Enhancing Multiple Sclerosis Lesion Segmentation in Multimodal MRI Scans with Diffusion Models

Alessia Rondinella
Dept. of Math and Computer Science
University of Catania
Catania, Italy
alessia.rondinella@phd.unict.it

Francesco Guarnera
Dept. of Math and Computer Science
University of Catania
Catania, Italy
francesco.guarnera@unict.it

Oliver Giudice
Dept. of Math and Computer Science
University of Catania
Catania, Italy
giudice@dmi.unict.it

Alessandro Ortis
Dept. of Math and Computer Science
University of Catania
Catania, Italy
alessandro.ortis@unict.it

Giulia Russo
Dept. of Drug and Health Sciences
University of Catania
Catania, Italy
giulia.russo@unict.it

Elena Crispino
Dept. of Biomedical and Biotechnological Sciences
University of Catania
Catania, Italy
elena.crispino@phd.unict.it

Francesco Pappalardo
Dept. of Drug and Health Sciences
University of Catania
Catania, Italy
francesco.pappalardo@unict.it

Sebastiano Battiato
Dept. of Math and Computer Science
University of Catania
Catania, Italy
sebastiano.battiato@unict.it

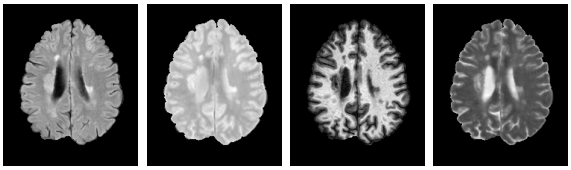
2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385334

***Abstract*—Accurate segmentation of Multiple Sclerosis (MS) lesions from Magnetic Resonance Imaging (MRI) scans is crucial for clinical diagnosis and effective treatment planning. In this work, we investigate the effectiveness of Diffusion Models (DM) in achieving pixel-wise segmentation of MS lesions. DM significantly improves segmentation sensitivity, especially in regions with subtle abnormalities. We conducted extensive experiments using the magnetic resonance volumes from a public dataset, encompassing various imaging modalities. Our analysis demonstrated how DM can achieve performance levels that are on par with state-of-the-art techniques, as evidenced by a mean Dice coefficient comparable to the best existing methods. Furthermore, some variants of standard DM exhibits robustness across various imaging modalities, showcasing its versatility in clinical settings.**

***Index Terms*—Multiple Sclerosis, Denoising Diffusion Models, Lesion segmentation, Medical image analysis**

I. Introduction

Multiple Sclerosis (MS) is a complex and debilitating chronic inflammatory demyelinating disease of the Central Nervous System (CNS) [1]. It is characterized by focal areas of inflammation accompanied by myelin and axonal loss, leading to a wide array of neurological symptoms and disabilities. Accurate detection and localization of MS lesions play a crucial role in clinical assessment and treatment planning. These lesions, observable via Magnetic Resonance Imaging (MRI), manifest in various regions of the brain and spinal cord, and their spatial distribution is indicative of the disease's progression and severity [2]. Differentiating lesions based on their specific locations, such as periventricular, cortical/iuxtacortical, brain stem/cerebellar, and spinal cord, holds importance in both diagnosis and monitoring of disease progression and therapeutic efficacy [3]. Manual annotation of MS lesions on MRI scans, while essential, is a resource-intensive endeavor requiring specialized expertise. Moreover, the inherent subjectivity introduces inter- and intra-operator variability, potentially impacting the accuracy and reproducibility of lesion segmentation [4]. This necessitates the development of automated tools to mitigate human-induced biases and ensure consistent and reliable clinical evaluations. The longitudinal brain MRI protocol encompasses a spectrum of sequences, each offering unique contrasts for delineating brain tissues. Notably, Fluid Attenuated Inversion Recovery (FLAIR), T1-weighted, T2-weighted, and PD-weighted images have become crucial in detecting MS lesions. Among these, FLAIR images stand out, providing a distinct and high-contrast view of lesions, enabling their clear demarcation from surrounding tissues. (Refer to Fig. 1 for illustrative examples.)



(a) FLAIR (a) PD-w (a) T1-w (a) T2-w

Fig. 1. Examples of axial brain MRI images in different modality of acquisition showing MS lesions: (a) FLAIR, (b) PD-weighted, (c) T1-weighted and (d) T2-weighted.

