A comparative evaluation of diffusion based networks for Multiple Sclerosis lesion segmentation

Alessia Rondinella[1*], Francesco Guarnera[1], Alessandro Ortis[1], Elena Crispino[2], Giulia Russo[3], Francesco Pappalardo[3], Sebastiano Battiato[1]

[1*]Department of Mathematics and Computer Science, University of Catania, Catania, 95125, Italy.
[2]Department of Biomedical and Biotechnological Sciences, University of Catania, Catania, 95125, Italy.
[3]Department of Drug and Health Sciences, University of Catania, Catania, 95125, Italy.

*Corresponding author(s). E-mail(s): alessia.rondinella@phd.unict.it;
Contributing authors: francesco.guarnera@unict.it;
alessandro.ortis@unict.it; elena.crispino@phd.unict.it;
giulia.russo@unict.it; francesco.pappalardo@unict.it;
sebastiano.battiato@unict.it;

Abstract

Semantic segmentation of Multiple Sclerosis (MS) lesions from longitudinal Magnetic Resonance Imaging (MRI) scans is crucial for the diagnosis and monitoring of disease progression. This study aims to evaluate the generalization performance of various deep learning segmentation models, which are commonly used in state-of-the-art medical image segmentation, when integrated into a diffusion model pipeline for segmenting MS lesions. Through an extensive set of experiments, we assess the performance of diffusion models with different architectural configurations to identify the optimal model for MS lesion segmentation. Additionally, we explored the robustness of diffusion model predictions by implementing various inference strategies to combine the diffusion model outputs obtained at each time step. Our results demonstrate the effectiveness of certain backbone architectures in enhancing diffusion model performance in MS lesion segmentation. Moreover, we demonstrate that accurate selection of inference strategies can further enhance the accuracy and robustness of diffusion model predictions. This study contributes to advancing the understanding of

diffusion models' applicability in clinical settings and provides insights for improving MS lesion segmentation in MRI. Our source code is freely available at https://github.com/alessiarondinella/MSSegDiff.

Keywords: Multiple Sclerosis, Denoising Diffusion Models, Lesion segmentation, Medical image analysis

1 Introduction

Multiple Sclerosis (MS) is an inflammatory disorder affecting the Central Nervous System (CNS). It is a chronic and debilitating demyelinating disease, characterized by focal areas of inflammation and subsequent myelin and axonal loss [1]. MS leads to a broad spectrum of neurological symptoms and disabilities. Typical symptoms of MS include fatigue, difficulty walking, numbness or tingling, muscle weakness, spasticity, vision problems, dizziness, bladder and bowel dysfunction, cognitive changes, and emotional disturbances [2]. Due to the difficulty in diagnosing multiple sclerosis, it is often identified and treated only after the disease has progressed to an advanced stage, causing significant neurophysiological damage. At this point, it is no longer possible to delay or halt the progression of the disease.

Diagnosing and analyzing MS in its early stages is quite challenging even for qualified and experienced radiologists due to the highly complex nature of the brain. Imaging plays a crucial role in the early diagnosis of MS, with Magnetic Resonance Imaging (MRI) scans assisting doctors in the diagnostic process. The longitudinal brain MRI protocol encompasses a variety of sequences, each providing unique contrasts for delineating brain tissues. Notably, Fluid Attenuated Inversion Recovery (FLAIR), T1-weighted (T1-w), T2-weighted (T2-w), and PD-weighted (PD-w) images are essential for detecting MS lesions. Among these, FLAIR images stand out, offering a high-contrast view of lesions and allowing for their clear demarcation from surrounding tissues. MS lesions appear hyperintense on FLAIR images, which are more sensitive than T2-w images in detecting juxtacortical and periventricular plaques. T1-w images, on the other hand, offer insights into chronic areas of atrophy or "black holes" [3], while T2-w images highlight regions with increased water content, making them more sensitive to infratentorial lesions. Acute lesions often appear with surrounding edema on T2-w images. PD-w images, in turn, are especially effective at detecting cervical spinal cord MS lesions, particularly when T2-w images fail to demonstrate them [4]. Accurately interpreting these results remains a complex and time-consuming task for physicians. Therefore, there is a need for a fully automated tool for diagnosing multiple sclerosis, particularly in its early stages. Indeed, accurate detection and localization of MS lesions are critical for clinical evaluation and treatment planning.

This study aims to evaluate the generalization capability of different recent deep learning segmentation models when integrated into MSSegDiff [5], a diffusion model based pipeline for segmenting MS lesions. Specifically, we vary this architecture by incorporating different main backbone commonly used in state-of-the-art medical

image segmentation. Through an extensive set of experiments, we evaluate the performance of diffusion models under different architectural configurations to determine the best model for MS lesion segmentation. Additionally, we examine the robustness of diffusion model predictions by implementing various inference strategies to combine the diffusion model outputs obtained at each time step.

Our results demonstrate the effectiveness of certain backbone architectures in enhancing diffusion model performance in MS lesion segmentation. Furthermore, we show that accurate selection of inference strategies can further improve the accuracy and robustness of diffusion model predictions.

The efficacy of the architectures was evaluated on the ISBI2015 dataset, employing a cross-validation scheme in patients with lesions in both baseline and follow-up scans.

The remainder of this paper is organized as follows. Section 2 provides an overview of the current state-of-the-art, while Section 3 describes the main modules of the architecture. Section 4 describes implementation details, metrics and configuration setup, with information on the dataset and its preprocessing. Section 5 reports the comparative study, and finally Section 6 concludes the paper.

