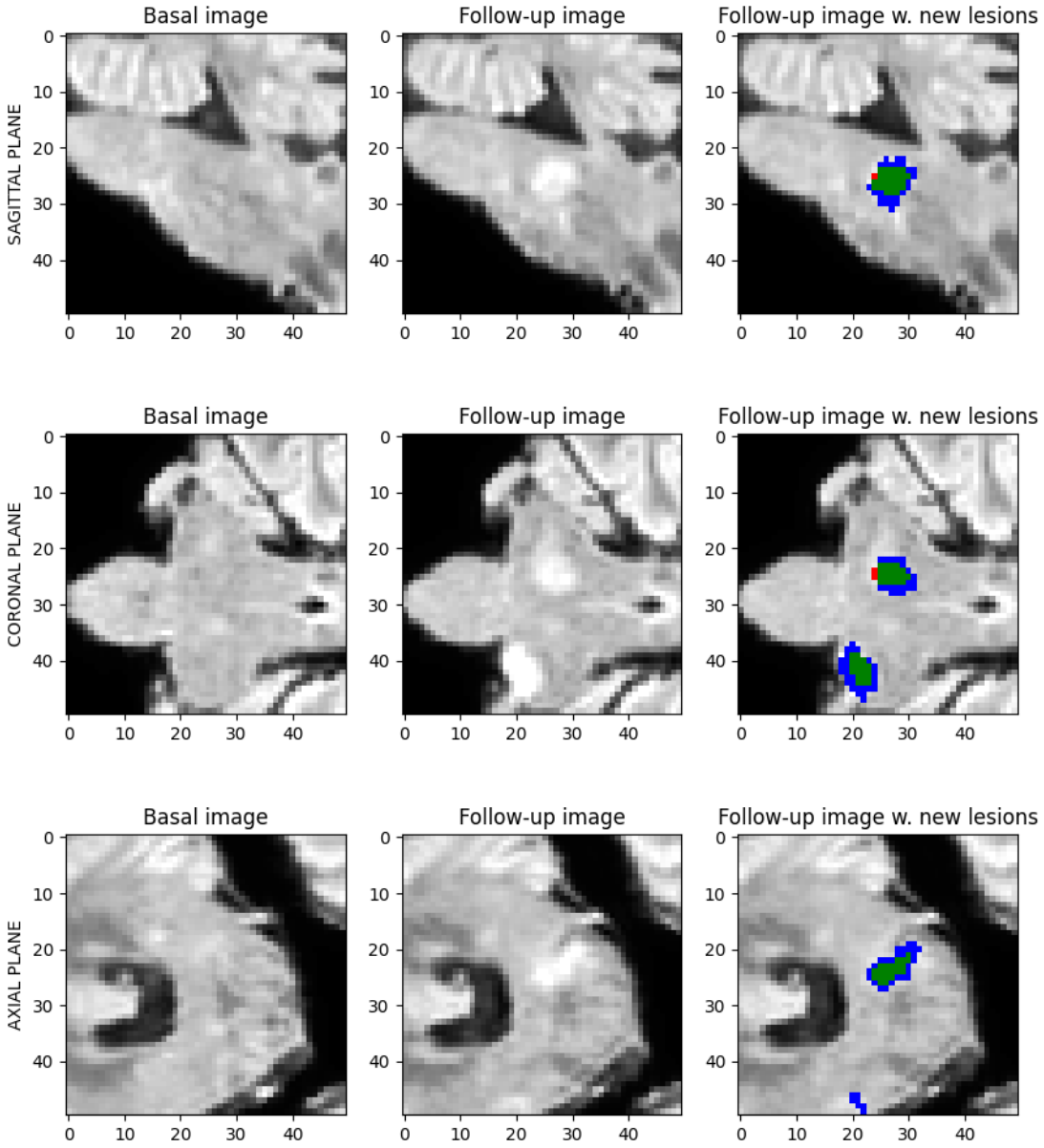

Figure 3.16: Example of the confusions the model makes when detecting new lesions on the MSSEG-2 dataset. TPs are shown in green, FPs in red and FNs in blue.

the contrast lower, making it really hard to correctly identify the lesions and their boundaries. Taking into account that the model has not seen images like these in the training, the lack of generalization of our model when applied to this dataset - evidenced by the much worse metrics - seems normal. Lastly, the amount of diminishing lesions, which our model has trouble detecting, seems to be higher in this dataset.