In this research, we explore the utilization of Diffusion Models (DM) in conjunction with various architectural en-

hancements to achieve more accurate segmentation of Multiple Sclerosis (MS) lesions. Our study illustrates how DM, when combined with these architectural variants, significantly improves sensitivity, especially in regions containing subtle abnormalities, thus elevating the overall accuracy of the segmentation algorithm. Extensive experiments on the magnetic resonance volumes from the ISBI2015 dataset [5] demonstrate the good performance of our method, achieving an important mean Dice Score (DSC) on the test set. Our tests leverages the power of DM for robust 3D medical image segmentation.

DMs represent a novel frontier in Deep Learning methodologies, introducing a unique approach that involves injecting controlled noise at the input and iteratively refining the segmentation label map, enhancing prediction stability. Building upon the architecture described in [18], the framework was extended in order to create a more powerful and tailored solutions. To harness the full capabilities of the DM, we conducted extensive testing on various 3D medical image segmentation algorithms based on them, incorporating modifications and additional components. Those architectures are equipped to handle medical imaging data effectively. In the best variant, in order to extract meaningful information from the input volume, a Denoising Attention U-Net Module and a standalone Encoder Module were added, which work jointly to learn and implement the denoising process. The output is a refined segmentation label map, free from noise artifacts.

The efficacy of the architectures were evaluated on ISBI2015 dataset, employing a leave-one-subject-out-cross-validation scheme in patients with lesions in both baseline and follow-up scans. Among the various models analyzed in the general results (Section IV-B), the architecture depicted in Fig. 3 (MS-SegDiff) emerged as the most promising, exhibiting superior performance.

The remainder of this paper is organized as follows. Section II provides an overview of the current state-of-the-art in medical image segmentation and highlights the use of Transformer and Denoising Diffusion Models in this context, while Section III delves into the employed dataset and details the proposed architecture. Experimental results and ablation studies are reported in Section IV, whereas Section V presents the concluding remarks of the paper.

TABLE I

CONFIGURATIONS USED TO PERFORM THE EXPERIMENT. A LEAVE-ONE-SUBJECT-OUT-CROSS-VALIDATION STRATEGY WAS IMPLEMENTED CONSIDERING 20 DIFFERENT FOLD, WITH 3 PATIENTS TO TRAIN THE NETWORK, 1 USED AT VALIDATION SET AND 1 SCAN TO TEST IT. NOTE THAT THE NUMBERS INDICATE THE PATIENT WITH ALL ITS TIME-POINT SCANS

Fold	Training	Validation	Test
Fold1	1, 2, 3	4	5
Fold2	1, 2, 4	3	5
Fold3	1, 3, 4	2	5
Fold4	2, 3, 4	1	5
Fold5	1, 2, 3	5	4
Fold6	1, 2, 5	3	4
Fold7	1, 3, 5	2	4
Fold8	2, 3, 5	1	4
Fold9	1, 2, 4	5	3
Fold10	1, 2, 5	4	3
Fold11	1, 4, 5	2	3
Fold12	2, 4, 5	1	3
Fold13	1, 3, 4	5	2
Fold14	1, 3, 5	4	2
Fold15	1, 4, 5	3	2
Fold16	3, 4, 5	1	2
Fold17	2, 3, 4	5	1
Fold18	2, 3, 5	4	1
Fold19	2, 4, 5	3	1
Fold20	3, 4, 5	2	1

TABLE II

TABLE DISPLAYING THE FIVE DATASET CONFIGURATIONS CHOSEN FOR CONDUCTING ALL TESTS. 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
Fold6	1, 2, 5	3	4
Fold11	1, 4, 5	2	3
Fold16	3, 4, 5	1	2
Fold17	2, 3, 4	5	1

II. STATE OF THE ART

A. Medical image segmentation

In recent years, there has been a notable surge in the utilization of deep learning techniques in medical image, spanning tasks such as classification [6] and segmentation. In the state-of-the-arts, there are various works dedicated to segmenting images from modalities like MRI and Computed Tomography (CT). These works vary in terms of the deep learning architectures employed and the methodologies adopted. Recently, [7] wrote a comprehensive review offering an insightful overview of deep learning techniques applied in MRI-based research, also identifying promising avenues for future development. Also [8] wrote a review in which summarised scientific articles that perform the detection and segmentation of MS lesions through deep learning. Numerous studies have harnessed the power of Convolutional Neural Networks (CNN) for precise MS segmentation. For instance, [9] recently propose a patch-wise CNN to extract brain lesion from MRI; authors in [19] introduce a CNN employing a dual-path architecture. This model incorporates an attention-driven interaction block, facilitating the exchange of information between two distinct time points. In [34] [35], authors employ a neural network-based automated approach to accurately identify MS lesions in 3D brain MRI scans.