2 Related Work

Recent advances in medical image segmentation, have leveraged a variety of deep learning architectures to enhance accuracy and efficiency. In the last few years, several notable approaches have emerged, each contributing to the field with unique methodologies and improvements. One significant development is the application of transformer-based models in medical imaging. The Vision Transformer (ViT), introduced in [6] has been adapted for segmentation tasks with promising results. Following this, the Swin Transformer, proposed in [7], has shown superior performance by utilizing hierarchical feature extraction with shifted windows, allowing for efficient global context modeling. Following these works, numerous studies have proposed modifying the backbone structures of medical image segmentation models by incorporating transformers, such as SwinUNet [8], a hybrid model combining Swin Transformer and U-Net that achieves robust performance in multi-organ segmentation task. Another noteworthy approach is the use of convolutional neural networks (CNNs) with attention mechanisms [9]. Authors in [10] proposed an Attention U-Net, integrates spatial attention gates to highlight relevant features, improving segmentation accuracy. This model has inspired the development of various attention-based architectural adaptations aiming to enhance segmentation performance, as done in [11] which propose a Fully Convolutional DenseNet with attention blocks for MS lesion segmentation in 2D images and in [12] where authors propose an Attention u-Net for the same purpose. Moreover, diffusion models have gained traction in the field of medical image analysis. These models, originally introduced for generative tasks, have been adapted for image segmentation by incorporating noise-injection and denoising processes to improve robustness and accuracy of prediction results. Authors in [13], proposed a segmentation model based on Diffusion Probabilistic Models (DDPM) [14] with a dynamic conditional encoding, which aims to learn segmentation by conditioning with the image prior. The same authors in [15] that integrate transformer into a diffusion

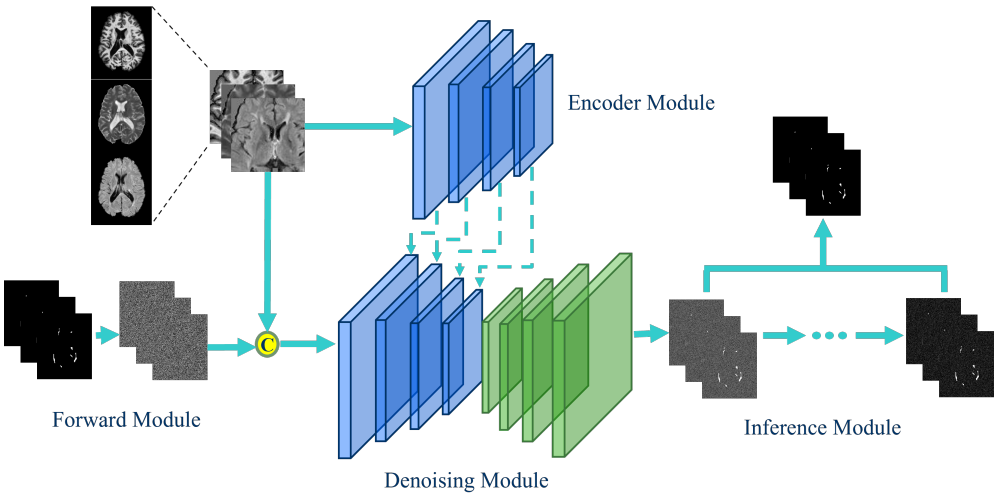


Fig. 1 Overview of the MSSegDiff [5] architecture. The architecture consists of four key modules: (1) the Forward Module, which implements the forward diffusion process by gradually adding Gaussian noise to the ground-truth segmentation mask; (2) the Encoder Module, which captures discriminative features from the channel-wise concatenated volumetric MRI sequences ('C' in the figure represents the channel-wise concatenation); (3) the Denoising Module, responsible for the reverse diffusion process, progressively removing the noise to produce a clean segmentation mask; and (4) the Inference Module, which combines predictions from each timestep to generate the final segmentation mask.

segmentation model with various conditional techniques over the denoising network to perform multi-organ segmentation.

These advancements highlight the ongoing evolution of segmentation techniques, with recent models focusing on integrating attention mechanisms, leveraging transformer architectures, and employing diffusion processes to achieve state-of-the-art performance in medical image segmentation tasks.

3 Overview of the MSSegDiff Architecture

In this section, the general scheme of MSSegDiff will be presented. As depicted in Figure 1 the overall pipeline comprises the following main modules: Forward Module, Encoder Module, Denoising Module and Inference Module. The main modules added to the architecture will be described below.

3.1 Forward Module

This module is responsible for implementing the forward diffusion process, which involves introducing T steps of Gaussian noise to the input image x_0 (Figure 2). This process gradually corrupts the original ground-truth x_{GT}, adding more noise until the information from the original image is completely destroyed and becomes just noise.

To formalize this diffusion process, we consider it as a fixed Markov chain with T steps, where the image at time step t maps to its subsequent state at timestep $t + 1$.

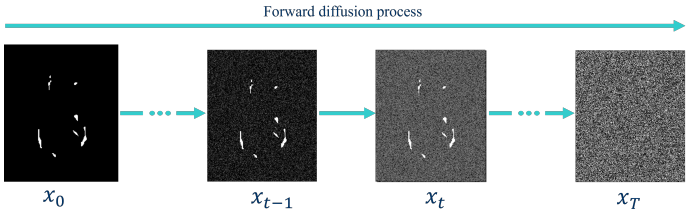


Fig. 2 This image depicts the forward module. The MS lesion segmentation mask is gradually perturbed with Gaussian noise, for a number of timesteps T, until it becomes completely noisy.

By modeling this as a Markov chain, it is possible to derive a formula to obtain the corrupted image at any time step directly, without the need for iterative computation. This streamlined approach greatly accelerates the diffusion process. Each step of the forward diffusion process is defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

Here, q represents the forward process, x_t is the output of the forward process at step t, with x_{t-1} being the output at the previous step. $\mathcal{N}$ is the normal distribution with mean $\sqrt{1-\beta_t}x_{t-1}$ and variance $\beta_t I$.

We set $T = 1000$ for all experiments, based on the original DDPM paper [14], where this value was selected for their experiments. Furthermore, the noise addition at each step follows a predetermined pattern determined by a β_t scheduler, with values ranging from $[0, 1]$. In line with the approach proposed in [14], we adopted a linear schedule for β_t, ranging from 1e$-$4 at timestep 0 to 0.02 at timestep T. This schedule ensures that the amount of noise added gradually increases over the course of the diffusion process, as described in the DDPM paper.