3.5 Deployment

The last step of the implementation has been the containerization of our best ensemble of models into a Docker container. The Docker container provides a lightweight, portable, resource efficient and isolated environment that encapsulates all the necessary dependencies and configurations required to run the segmentation model. Obtaining the predictions of the model on a set of examples becomes really accessible and user-friendly using the container: it only requires adapting the names of the images to the nnU-Net format (which is necessary for the model to know which images are basal and which are of the follow-up) and running one terminal command. The exact instructions are detailed in the Github repository of the project. Access to the Docker container will be granted upon a fair request.

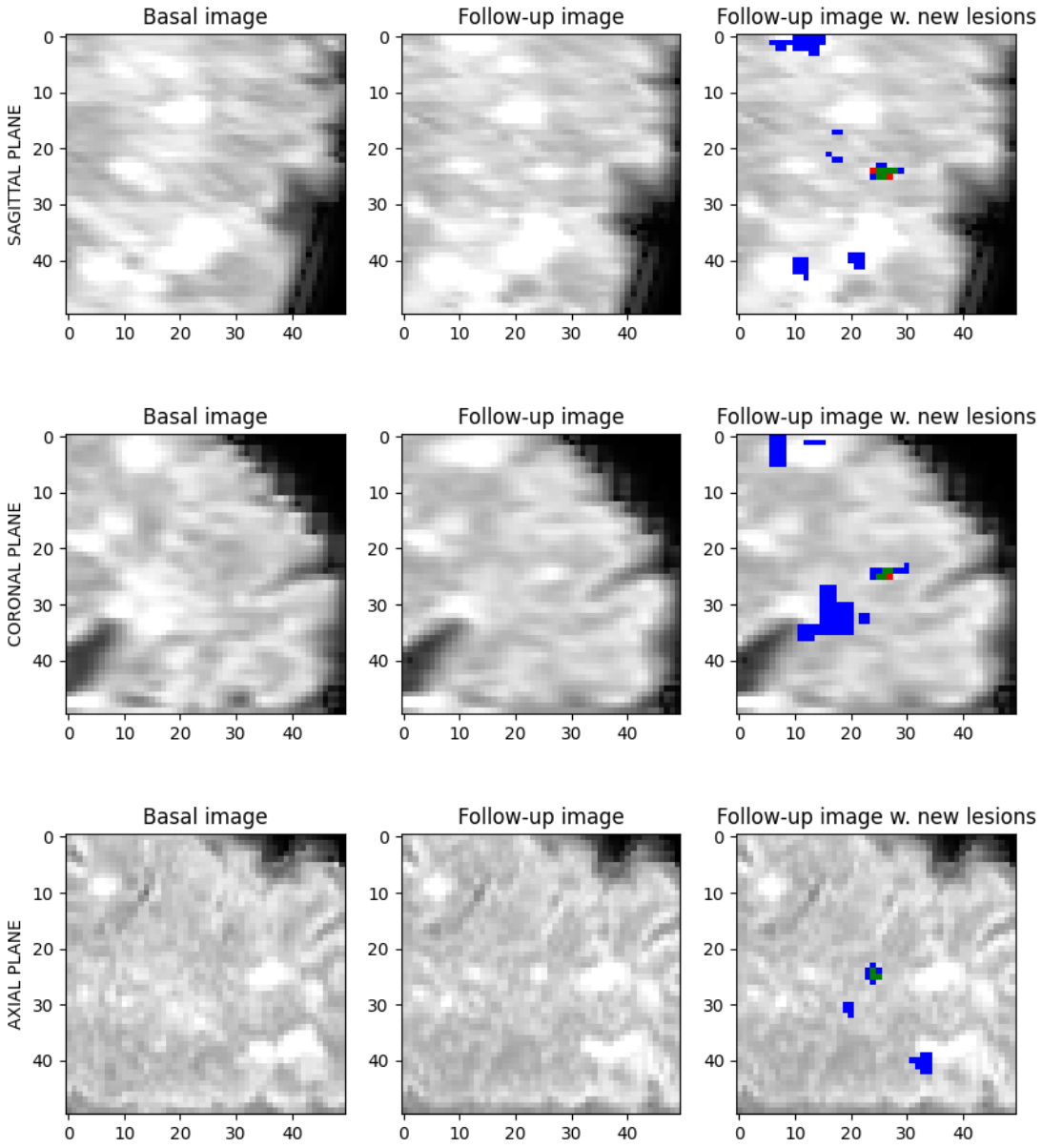

Figure 3.17: Example of the confusions the model makes when detecting new or evolving lesions on the Open MS Longitudinal Data dataset. TPs are shown in green, FPs in red and FNs in blue.

Chapter 4

Summary, Conclusions, and Future Work

In this chapter we summarize our work, draw conclusions and think about future work.

4.1 Summary

In this thesis we have generalized the previous work on the detection of new Multiple Sclerosis lesions on Longitudinal MRI images to the multi-class detection of both new or evolving and basal lesions.

For this purpose we have used a dataset provided by the IDIBAPS through an agreement with the UOC, consisting on 117 basal and follow-up FLAIR images and their corresponding basal and new or evolving lesion labels.

Since there is no published research aiming at our concrete goal, we have used the State of the Art in the new MS lesion detection task as basis for the design of our solution. Most of this research was published within the MICCAI 2021 MSSEG-2 challenge or is closely related to it, so we have given special attention to this challenge. In addition, the dataset used in it, named MSSEG-2 dataset, is public and commonly used as benchmark. From the study of the State of the Art we have identified the challenges of the new lesion detection task and which techniques are most frequently used when facing these challenges, most of which also apply to the multi-class classification extension, together with the commonly used network architectures.