B. U-Net-based Architecture

Across medical images, conventional volumetric segmentation algorithms are often relied on a U-shaped architecture, integrating encoder-decoder frameworks with skip connections that enable the decoder to reconstruct features derived from the encoder. The majority of segmentation algorithms for volumet-

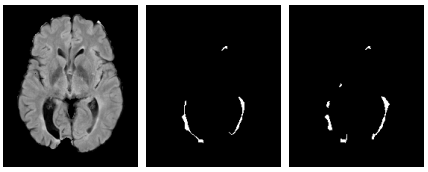
ric medical images have adopted these structure, achieving promising results in this field. Furthermore, the integration of attention mechanism has been shown that significantly improves the performance of basic architecture. Specifically, authors in [10] introduces a U-Net variant that incorporates global attention to perform segmentation; Additionally, in [11] a Fully Convolutional Densely Network is presented, featuring attention blocks that enable lesion segmentation applied to 2D images. In a recent study [20] conducted a segmentation on FLAIR and T2 MRI images, exploiting two different U-Net-based network architectures. In these recent studies, [20] [21] the authors perform MRI image segmentation with U-Net-based architectures integrating attention gate and channel attention techniques.

C. Transformer-based Architecture

Recent advancements in the field of deep learning have witnessed the emergence of the transformer architecture [12], originally designed for natural language processing tasks but also adapted for images with Visual Transformer (ViT) [13]. For instance, studies have successfully applied transformer-based models for medical image segmentation. [14] employ a 3D CNN that model local and global feature for 3D multimodal brain tumor segmentation; In [15] authors propose BiTr-UNet, that consist of an attention module that refine UNet feature also with two ViT layers to segment tumors in BraTS 2021 [16]. Authors in [17] present SwinUNETR, a novel architecture designed for the segmentation of brain tumours. This approach frames the task as a sequence-to-sequence prediction problem, with an encoder acting as a multi-resolution feature extractor connected to an FCNN decoder, which finally generates the segmentation output.

D. Denoising Diffusion Model-based Architecture

The integration of Diffusion Models in medical image segmentation represents a cutting-edge advancement in the field. These models, initially designed for generative tasks, have found substantial success in enhancing segmentation accuracy and precision, including in medical imaging. For instance, MedSegDiff [22] has demonstrated notable proficiency in 2D medical image segmentation by incorporating a segmentation Denoising UNet and inter-structural information via Fourier transform. Additionally, the same authors have proposed an advanced version, MedSegDiff-V2 [23], a transformer-based conditional UNet framework a transformer-based conditional UNet framework that exploits conditioning techniques to improve segmentation, leads a significantly enhancing the performance of MedSegDiff. Additionally, [24] have leveraged Diffusion Models to refine 2D medical image segmentation results, emphasizing the importance of fusing outputs from each diffusion step for improved robustness. Furthermore, [25] propose BerDiff, a model that uses Bernoulli noise to produce a series of segmentation masks, which can help highlight regions of interest in order to improve binary segmentation tasks. Inspired by the recent success of Denoising Diffusion Model, we designed a segmentation model for MS segmentation. To the best of our knowledge, this is the first work which use DM for MS lesions segmentation.



(a) FLAIR (a) Rater 1 (a) Rater 2

Fig. 2. Examples of FLAIR image with the corresponding masks annotated by Rater 1 (a) and Rater 2 (b), respectively. To note that, different Raters have different decision about pixel location in the mask.