3.2 Encoder Module

The primary objective of the Encoder Module (EM) of MSSegDiff (Figure 3) is to extract an embedding that captures the discriminative features from the volumetric MRI sequences used in this study (T1-w, T2-w, and FLAIR modalities), which are concatenated along the channel dimension. The input to the EM is a volumetric image I of size $M \times D \times W \times H$, where M represents the number of MRI modalities, and D, W, and H correspond to the depth, width, and height of the volumetric image, respectively. By concatenating multiple MRI sequences in this manner and passing them through a 3D encoder, the module aims to leverage complementary information from the different imaging modalities to enhance the segmentation process. Each MRI sequence provides unique insights regarding tissue characteristics and neurodegenerative pathology. For instance, FLAIR sequences are sensitive to inflammation enabling for increased sensitivity in detecting hyperintense lesion in the periventricular areas, T1-w images offer anatomical details revealing any old areas of atrophy or black holes, and T2-w images highlight areas of increased water content, particularly sensitive in

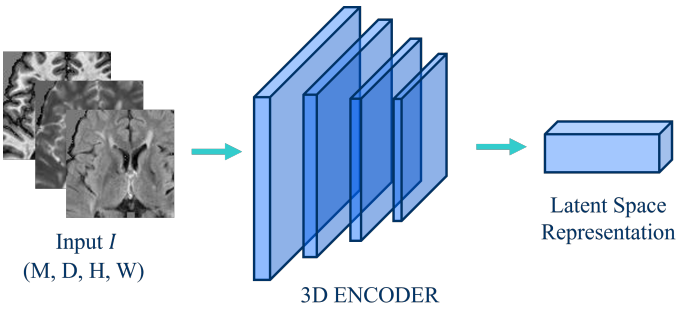


Fig. 3 This image depicts the structure of the encoder module. It takes as input a concatenated volumetric MRI and outputs a segmentation feature vector.

detecting infratentorial lesions. By combining these modalities, our feature extraction module aims to capture a comprehensive representation of the underlying tissue properties, enhancing the discriminative power of the segmentation model. Moreover, the use of volumetric images allows for the preservation of spatial information across different slices, enabling the model to consider 3D structural context of MS lesions during feature extraction. This is crucial in medical imaging tasks where lesion size and distribution can vary significantly in different anatomical regions.

In general, the extraction of segmentation features from concatenated volumetric MRI sequences enables our model to effectively capture both spatial and multimodal information, leading to improved lesion segmentation performance and enhanced clinical utility in the diagnosis and management of multiple sclerosis.

3.3 Denoising Module

The Denoising Module (DM), shown in Figure 4, aims to generate a "clean" MS lesion segmentation mask x_0 by reversing the diffusion process. The output x_T, produced in the Forward Module (Section 3.1), undergoes an iterative denoising procedure. At each timestep, the DM predicts the amount of noise present in the input and subtracts a portion of it according to a predefined schedule. Initially, only a small fraction of the predicted noise is subtracted, but as the process continues, progressively larger portions are removed. This iterative approach allows for a gradual refinement of the segmentation, with each step producing a result that is increasingly accurate and contains less noise. Unlike generative models such as GANs, which generate a clean image in a single step, diffusion models reconstruct the input over multiple timesteps, resulting in a segmentation mask that converges to the true segmentation as the iterations proceed. This involves employing probabilities to make informed estimations about the appearance of the data before noise introduction. This capability is fundamental for the model to accurately reconstruct data, ensuring that the outputs are not only devoid of noise but also closely resemble the original data.

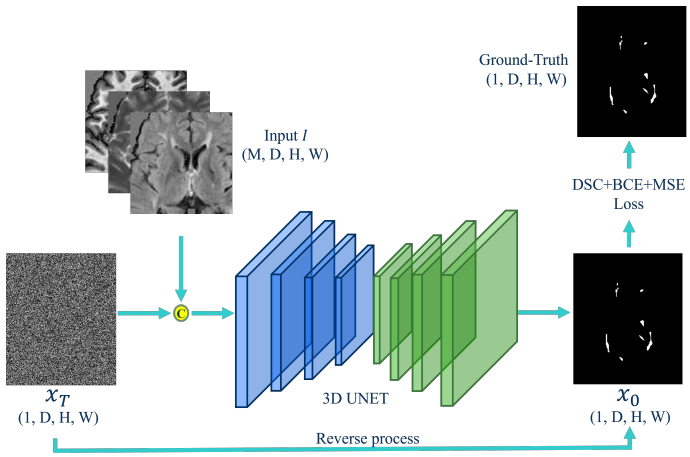


Fig. 4 This image shows the structure of the Denoising Module. It implements the reverse process of the diffusion model in order to generate a clean segmentation mask.

3.4 Inference Module

At test time, the diffusion model iterates through T steps using the Denoising Diffusion Implicit Models (DDIM) method [16]. The Inference Module (IM) at each step generates an increasingly refined segmentation mask. Building on the understanding that the accuracy of predictions improves and uncertainty decreases with an increasing number of testing steps, the segmentation masks obtained from each iteration were combined. Figure 8 represents an example of mask produced by the IM.

4 Overall Pipeline

In MSSegDiff, the noise has been introduced to the segmentation masks since they represent what we aim to learn. The objective of this work is the evaluation of the performance of the diffusion process applied to different backbone for multiple sclerosis lesion segmentation in volumetric MRI data. MSSegDiff takes multimodal MRI and the corresponding ground-truth as input and learns to gradually remove random noise with the aim of generating clear segmentation maps of MS lesions. MSSegDiff implements a U-Net backbone, an encoder-decoder architecture widely used for segmentation in the medical domain. This network takes as input: the input volume I, consisting of three MRI modalities, with dimensions $M \times D \times W \times H$, and the noisy ground-truth x_T after undergoing the forward process (3.1), concatenated channel-wise.

As discussed, DM is conditioned by the segmentation feature extracted from the raw MRI data. In fact, the input volume is passed through the Encoder Module that extract the latent space representation, a projection of the input data in the latent space, representing the most significant segmentation features. This features helps to minimize the variability in the diffusion process, leading to more consistent and accurate segmentation results. The Encoder Module has the same dimensions as the encoder of the U-Net, allowing for the addition of conditional features extracted from the input volume by the Encoder Module with the features extracted at each

step of the downsampling path of the U-Net, as they have the same dimension and number. Once this is done, the upsampling path of the U-Net, identical and opposite to the downsampling path, will reconstruct the original image, leveraging the skip connection to recover high-level details lost during the downsampling process. It takes these combined features as input and returns the clean segmentation mask x_0 as output.