Taking these challenges and techniques in mind we have iteratively designed and developed a semantic segmentation pipeline that automatically detects basal and new or evolving lesions. In brief, we have relied on the nnU-Net semantic segmentation pipeline as a baseline, which includes some preprocessing, data augmentation and oversampling strategies, and have

achieved excellent results in both the basal and new or evolving lesion detection by slightly tuning the oversampling strategy and the training process. Furthermore, the new or evolving lesion detection seems to generalize very well to the MSSEG-2 dataset, where our model ranks among the best solutions in all four metrics and achieves the perfect score in two of them. The generalization results are not as good when tested on the MS Longitudinal Data dataset, but this drop in performance is consistent with the worse image quality of this dataset.

To conclude the project we have also containerized the pipeline and model using Docker, which gives a user-friendly way to obtain basal and new or evolving lesion masks.

4.2 Conclusions

With the work we have done and the results in mind, and careful consideration, we have arrived to the following conclusions.

It is crystal clear the strength and usefulness of nnU-Net as a baseline model in our task. The results obtained with almost no decisions from the user are remarkable, and having such a well-performing baseline with an easy-to-setup tool really speeds up the first iteration of the project, thus allowing to rapidly move on to possible improvements. Nevertheless, tuning the most relevant pieces in the pipeline, such as the oversampling strategy and the training process in our case, is prone to boost the performance, as shown by our results.

As we have seen, the extension to multi-class segmentation makes tuning the training slightly harder. On the other hand, seems that having labels for both basal and new or evolving lesions improves the models' results on both categories when compared to only having one of them (as exemplified, in particular, by the results on the MSSEG-2 dataset), i.e., the model learns from basal lesion labels to better predict new or evolving lesions and vice-versa. Taking both things into consideration, seems that the extra effort spent into labeling both types of lesions and tuning the training is well worth it.

Finally, it is also noteworthy that even with last generation GPUs (the GPU used in this project was released $\sim$6 months ago) the training of 3D semantic segmentation models is still a computational challenge, making the development of projects such as ours in a short time-span much harder.