III. Methodology

Diffusion Models are a class of generative models designed to understand and simulate a diffusion process for generating data that closely resembles the original data distribution. This process comprises two stages: the forward process and the reverse process. During the forward process, the model starts by introducing incremental noise, typically Gaussian noise, to the initial image. This is done incrementally, step by step, until the input image is gradually transformed into a progressively noisier representation. In the reverse process, inverse transformations are applied to recover the original input image. In this way, DM are able to generate new images that have a similarity to the original data distribution. Diffusion Models have gained prominence among deep learning architectures due to their stability during training, especially when compared to other generative architecture, considering the fact that DM adhere to likelihood-based training, ensuring a more robust and stable training process that effectively mitigates the risks associated with mode collapse. Additionally, DM exhibit a heightened resilience to overfitting. This resilience is important to be sure that the model generalizes the results to unseen data. The intrinsic qualities of DM make them particularly suitable for a wide range of tasks in the field of deep learning. These tasks include, but are not limited to, image generation, data completion, denoising, and image segmentation, among others. We conducted an analysis of the application of DM for the segmentation of MS lesions in brain MRI images. First, this choice is driven by the remarkable ability of DM to capture the underlying complexity of data distributions, resulting in high-quality images characterized by fine details and realistic textures. Second, it is worth emphasizing that our work represents the first attempt to employ DM for the specific segmentation of multiple sclerosis lesions.

A. Dataset and preprocessing

In order to assess the effectiveness of our proposed method in segmenting MS lesions, we employed the publicly accessible ISBI 2015 Longitudinal MS Lesion Segmentation Challenge dataset [5]. The dataset consists of 19 MRI scans from patients acquired over multiple time points using a 3.0 T MR scanner. However, only five patients have corresponding segmentation masks, each created by two expert human raters.

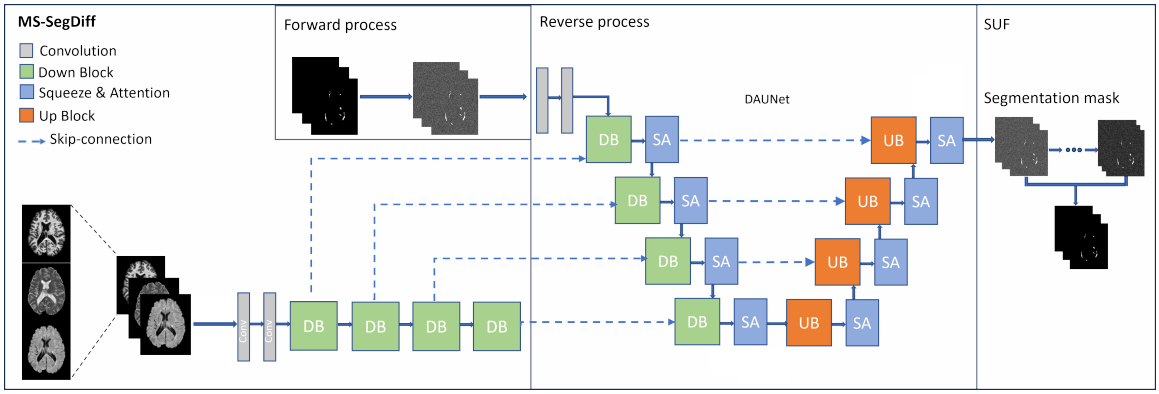


Fig. 3. Schematic representation of the MS-SegDiff architecture. This model takes as input the multimodal MRI volume and its noisy mask and returns the expected clean segmentation mask as output.

It's worth noting that there are discrepancies between these masks, highlighting the challenge even for MS experts, as shown in Fig. 2. Four patients underwent 4 time points, while one patient underwent 5 time points. The different acquisitions were performed one year apart. Each scan contains the original MRI, the MRI itself after a preliminary preprocessing step (this include co-registration, brain extraction and non-uniformity correction) and the masks annotated by two different raters. In addition, there are several MRI modalities: T1-weighted, T2-weighted, PD-weighted and FLAIR. Each scans have the size of $181 \times 217 \times 181$. In many of the methods mentioned in the state of the art, the ISBI2015 dataset is used for MS lesion segmentation [11] [20] [21] [34] [35]. In this work, only the preprocessed FLAIR, T1-weighted and T2-weighted images were used for train and test the architectures. In order to evaluate the stability of the model, the experiments were conducted using only masks annotated by the Rater 1. The experiments were conducted using the 20 configurations used in [34] [35] and listed in Table I: in particular a subset of them (listed in Table II) have been used to compare the variants of the model while in the best variant all the configurations have been trained and tested. In the preprocessing phase, MONAI [29] transformations were implemented, including foreground cropping, padding to $96 \times 96 \times 96$, and normalization of input intensity. Additionally, data augmentation techniques were exploited, encompassing random clipping, flips, scaling and random intensity adjustments.