4.1 Implementation details

MSSegDiff was implemented and tested using the PyTorch [17] and MONAI frameworks [18]. Training was conducted on a single NVIDIA A100 GPU, using a batch size of 2 and the AdamW optimizer with a base learning rate of 1e−4 and a weight decay of 1e−3. To dynamically adjust the learning rate (LR), a Cosine Annealing LR schedule [19] was employed. Initially, the LR linearly increased and was gradually reduced following a cosine function after an initial warmup period. To diversify the training data, a patch size of $96 \times 96 \times 96$ was randomly selected in each epoch. This approach ensured that every batch could contain different parts of the same patient, enhancing the model's ability to generalize. During inference, a sliding windows algorithm with overlapping patches of 0.5 was employed. This strategy ensured that no detail was missed while providing intrinsic data augmentation. Additionally, it helped reduce the amount of memory required while maintaining the detail of the original input image. This model is trained with a combination of Dice (DSC) Loss, Binary Cross Entropy (BCE) Loss and Mean Squared Error (MSE) Loss, and thus the finally Loss of our model is:

$$Loss(x_0, x_{GT}) = Loss_{DSC} + Loss_{BCE} + Loss_{MSE} \tag{2}$$

4.2 Evaluation metrics

The model evaluation process entailed a comparison between the predicted segmentation masks and the provided ground-truth data annotated by Rater 1.

The Dice Score (DSC) [20] served as a primary evaluation metric, quantifying the spatial overlap between the predicted and ground-truth segmentation masks. This metric measures the similarity between the two masks, with a higher Dice Score indicating greater agreement between the predicted and true lesion regions.

In addition, the True Positive Rate (TPR) and the Positive Predictive Rate (PPV) has been employed as complementary evaluation metrics. The TPR, also known as sensitivity, measures the proportion of true positive predictions among all actual positive instances in the ground-truth data. It reflects the model's ability to correctly identify lesion regions. Conversely, the PPV, also referred to as precision, assesses the accuracy of positive predictions made by the model. It represents the proportion of true positive predictions among all positive predictions made by the model.

Furthermore, the evaluation incorporated Lesion False Positive Rate (LFPR) and Lesion True Positive Rate (LTPR) to assess the model's performance at a lesion-level, where LFPR quantifies the rate of false positive lesions and LTPR measures the

Table 1 Table displaying the five dataset folds chosen for all training. A leave-one-subject-out-cross-validation approach is consistently employed, ensuring that a different patient is retained for testing in each fold during the configuration selection process.

Fold	Training	Validation	Test
Fold1	1, 2, 3	4	5
Fold2	1, 2, 5	3	4
Fold3	1, 4, 5	2	3
Fold4	3, 4, 5	1	2
Fold5	2, 3, 4	5	1

rate of correctly identified lesions. These lesion-wise metrics provide a more detailed evaluation of the model's capacity to distinguish between lesion and non-lesion regions. Additionally, the Absolute Volume Difference (AVD) and Average Symmetric Surface Distance (ASSD) were used to assess the volumetric and surface differences between the predicted and ground-truth masks. The AVD measures the absolute difference in lesion volume between the predicted and ground-truth segmentations, while the ASSD evaluates the average distance between corresponding surfaces of the predicted and actual lesion masks. These metrics offer further insights into the accuracy and clinical relevance of the segmentation results.

4.3 Configuration Setup

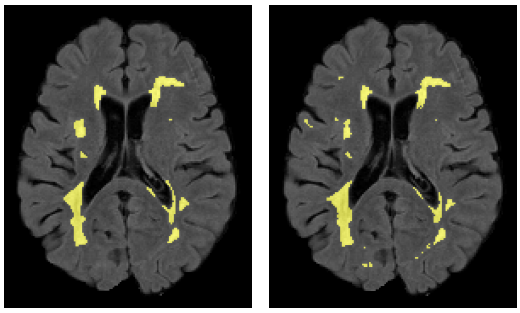
Table 1 details the distribution of patients across different folds used in the cross-validation process. Specifically, all configurations were trained using a Leave-One-Subject-Out-Cross-Validation (LOSO-CV). In LOSO-CV a patient is reserved for the evaluation, an other for the test and the model is trained on remaining patients. This ensures that each validation and test set includes an entire patient (a patient with all of its time points) and also that the network did not have seen the test patient's data during training, nor any of its time points.

This method contrasts with some state-of-the-art approaches ([11] [12] [21]), where configurations involve using the entire patient for model training, reserving only one time point for testing. Such practices introduce a significant risk of overfitting due to the model's intrinsic knowledge of the tested patient.

5 distinct folds has been choosen, each with 3 patients in the training set, 1 patient in the validation set and 1 in the test set. The numerical values in Table 1 represent the specific patient along with all their corresponding time point. By changing the patient used for the evaluation in each fold of the cross validation, LOSO-CV provided a subject-wise estimate of the performance for new patients.

4.4 Dataset and preprocessing

The dataset employed is a subset of the ISBI 2015 challenge dataset, which was publicly presented at the Longitudinal MS Lesion Segmentation Challenge [22], organized in conjunction with the ISBI 2015 conference. While the full dataset comprises MRI scans from 19 patients acquired at multiple time points on a 3.0 Tesla MR scanner,



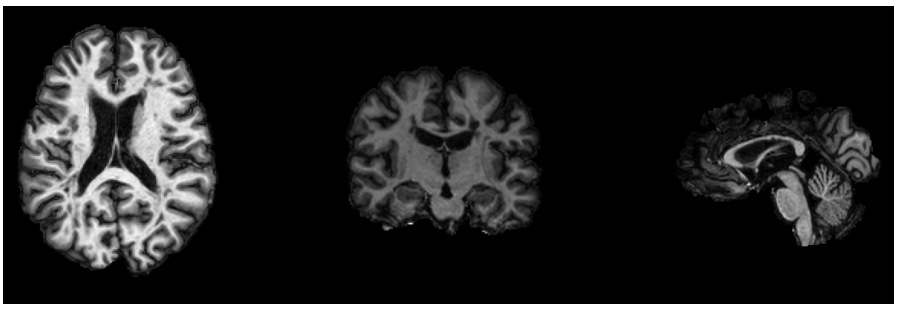
(a) Mask made by Rater 1 (b) Mask made by Rater 2

Fig. 5 The image shows the same FLAIR slices of a patient belonging to the ISBI 2015 dataset, labeled by two different experts, 5(a) Rater 1 and 5(b) Rater 2. It's evident that the two generated masks (overlaid and highlighted in yellow) are different in many pixels, highlighting the difficulty of manually creating masks that reliably delineate MS lesions.