4.3 Future work

Taking into account the exceptional results we have obtained we believe that the following possible future work lines are specially interesting.

The first and most obvious future step is to try to apply the model to the detection of Slowly Expanding Lesions (SEL). This step initially fell within the scope of this project, as explained in section 1.2, but due to the difficulties with the models' training has finally been left as future work.

It also seems interesting to include the evaluation of the medical criteria - in our case, the McDonald criterion - in the evaluation of models. The final objective of our model is to be helpful for the clinicians using it, and one way to evaluate this is to measure how often the predictions give a correct evaluation of the medical diagnosis criterion.

Although the concrete task we have addressed does not seem specially sensitive to sex and race differences, in order to guarantee a fair model it would be interesting to check that the algorithm performs similarly in all sex and race subgroups in the dataset.

Another possible future work would be that of separating the detection of diminishing (shrinking or disappearing) lesions. We have noticed that our model has difficulties when detecting diminishing lesions, since they are scarce and significantly different from new lesions, so it would be interesting to separate their detection from that of new or evolving lesions.

We believe that it would also be helpful to include the model's uncertainty in the output we give to end-users. Semantic segmentation models output the estimated "probabilities"[1] of a pixel (or voxel, in our case) being of each class, and the class with the biggest probability is selected as the predicted one. But the values themselves are also informative, since they serve to measure the confidence of the model in its predictions, e.g. by computing the entropy of the predicted probabilities, and thus give a user the ability to look more closely at those predictions where the model has not been very sure. The difficulty of this task lies, in our opinion, in finding a user-friendly way of sharing this information with the end-users.

It would also be interesting to assess the impact of fine-tuning for improving the performance on other datasets, such as the MSSEG-2 and the MS Open Data dataset, or the impact of including some of their images in the training. For this purpose we would probably need to label the basal lesions in these images, for which we could precisely use the model's predictions as pre-labels.

Related to the previous point, using the model to pre-label unlabeled data would definitely improve labeling speed, which would increase the amount of data for training and thus the performance of the models. This would be easier with a suitable labeling tool, such as Label Studio, which is already working on compatibility with NIFTI images.

[1] Although the output of classification and semantic segmentation models resemble probabilities, their values normally are not probabilities in the usual sense, meaning that a 0.2 score for a class does not mean a that 20% of similar cases will be of this class. This can be improved through calibration [79].

To conclude with the possible extensions of the project, and related to the last three points, it would be interesting to apply active learning and human-in-the-loop techniques, which are specially interesting in medical tasks [80] and we believe that would definitely improve the performance of the model.

Finally, and as so usually happens in deep learning, further experimentation with the training process could also improve the performance of the model. Concretely, tuning more the combination of learning rate and learning rate decay, using only the cross-entropy loss in the early-stopping (so that the training does not stop as long as the softmax outputs do not worsen) and revisiting the dataset splits are experiments that due to time constraints we have not been able to do and seem promising.

Bibliography

[1] Marcos Diaz-Hurtado, Eloy Martínez-Heras, Elisabeth Solana, Jordi Casas-Roma, Sara Llufriu, Baris Kanber, and Ferran Prados. Recent advances in the longitudinal segmentation of multiple sclerosis lesions on magnetic resonance imaging: a review. *Neuroradiology*, 64(11):2103–2117, 2022.

[2] Alexandre Fenneteau, Pascal Bourdon, David Helbert, Christine Fernandez-Maloigne, Christophe Habas, and Rémy Guillevin. Siamese convolutional neural network for new multiple sclerosis lesion segmentation. In *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 13. Siemens Healthcare, 2021.

[3] Domen Preloznik and Žiga Špiclin. Double Pathway Method For MSSEG-2 Challenge. In *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 41, 2021.

[4] Tiziano Dalbis, Thomas Fritz, Joana Grilo, Sebastian Hitziger, and Wen Xin Ling. 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*, page 57, 2021.