B. Architecture

The objective of this work was the evaluation of performances related to the application of DM to MS lesion segmentation in MRI data. The denoising architecture, takes multimodal MRI data and the corresponding ground truth as input and adding random noise learns to gradually remove it with the aim of generating clear segmentation maps. Fig. 3 provides an overview of MS-SegDiff (the best DM variants tested) for MS lesion segmentation. The architecture started from [18] employing a BasicUNet [31] as a backbone, a widely adopted architecture for image segmentation. To enhance its performance to detect MS lesions some attention mechanisms were integrated. The forward process in this pipeline starts with the introduction of 't' steps of noise to the ground-truth masks. Following the approach outlined in [18], the input volume containing various modalities is fed through a first Encoder Module (EM) to extract the embedding representing the discriminative features. These features are subsequently combined channel-wise, with the noisy labels in the downsampling path of the Denoising Attention U-Net Module (DAUNet). This network perform the reverse process and produce the clear segmentation mask. A demonstrated improvement of the architecture is the addiction of Squeeze-and-Attention blocks [32] after each module of the U-Net. These components has been customised to work with 3D images. It introduce a pixel-group attention mechanism through a convolutional attention channel. This upgrading allows the network to selectively focus on the most relevant groups of pixels while ignoring extraneous information. This essential component is strategically positioned after each down block (DB), ensuring a comprehensive treatment of the input data. The multi-scale features output from each Squeeze-and-Attention block serve as input for the next down block. It is worth noting that both encoders keep the same number of features and the same size to allow for feature merging. Finally, the upsampling path is symmetical to the downsampling, featuring Squeeze and attention blocks after each upsampling block (UB).

Notably, the evaluation process generate different segmentation mask at each step of the network through the Denoising Diffusion Implicit Model (DDIM) approach [27]. Leveraging the insight that with an increasing number of testing steps, the prediction becomes progressively more accurate and the prediction uncertainty decreases, we merge the segmentation masks obtained from each iteration. We used the Step-Uncertainty based Fusion (SUF) module proposed in [18]. This module merge the segmentation masks obtained from each iteration, ensuring more stable segmentation results during testing.

C. Evaluation metrics

The model evaluation was done comparing the predicted segmentation masks with the provided ground-truth. As evalu-

TABLE III
THIS TABLE PRESENTS THE RESULTS OBTAINED FROM THE STUDIES PERFORMED CONSIDERING DIFFERENT MODEL CONFIGURATIONS, EVALUATED ON FIVE FOLDS.

Fold	Baseline			MS-SegDiff			Baseline-Encoder			Baseline-Encoder+attention		
	DSC	TPR	PPV	DSC	TPR	PPV	DSC	TPR	PPV	DSC	TPR	PPV
Fold1	0.7128	0.7789	0.6598	0.7200	0.7621	0.6891	0.7312	0.7552	0.7123	0.7099	0.7630	0.6674
Fold6	0.7123	0.8283	0.6304	0.7671	0.7908	0.7490	0.7510	0.8005	0.7089	0.7528	0.8306	0.6923
Fold11	0.6420	0.9411	0.4884	0.7498	0.8716	0.6590	0.6549	0.9379	0.5047	0.6997	0.8608	0.5914
Fold16	0.7606	0.6481	0.9216	0.8337	0.7603	0.9231	0.7159	0.5707	0.9614	0.7195	0.5777	0.9544
Fold17	0.7468	0.6896	0.8298	0.7233	0.6431	0.8479	0.7226	0.6816	0.7870	0.7475	0.6891	0.8299
Mean	**0.7149**	**0.7772**	**0.7060**	**0.7588**	**0.7656**	**0.7736**	**0.7151**	**0.7494**	**0.7349**	**0.7259**	**0.7442**	**0.7471**

TABLE IV
THIS TABLE PRESENTS THE RESULTS OBTAINED FROM THE OPTIMAL MODEL CONFIGURATION, MS-SEGDIFF, EVALUATED ACROSS ALL 20 FOLDS.