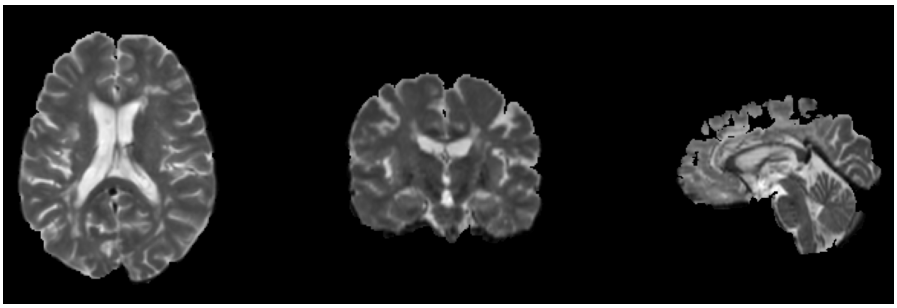
only 5 patients have corresponding segmentation masks. Each patient's MRI scans were annotated by two expert human raters, denominated Rater 1 and Rater 2, resulting in two different segmentation masks per patient. It's noteworthy that there are discrepancies between these masks, indicative of the complexities even for MS experts in accurately delineating MS lesions. As illustrated in Figure 5, the differences in mask annotations can be visually observed.

Among the 5 annotated patients in the dataset, four underwent longitudinal scans at four time points, while one patient had scans at five time points, totaling 21 MRI acquisitions. The time interval between consecutive acquisition time points was approximately one year. It's important to consider the highly variable nature of multiple sclerosis progression; follow-up scans may not necessarily correspond to disease progression, as lesions can appear at different times and in different brain regions. Each acquisition includes original MR images, as well as images after rigid registration to a 1mm isotropic MNI template, brain extraction, and non-uniformity correction. The MRI sequences consist of T1-w, T2-w, PD-w, and FLAIR images. To assess the stability of the model, experiments using masks labeled by Rater 1 has been conducted. Moreover, only FLAIR, T1-w and T2-w images, as MS lesion are more visible in these sequences. Each sequence has the spatial dimension of $181 \times 181 \times 217$. Figure 6 illustrates an example of a patient from the ISBI dataset.

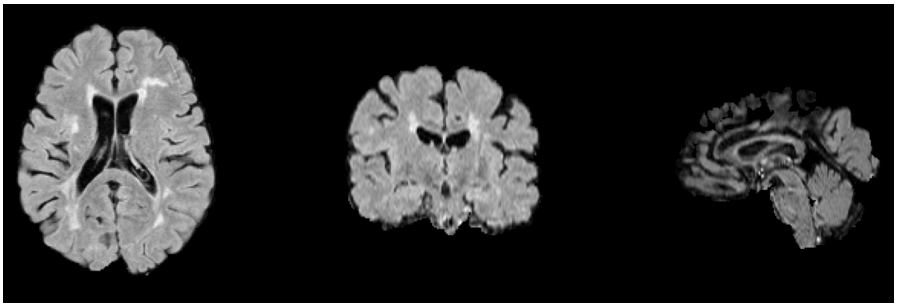
In our preprocessing phase, we employed transformations available in MONAI, encompassing foreground cropping, padding, and intensity normalization. These transformation assist the training and validation phases of the model, considering that pixels corresponding to lesions are often small compared to the entire image. Additionally, cropping and padding transformations help reduce the number of black pixels in the image, as they constitute the majority compared to the white pixels identifying lesions in the ground-truth mask. Intensity normalization is an essential preprocessing step, particularly considering that MRIs may be acquired from different patients or the same patient at different times using different scanners or parameters. This variability can lead to significant intensity variations among MRI modalities, and normalization



(a) Axial, coronal and sagittal view of the central slice extracted of a T1-w



(b) Axial, coronal and sagittal view of the central slice extracted of a T2-w



(c) Axial, coronal and sagittal view of the central slice extracted of a FLAIR

Fig. 6 Central slice of a MRI scan in three different modalities: 6(a) T1-w, 6(b) T2-w, and 6(c) FLAIR, showing axial, coronal, and sagittal views.

is therefore necessary to ensure greater consistency in pixel intensities, thereby facilitating model learning.In addition to the aforementioned preprocessing steps, we also employed data augmentation, which was crucial given the limited number of training samples available in our dataset. Data augmentation techniques were applied to artificially increase the diversity of the training data, helping to prevent overfitting and improve the generalization capability of the model and important when dealing with medical imaging datasets where the number of annotated samples may be limited.

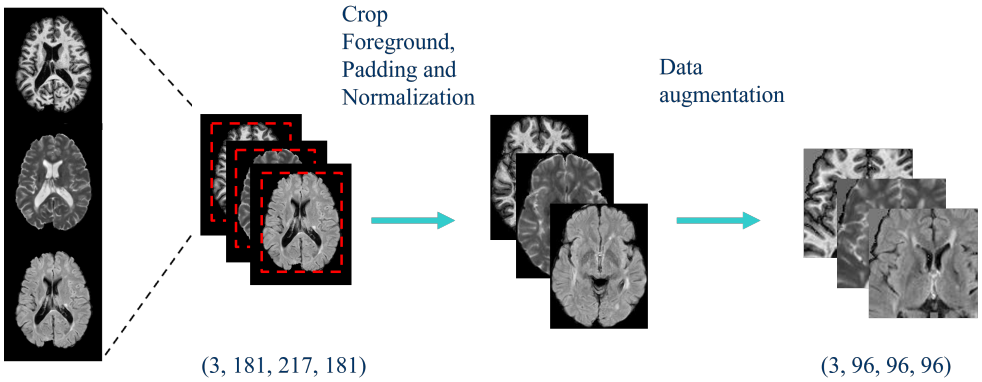


Fig. 7 Overview of the preprocessing step. The three volumes, corresponding to the three MRI modalities T1-w, T2-w and FLAIR are concatenated channel-wise, after which transformations are applied and data augmentation is performed.