[5] Francesco La Rosa, Jean-Philippe Thiran, and Meritxell Bach. A subtraction image-based method to detect new appearing multiple sclerosis lesions on single-contrast flair mri. In *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 65, 2021.

[6] Brennan Nichyporuk, Kirill Vasilevski, Anjun Hu, Chelsea Myers-Colet, Jillian Cardinell, Justin Szeto, Jean-Pierre Falet, Eric Zimmermann, Julien Schroeter, Douglas L Arnold, et al. Consensus learning with multi-rater labels for segmenting and detecting new lesions. In *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, volume 85, 2021.

[7] Mariem Hamzaoui, Théodore Soulier, Arya Yazdan-Panah, Marius Schmidt-Mengin, Olivier Colliot, Nicholas Ayache, and Bruno Stankoff. Intensity based Regions Of Interest (ROIs) preselection followed by Convolutional Neuronal Network (CNN) based segmentation for new lesions detection in Multiple Sclerosis. In *2nd MICCAI challenge on multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure—MICCAI-MSSEG-2*, 2021.

[8] Mariano Cabezas, Yuling Luo, Kain Kyle, Linda Ly, Chenyu Wang, and Michael Barnett. 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*, page 107, 2021.

[9] Julia Andresen, Hristina Uzunova, Jan Ehrhardt, and Heinz Handels. New multiple sclerosis lesion detection with convolutional neural registration networks. In *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 111, 2021.

[10] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010.

[11] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[12] Eloy MARTÍNEZ-HERAS, Adrian VICENTE-GOMEZ, Francesc VIVÓ, Marcos DIAZ-HURTADO, Baris KANBER, Jordi CASAS-ROMA, Sara LLUFRIU, and Ferran PRADOS. 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.

[13] Mostafa Salem, Sergi Valverde, Mariano Cabezas, Deborah Pareto, Arnau Oliver, Joaquim Salvi, Àlex Rovira, and Xavier Lladó. A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. *NeuroImage: Clinical*, 25:102149, 2020. ISSN 2213-1582. doi: https://doi.org/10.1016/j.nicl.2019.102149. URL https://www.sciencedirect.com/science/article/pii/S2213158219304954.

[14] Berke Doga Basaran, Paul M Matthews, and Wenjia Bai. MSSEG-2 Challenge, Team New Brain: Cascaded networks for new MS lesion detection. In *MSSEG-2 challenge proceed-*

ings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure, page 77, 2021.

[15] Arthur Masson, Brandon Le Bon, Anne Kerbrat, Gilles Edan, Francesca Galassi, and Benoit Combes. 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*, page 5, 2021.

[16] Timo Löhr, Johannes C Paetzold, Anjany Sekobouyina, Suprosanna Shit, Ivan Ezhov, Benedikt Wiestler, and Bjoern H Menze. MSSEG-2 new MS lesions detection and segmentation challenge using a data management and processing infrastructure. In *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 21, 2021.

[17] Julia Krüger, Roland Opfer, Nils Gessert, Ann-Christin Ostwaldt, Praveena Manogaran, Hagen H Kitzler, Alexander Schlaefer, and Sven Schippling. Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3D convolutional neural networks. *NeuroImage: Clinical*, 28:102445, 2020.

[18] Marius Schmidt-Mengin, Arya Yazdan-Panah, Théodore Soulier, Mariem Hamzaoui, Nicholas Ayache, and Olivier Colliot. Segmentation of new multiple sclerosis lesions on FLAIR MRI using online hard example mining. In *2nd MICCAI challenge on multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure—MICCAI-MSSEG-2*, 2021.

[19] Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017.

[20] Olivier Commowick, Audrey Istace, Michael Kain, Baptiste Laurent, Florent Leray, Mathieu Simon, Sorina Camarasu Pop, Pascal Girard, Roxana Ameli, Jean-Christophe Ferré, et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific reports*, 8(1):13650, 2018.

[21] Olivier Commowick, Frédéric Cervenansky, François Cotton, and Michel Dojat. 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, page 126, http://portal.fli-iam.irisa.fr/msseg-2, 2021.