We applied random cropping to $96 \times 96 \times 96$, flips, scaling, and random intensity adjustments. These adjustments can enhance the robustness of our model and improve its performance on unseen data. Figure 7 shows the preprocessing step performed.

5 Comparative studies

We explore different configurations of the MSSegDiff architecture, depending on which backbone architectures are used as the Encoder Module and Denoising Module. We chose to use network architectures based on the U-Net model in all configurations due to their proven capability to produce accurate segmentation maps, even with limited input data. This is particularly important in medical imaging, where access to large datasets is often restricted.

State-of-the-art medical image segmentation often incorporates custom modifications to basic architectures to enhance performance. Each configuration is trained for 1200 epochs. Following the completion of training, we employ the validation set to select the model obtained at the epoch with the highest validation Dice Score. For our studies we use the appropriately customized implementation of architectures available in MONAI. All the trained configurations will be explained below.

5.1 Segmentation models

The experiments were carried out on ISBI2015 dataset using three representative and recent network structure in medical image segmentation, namely BasicUNet [23], SegResNet [24] and SwinUNETR [25]. In the following, a briefly description of the network architectures employed; all the implementations of these networks are available in MONAI framework.

BasicUNet

The BasicUNet [23] is a convolutional network architecture designed for biomedical image segmentation. It consists of a symmetric encoder-decoder structure with skip connections that directly link corresponding layers in the encoder and decoder. The encoder, or contracting path, captures context through successive convolutional and downsampling layers, while the decoder, or expansive path, reconstructs the image using upsampling operations. Skip connections ensure that spatial information lost during downsampling is recovered, facilitating precise localization. This architecture allows for effective segmentation with relatively small amounts of data, making it suitable for medical imaging tasks.

SegResNet

SegResNet is an architecture specifically designed for medical image segmentation. This architecture was originally proposed in [24] for brain tumor segmentation from 3D MRIs. It is based on an encoder-decoder CNN architecture, featuring a large encoder to extract high-level image features and a small decoder to reconstruct the segmentation mask. The encoder is based on Residual Network (ResNet) blocks [26], while the decoder is similar to the encoder, except that it uses a unique block for each spatial level. The original implementation in [24] employs a variational autoencoder (VAE) [27] to reconstruct the original input image. In this work, we use the MONAI implementation without the VAE part.

SwinUNETR

SwinUNETR [25] is an advanced network architecture that combines the strengths of the Swin Transformer and UNETR [28] models, specifically designed for 3D brain tumor semantic segmentation. The Swin Transformer is utilized as the encoder, leveraging hierarchical feature extraction with a shifted windows mechanism. This allows for efficient and scalable computation of global self-attention across the image. This architecture excels at capturing both local and global context, making it particularly effective for complex segmentation tasks. The output of the encoder is connected to a CNN-based architecture used as the decoder at different resolutions through skip connections.

5.2 Performance of different configuration models

Table 2 shows the segmentation performances of the different configuration models tested. We began our study by evaluating the performance of the BasicUNet architecture. This first configuration follows the implementation proposed in [5], for this also attention mechanisms have been added in this first test.

This configuration, named **MSSegDiff** in Table 2, we use the encoder part of the BasicUNet as the EM to extract high-level features from the input volume I. As for the DM, we adopted the whole BasicUNet architecture, enhanced with attention mechanisms. The features extracted by the EM were then concatenated with the

Table 2 Results obtained from the comparative studies performed considering different model configurations. The results report the average values of evaluation metrics among all folds.

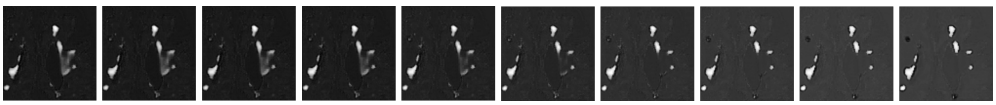
	Mean on 5 Fold						
Model	**DSC↑**	**TPR↑**	**PPV↑**	**LTPR↑**	**LFPR↓**	**AVD↓**	**ASSD↓**
MSSegDiff	**0,7526**	**0,7617**	**0,7700**	0,6652	0,2782	**0,2251**	**0,7699**
MSSegDiff+EncoderSegResNet	0,7167	0,7103	0,7656	0,6467	**0,2698**	0,2765	1,1371
MSSegDiff+SegResNet	0,7140	0,7485	0,7602	0,6237	0,2705	0,3800	1,1888
MSSegDiff+SwinUNETR	0,7093	0,7364	0,7345	**0,6654**	0,3261	0,3409	0,9624
MSSegDiff+MultiEncoder	0,7024	0,7598	0,6890	0,6244	0,3275	0,4096	1,2689

output features from each downsampling block in the denoising U-Net. Squeeze-and-attention (SA) [29] layers were introduced after each block, in both upsampling and downsampling paths, to refine feature selection and enhance the network's capability to focus on relevant information. We customized the layers proposed in the original paper to work with 3D images. As demonstrated by the obtained results in Table 2, this first configuration achieves high values, showcasing the potential of this attention-enhanced network architecture when integrated into MSSegDiff.

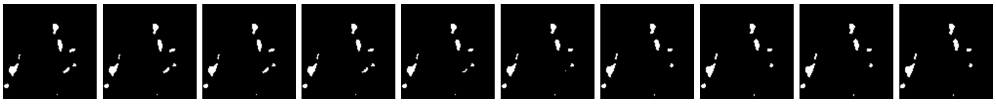
The second configuration, named **MSSegDiff+EncoderSegResNet**, involves a change in the encoder (EM). Specifically, we used the encoder section of SegResNet as EM, while keeping the DM based on BasicUNet enhanced with SA layers, unchanged. The goal was to assess the performance of SegResNet as an encoder for extracting relevant features from volumetric MRI images, while leveraging the capability of BasicUNet+SA to generate accurate segmentation masks through the denoising process. The experimental results indicated that integrating SegResNet as the EM did not lead to a significant improvement compared to the first configuration, which uses the BasicUNet for both the Encoder and Denoising Modules. Specifically, the evaluation metrics did not show substantial increases, suggesting that adopting SegResNet as the encoder does not offer considerable advantages over the baseline BasicUNet enhanced with attention mechanisms.