[22] Xavier Lladó, Onur Ganiler, Arnau Oliver, Robert Martí, Jordi Freixenet, Laia Valls, Joan C Vilanova, Lluís Ramió-Torrentà, and Àlex Rovira. Automated detection of multiple sclerosis lesions in serial brain MRI. *Neuroradiology*, 54:787–807, 2012.

[23] Alan J Thompson, Brenda L Banwell, Frederik Barkhof, William M Carroll, Timothy Coetzee, Giancarlo Comi, Jorge Correale, Franz Fazekas, Massimo Filippi, Mark S Freedman, et al. Diagnosis of multiple sclerosis: 2017 revisions of the mcdonald criteria. *The Lancet Neurology*, 17(2):162–173, 2018.

[24] Žiga Lesjak, Franjo Pernuš, Boštjan Likar, and Žiga Špiclin. Validation of white-matter lesion change detection methods on a novel publicly available MRI image database. https://github.com/muschellij2/open_ms_data. *Neuroinformatics*, 14(4):403–420, 2016.

[25] Ferran Prados and Baris Kanber. Detecting new multiple sclerosis lesions using a mixed approach. In *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 1, 2021.

[26] Junghwa Kang, Siyun Jung, Jeongmin Yim, Hyebin Lee, Jinhee Jang, and Yoonho Nam. Segmentation of New Multiple Sclerosis Lesions in Longitudinal MRI Analysis Using a Multi-Stage 3D patch-wise Deep Learning Algorithm. In *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, 2021.

[27] Alexandre Fenneteau, Pascal Bourdon, David Helbert, Christine Fernandez-Maloigne, Christophe Habas, and Rémy Guillevin. Investigating efficient CNN architecture for multiple sclerosis lesion segmentation. *Journal of Medical Imaging*, 8(1):014504–014504, 2021.

[28] Adam Gibicar, Samir Mitha, and April Khademi. Segmentation of New Multiple Sclerosis Lesions using an Ensemble of SC U-Nets with Multi Channel Patch-Based Inputs. In *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 17, 2021.

[29] Jiong Wu, Yue Zhang, Kai Wang, and Xiaoying Tang. Skip connection U-Net for white matter hyperintensities segmentation from MRI. *IEEE Access*, 7:155194–155202, 2019.

[30] Hongwei Li, Jianguo Zhang, Mark Muehlau, Jan Kirschke, and Bjoern Menze. Multi-scale convolutional-stack aggregation for robust white matter hyperintensities segmentation. In

Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4, pages 199–207. Springer, 2019.

[31] Anima scripts: RRID:SCR_017072. URL https://anima.irisa.fr.

[32] Beytullah Sarica and Dursun Zafer Seker. 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*, page 25, 2021.

[33] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[34] Reda Abdellah Kamraoui, Vinh-Thong Ta, José V Manjon, and Pierrick Coupé. New MS lesion Segmentation with Lesion-wise Metrics Learning. In *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 29, 2021.

[35] Reda Abdellah Kamraoui, Vinh-Thong Ta, José V Manjon, and Pierrick Coupé. Draw and Erase to Learn Better. In *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 33, 2021.

[36] Reda Abdellah Kamraoui, Vinh-Thong Ta, José V Manjon, and Pierrick Coupé. 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*, page 37, 2021.

[37] Reda Abdellah Kamraoui, Vinh-Thong Ta, Thomas Tourdias, Boris Mansencal, José V Manjon, and Pierrick Coupé. Towards broader generalization of deep learning methods for multiple sclerosis lesion segmentation. *arXiv preprint arXiv:2012.07950*, 2020.

[38] Pooya Ashtari, Berardino Barile, Sabine Van Huffel, and Dominique Sappey-Marinier. 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*, page 45, 2021.

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.

[40] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570. PMLR, 2015.

[41] Medical Open Network for AI (MONAI). URL https://github.com/Project-MONAI/MONAI.

[42] Cory Efird, Dylan Miller, and Dana Cobzas. A UNet Pipeline for Segmentation of New MS Lesions. In *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 53, 2021.

[43] Advanced normalization tools. URL http://stnava.github.io/ANTs/.

[44] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Bridging category-level and instance-level semantic image segmentation. *arXiv preprint arXiv:1605.06885*, 2016.

[45] Huahong Zhang, Hao Li, and Ipek Oguz. 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*, page 61, 2021.

[46] Huahong Zhang, Alessandra M Valcarcel, Rohit Bakshi, Renxin Chu, Francesca Bagnato, Russell T Shinohara, Kilian Hett, and Ipek Oguz. Multiple sclerosis lesion segmentation with tiramisu and 2.5 d stacked slices. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pages 338–346. Springer, 2019.

[47] Stefano Cerri, Andrew Hoopes, Douglas N Greve, Mark Mühlau, and Koen Van Leemput. A longitudinal method for simultaneous whole-brain and lesion segmentation in multiple sclerosis. In *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology: Third International Workshop, MLCN 2020, and Second International Workshop, RNO-AI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3*, pages 119–128. Springer, 2020.

[48] Uzay Macar, Enamundram Naga Karthik, Charley Gros, Andréanne Lemay, and Julien Cohen-Adad. Team NeuroPoly: Description of the Pipelines for the MICCAI 2021 MS New

Lesions Segmentation Challenge. https://github.com/ivadomed/ms-challenge-2021. In *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 69, 2021.

[49] Benjamin De Leener, Simon Lévy, Sara M Dupont, Vladimir S Fonov, Nikola Stikov, D Louis Collins, Virginie Callot, and Julien Cohen-Adad. Sct: Spinal cord toolbox, an open-source software for processing spinal cord mri data. *Neuroimage*, 145:24–43, 2017.

[50] Medical Image Registration ToolKit (MIRTK). URL https://mirtk.github.io/.

[51] Md Mahfuzur Rahman Siddiquee and Andriy Myronenko. Robust 3D MRI Segmentation of Multiple Sclerosis Lesions. In *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 81, 2021.

[52] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pages 311–320. Springer, 2019.

[53] Martin Holm Jensen, Dan Richter Jørgensen, Raluca Jalaboi, Mads Eiler Hansen, and Martin Aastrup Olsen. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 540–548. Springer, 2019.

[54] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12341–12351, 2021.

[55] Brennan Nichyporuk, Justin Szeto, Douglas Arnold, and Tal Arbel. Optimizing operating points for high performance lesion detection and segmentation using lesion size reweighting. In *Medical Imaging with Deep Learning*, 2021.

[56] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[57] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008.

[58] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

[59] Richard McKinley, Franca Wagner, and Roland Wiest. Detection of lesion change in multiple sclerosis using a cascade of 3D-to-2D networks. In *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 97, 2021.

[60] Ulrike W Kaunzner and Susan A Gauthier. Mri in the assessment and monitoring of multiple sclerosis: an update on best practice. *Therapeutic advances in neurological disorders*, 10(6):247–261, 2017.

[61] Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40(17):4952–4964, 2019.

[62] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[63] Julia Andresen, Timo Kepp, Jan Ehrhardt, Claus von der Burchard, Johann Roider, and Heinz Handels. Deep learning-based simultaneous registration and unsupervised non-correspondence segmentation of medical images with pathologies. *International Journal of Computer Assisted Radiology and Surgery*, 17(4):699–710, 2022.

[64] Mostafa Salem, Arnau Oliver, Joaquim Salvi, and Xavier Lladó. Msdetector: A fully convolutional neural network for the detection of new t2-w lesion in multiple sclerosis. *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, page 115, 2021.

[65] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

[66] Juan Eugenio Iglesias, Cheng-Yi Liu, Paul M Thompson, and Zhuowen Tu. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE transactions on medical imaging*, 30(9):1617–1634, 2011.

[67] László G Nyúl, Jayaram K Udupa, and Xuan Zhang. New variants of a method of mri scale standardization. *IEEE transactions on medical imaging*, 19(2):143–150, 2000.