Given these findings, we tested the configuration **MSSegDiff+SegResNet**, where we employ SegResNet throughout the entire architecture, both in the EM and the DM. The rationale behind this approach is to determine whether utilizing SegResNet for the entire pipeline might enhance overall performance. This approach seeks to maximize the utilization of SegResNet's strengths in feature extraction and representation, which may not have been fully harnessed when it was employed solely as an encoder. However, the outcomes from this configuration, which are practically identical to those of the second setup, indicate that our initial expectations were unfounded. The results obtained highlight how the use of SegResNet did not confer advantages over BasicUNet when integrated into a pipeline based on the diffusion model. This finding underscores the importance of carefully considering the suitability of different architectures within specific frameworks.

In the fourth configuration, named **MSSegDiff+SwinUNETR**, we thought of evaluating the performance of SwinUNETR. This decision stems from transformer-based model recent advancements in medical image segmentation tasks, particularly with Vision Transformers (ViTs), due to the effectiveness of their self-attention in capturing long-range dependencies while maintaining computational efficiency. Given



(a) All model outputs



(b) All prediction samples

Fig. 8 Image shows 8(a) model outputs and 8(b) the prediction for x_0 through $T = 10$ steps.

the intricate nature of MS lesion segmentation from MRI scans, SwinUNETR's ability to capture both local and global contextual information may prove beneficial. Thus, exploring its performance within the MSSegDiff pipeline could provide valuable insights into its suitability for this task. However, from the results obtained, it can be concluded that this configuration performed worse than all previous configurations. One possible explanation for this outcome could be attributed to the inherent limitations of utilizing transformers with limited data. Considering the superior performance observed in the first configuration utilizing the BasicUNet architecture, we decided to conduct a final experiment. The idea of this experiment is to improve our approach by refining the EM. This refinement involved splitting the encoder, originally responsible for processing the concatenated MRI modalities into three distinct encoders, each dedicated to a specific modality. Subsequently, we aggregated the outputs derived from the three encoders before integrating them with the output of the encoder from the Denoising Module. Despite our efforts, this adjustment, termed **MSSegDiff+MultiEncoder**, did not improve performance. In fact, it yielded the lowest Dice score, as evident in Table 2.

5.3 Performance of different inference methods

At test time, the diffusion model iterates through T steps using the Denoising Diffusion Implicit Models (DDIM) method [16]. Each step generates an increasingly refined segmentation mask. As described in Section 3.4 DDIM is able to produce refined masks over iterations, so we decided to use $T = 10$ to test the generalization capabilities of the model. Figure 8 shows an example of prediction masks obtained from each iteration of DDIM with $T = 10$ step. Leveraging the insight that with an increasing number of testing steps, the prediction becomes progressively more accurate and the prediction uncertainty decreases, we decided to assess various inference methods. This methods leverage the fusion of segmentation masks obtained from each iteration. Merging the predictions generated at each iteration can ensure more robust segmentation results compared to using only the final segmentation, as done in traditional generative tasks.

As the first method, we decided to calculate the mean of the segmentation masks obtained at each step to generate the final segmentation mask. Additionally, we calculated the variance of the output predictions. Figure 9(a) shows an example of the mean mask obtained, while Figure 9(b) illustrates the variance. From the variance

(a) Mean (b) Variance (c) Mean-Variance

Fig. 9 Example of segmentation masks obtained by 9(a) the mean of the masks obtained from each iteration, 9(b) their variance, and 9(c) the difference between the mean and the variance.

image, it is noticeable that the pixels most challenging to predict are those delineating the boundaries of MS lesions. As an additional inference method, we decided to compute the difference between the mean and the variance. An example of the result is shown in Figure 9(c). This stems from the insight that is particularly challenging for our models to accurately predict the boundaries of the lesions. These boundaries are already difficult to identify due to the small and complex structure of the lesions, making precise segmentation a significant challenge for the network. To further enhance the fusion process, we decided to use two additional methods. The first method, originally proposed in [30] and utilized in [5], introduces a Step-Uncertainty based Fusion (SUF) module to fuse the segmentation masks based on the number of steps and the prediction uncertainty. The uncertainty is estimated by performing multiple forward passes through the diffusion model. Each step produces different outputs, generated from different random noise, which are then used to calculate the uncertainty map. The second method we tested is the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm, proposed in [31], a widely used method in medical imaging also employed in [13] [15]. STAPLE is a voting algorithm that creates a final segmentation mask by performing pixel-by-pixel voting of all the segmentation masks. It has been demonstrated that the accuracy of STAPLE improves with an increasing number of segmentation masks considered for voting. In Table 3 we report the results of the different inference methods in terms of the mean Dice score, evaluated over all folds (best results are denoted as bold). From the results, it can be seen that the inference methods Mean, Mean-var, and SUF achieve similar results, showing consistent performance across different network configurations. In contrast, the STAPLE algorithm exhibits poor performance across all network configurations. A possible reason for this could be the low number of predictions used for the voting process. A higher number of predictions might lead to better results. Additionally, it is evident that the MSSegDiff configuration provides the highest Dice scores across all inference methods, confirming it as the best configuration. Specifically, the SUF inference method achieves the highest Dice score in this configuration. However, the SUF method does not perform better in all configurations, while Mean-var method achieves a Dice score of 0.7213, the highest across all network configurations, slightly surpassing the SUF method. Given the fact that the best network configuration MSSegDiff obtains the highest Dice scores and achieves the best performance using SUF as the inference method, we decided to use the SUF method and report the results in Table 2 using this inference method.

Table 3 Results obtained from methods evaluated by mean Dice Score among all folds.