[68] Reda Abdellah Kamraoui, Boris Mansencal, José V. Manjon, and Pierrick Coupé. Longitudinal detection of new ms lesions using deep learning. *Frontiers in Neuroimaging*, 1, 2022. ISSN 2813-1193. doi: 10.3389/fnimg.2022.948235. URL https://www.frontiersin.org/articles/10.3389/fnimg.2022.948235.

[69] Marius Schmidt-Mengin, Théodore Soulier, Mariem Hamzaoui, Arya Yazdan-Panah, Benedetta Bodini, Nicholas Ayache, Bruno Stankoff, and Olivier Colliot. Online hard example mining vs. fixed oversampling strategy for segmentation of new multiple sclerosis lesions from longitudinal flair mri. *Frontiers in Neuroscience*, 16, 2022. ISSN 1662-453X. doi: 10.3389/fnins.2022.1004050. URL https://www.frontiersin.org/articles/10.3389/fnins.2022.1004050.

[70] Sebastian Hitziger, Wen Xin Ling, Thomas Fritz, Tiziano D'Albis, Andreas Lemke, and Joana Grilo. Triplanar u-net with lesion-wise voting for the segmentation of new lesions on longitudinal mri studies. *Frontiers in Neuroscience*, 16, 2022. ISSN 1662-453X. doi: 10.3389/fnins.2022.964250. URL https://www.frontiersin.org/articles/10.3389/fnins.2022.964250.

[71] Mostafa Salem, Marwa Ahmed Ryan, Arnau Oliver, Khaled Fathy Hussain, and Xavier Lladó. Improving the detection of new lesions in multiple sclerosis with a cascaded 3d fully convolutional neural network approach. *Frontiers in Neuroscience*, 16, 2022. ISSN 1662-453X. doi: 10.3389/fnins.2022.1007619. URL https://www.frontiersin.org/articles/10.3389/fnins.2022.1007619.

[72] Berke Doga Basaran, Paul M. Matthews, and Wenjia Bai. New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation. *Frontiers in Neuroscience*, 16, 2022. ISSN 1662-453X. doi: 10.3389/fnins.2022.1007453. URL https://www.frontiersin.org/articles/10.3389/fnins.2022.1007453.

[73] Xinru Zhang, Chenghao Liu, Ni Ou, Xiangzhu Zeng, Xiaoliang Xiong, Yizhou Yu, Zhiwen Liu, and Chuyang Ye. Carvemix: A simple data augmentation method for brain lesion segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*

2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I, pages 196–205, 2021.

[74] Liliana Valencia, Albert Clèrigues, Sergi Valverde, Mostafa Salem, Arnau Oliver, Àlex Rovira, and Xavier Lladó. Evaluating the use of synthetic t1-w images in new t2 lesion detection in multiple sclerosis. *Frontiers in Neuroscience*, 16, 2022. ISSN 1662-453X. doi: 10.3389/fnins.2022.954662. URL https://www.frontiersin.org/articles/10.3389/fnins.2022.954662.

[75] Maryam Hashemi, Mahsa Akhbari, and Christian Jutten. Delve into multiple sclerosis (ms) lesion exploration: A modified attention u-net for ms lesion segmentation in brain mri. *Computers in Biology and Medicine*, 145:105402, 2022. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2022.105402. URL https://www.sciencedirect.com/science/article/pii/S0010482522001949.

[76] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. FSL. https://fsl.fmrib.ox.ac.uk/fsl/fslwiki. *Neuroimage*, 62 (2):782–790, 2012.

[77] Marco Ganzetti, Nicole Wenderoth, and Dante Mantini. Intensity inhomogeneity correction of structural mr images: a data-driven approach to define input algorithm parameters. *Frontiers in neuroinformatics*, 10:10, 2016.

[78] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016.

[79] Dongdong Wang, Boqing Gong, and Liqiang Wang. On calibrating semantic segmentation models: Analyses and an algorithm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23652–23662, 2023.

[80] Samuel Budd, Emma C. Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71: 102062, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2021.102062. URL https://www.sciencedirect.com/science/article/pii/S1361841521001080.