	Mean DSC on 5 Fold			
Model	**Mean**	**Mean-var**	**SUF**	**STAPLE**
MSSegDiff	**0,7517**	**0,7495**	**0,7526**	**0,6690**
MSSegDiff+EncoderSegResNet	0,7133	0,7246	0,7167	0,6381
MSSegDiff+SegResNet	0,7149	0,7169	0,7140	0,5655
MSSegDiff+SwinUNETR	0,7073	0,7108	0,7093	0,6246
MSSegDiff+MultiEncoder	0,7023	0,7049	0,7025	0,5511
Mean	0,7179	**0,7213**	0,7190	0,6097

6 Conclusion

In this study, we explored the efficacy of various deep learning architectures integrated into a diffusion model pipeline for MS lesion segmentation from longitudinal MRI scans. Through an extensive set of tests, we evaluated the performances of different network configurations within the MSSegDiff framework. Our findings demonstrate that while incorporating advanced architectures holds promise, the traditional U-Net-based configuration consistently outperforms other models. The MSSegDiff configuration leverages attention mechanisms to enhance segmentation accuracy, confirming its robustness in MS lesion segmentation. Moreover, our exploration of different inference methods revealed that the SUF approach consistently produces high-quality segmentation results, indicating its effectiveness in handling prediction uncertainty. Overall, our study contributes to advancing the understanding of deep learning architectures in MS lesion segmentation and the MSSegDiff framework provides valuable insights for future research directions. Future work could focus on optimizing the MSSegDiff framework to improve its computational efficiency, making it more suitable for real-time applications in clinical settings. Additionally, extending the approach to incorporate multi-site and heterogeneous data would enhance the model's robustness across various patient populations and imaging devices, paving the way for broader clinical adoption.

Declarations

Ethics approval and consent to participate. Not applicable

Consent for publication. Not applicable

Availability of data and materials. The dataset used in this work comes from the Longitudinal MS Lesion Segmentation Challenge held at the 2015 International Symposium on Biomedical Imaging in New York, NY, April 16-19 and is directly available for download at https://smart-stats-tools.org/lesion-challenge.

Competing interests. The authors declare that they have no competing interests.

Funding. FG is funded by the PNRR MUR project PE0000013-FAIR. FP is funded through the Programma di ricerca CN00000013 "National Centre for HPC, Big Data and Quantum Computing", finanziato dal Decreto Direttoriale di concessione del

finanziamento n.1031 del 17.06.2022 sulle risorse del PNRR–M4C2—Investimento 1.4— Avviso "Centri Nazionali"—D.D. n. 3138 del 16 dicembre 2021.

Authors' contributions. AR and FR initiated the study, developed the code, and conducted the experiments. AR, FR, AO and SB wrote the manuscript. EC, GR and FP validated the data and results. All authors read and approved the final manuscript.

Acknowledgements. Alessia Rondinella is a PhD candidate in the National PhD in Artificial Intelligence, XXXVII cycle, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma. Francesco Guarnera is funded by the PNRR MUR project PE0000013-FAIR. Francesco Pappalardo is funded through the Programma di ricerca CN00000013 "National Centre for HPC, Big Data and Quantum Computing", finanziato dal Decreto Direttoriale di concessione del finanziamento n.1031 del 17.06.2022 a valere sulle risorse del PNRR–M4C2—Investimento 1.4— Avviso "Centri Nazionali"—D.D. n. 3138 del 16 dicembre 2021.

References

[1] Lassmann, H.: Multiple sclerosis pathology. Cold Spring Harbor perspectives in medicine **8**(3), 028936 (2018)

[2] Coles, A.: Alastair compston, alasdair coles. Lancet **372**, 1502–1517 (2008)

[3] Janardhan, V., Suri, S., Bakshi, R.: Multiple sclerosis: hyperintense lesions in the brain on nonenhanced t1-weighted mr images evidenced as areas of t1 shortening. Radiology **244**(3), 823–831 (2007)

[4] Chong, A., Chandra, R.V., Chuah, K., Roberts, E., Stuckey, S.: Proton density mri increases detection of cervical spinal cord multiple sclerosis lesions compared with t2-weighted fast spin-echo. American Journal of Neuroradiology **37**(1), 180–184 (2016)

[5] Rondinella, A., Guarnera, F., Giudice, O., Ortis, A., Russo, G., Crispino, E., Pappalardo, F., Battiato, S.: Enhancing multiple sclerosis lesion segmentation in multimodal mri scans with diffusion models. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 3733–3740 (2023). https://doi.org/10.1109/BIBM58861.2023.10385334

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[7] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

[8] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205–218 (2022). Springer

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[10] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)

[11] Rondinella, A., Crispino, E., Guarnera, F., Giudice, O., Ortis, A., Russo, G., Di Lorenzo, C., Maimone, D., Pappalardo, F., Battiato, S.: Boosting multiple sclerosis lesion segmentation through attention mechanism. Computers in Biology and Medicine **161**, 107021 (2023)

[12] Hashemi, M., Akhbari, M., Jutten, C.: 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)

[13] Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., Liu, H., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. In: Medical Imaging with Deep Learning, pp. 1623–1639 (2024). PMLR

[14] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

[15] Wu, J., Ji, W., Fu, H., Xu, M., Jin, Y., Xu, Y.: Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 6030–6038 (2024)

[16] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

[17] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

[18] Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)

[19] Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)

[20] Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)

[21] Raab, F., Wein, S., Greenlee, M., Malloni, W., Lang, E.: A multimodal 2d convolutional neural network for multiple sclerosis lesion detection. Authorea Preprints (2023)

[22] Carass, A., Roy, S., Jog, A., Cuzzocreo, J.L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C.H., *et al.*: Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. NeuroImage **148**, 77–102 (2017)

[23] Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., *et al.*: U-net: deep learning for cell counting, detection, and morphometry. Nature methods **16**(1), 67–70 (2019)

[24] Myronenko, A.: 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. Revised Selected Papers, Part II 4, pp. 311–320 (2019). Springer

[25] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop, pp. 272–284 (2021). Springer

[26] He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, October 11–14, 2016, Proceedings, Part IV 14, pp. 630–645 (2016). Springer

[27] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

[28] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)

[29] Zhong, Z., Lin, Z.Q., Bidart, R., Hu, X., Daya, I.B., Li, Z., Zheng, W.-S., Li, J., Wong, A.: Squeeze-and-attention networks for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13065–13074 (2020)

[30] Xing, Z., Wan, L., Fu, H., Yang, G., Zhu, L.: Diff-unet: A diffusion embedded network for volumetric segmentation. arXiv preprint arXiv:2303.10326 (2023)

[31] Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging **23**(7), 903–921 (2004)