Fold	DSC		TPR		PPV	
	Best	Final	Best	Final	Best	Final
Fold1	0.7200	0.7180	0.7621	0.7436	0.6891	0.6999
Fold2	0.7520	0.7495	0.7371	0.7798	0.7721	0.7265
Fold3	0.5854	0.7277	0.9155	0.6863	0.4312	0.7796
Fold4	0.7011	0.7132	0.7294	0.6964	0.6782	0.7331
Fold5	0.7875	0.7850	0.7608	0.7537	0.8193	0.8229
Fold6	0.7671	0.7809	0.7908	0.7729	0.7490	0.7936
Fold7	0.6575	0.7883	0.8560	0.7385	0.5352	0.8484
Fold8	0.7629	0.7800	0.7709	0.7484	0.7574	0.8183
Fold9	0.8045	0.8045	0.8156	0.8145	0.7950	0.7958
Fold10	0.8002	0.7992	0.8066	0.8355	0.7952	0.7665
Fold11	0.7498	0.8089	0.8716	0.8262	0.6590	0.7934
Fold12	0.7419	0.7750	0.8640	0.8200	0.6507	0.7365
Fold13	0.8191	0.8131	0.7121	0.7007	0.9640	0.9688
Fold14	0.8076	0.7668	0.6986	0.6353	0.9581	0.9699
Fold15	0.8162	0.7678	0.7142	0.6365	0.9530	0.9700
Fold16	0.8337	0.6080	0.7603	0.4422	0.9231	0.9760
Fold17	0.7233	0.6969	0.6431	0.5830	0.8479	0.8855
Fold18	0.7675	0.7243	0.7248	0.6435	0.8247	0.8506
Fold19	0.7770	0.7552	0.7393	0.6875	0.8251	0.8473
Fold20	0.4574	0.4885	0.5699	0.3338	0.3831	0.9130
Mean	**0.7416**	**0.7425**	**0.7622**	**0.6939**	**0.7505**	**0.8348**

ation metrics, we include the Dice Score [30], the True Positive Rate (TPR) and the Positive Predictive Rate (PPV).

D. Implementation details

The proposed architecture has been implemented and tested using Pytorch [28] and MONAI [29] frameworks. Training was performed on a single NVIDIA A100 GPU. We employ a batch size of 2 and Adam optimizer with a base learning rate of $1e{-}4$ and a weight decay of $1e{-}5$. The learning rate (LR) was dynamically adjusted using a Cosine Annealing LR schedule [33]. Initially, the LR grows linearly, which was then gradually reduced after an initial warmup period, following a cosine function. All the network configuration was trained for 1200 iterations. In order to increase the number of training a patch $96\times96\times96$ have been chosen randomly in each epoch: in this way every batch contains different parts of same patient. Overlapping patches of 0.5 was also used during inference, to give intrinsic data augmentation and reduce the amount of memory required, while maintaining the detail of the original input image.

IV. EXPERIMENTS AND RESULTS

A. Experimental setup

To evaluate the performance of the proposed approach, a leave-one-subject-out-cross-validation on annotated subject was performed, following the scheme illustrated in Table II. The table depicts the patient distribution across different folds employed in cross-validation. Five distinct setups (Table II) were selected from the 20 configurations in Table I to guarantee that each validation and test set comprises a unique patient. Within each fold, there are three patients in the training set, one in the validation set, and one in the test set and the numerical values represent the specific subject along with its corresponding time-points. The choice of employing a leave-one-subject-out-cross-validation is significant, as it ensures a subject-wise subdivision of patients into training, validation, and test sets. This methodology guarantees that the entire

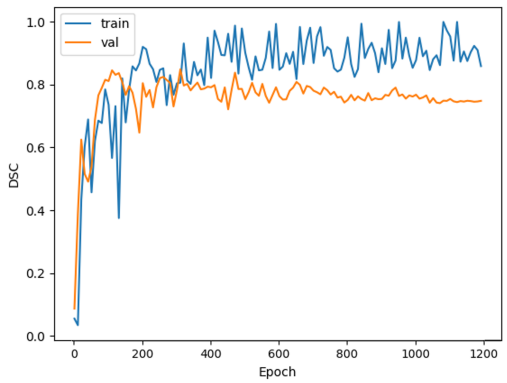


Fig. 4. Dice score produced after 1200 epoch of training by *Fold11*

patient, along with all their time-points, is included in each set. Consequently, the network is never exposed to the patient during training, nor to any of their previous time points. This approach stands in contrast to some works in the state-of-the-art ([11] [20] [34]), where configurations are tested involving the use of the entire patient for model training, reserving only one time point for testing. This introduces a substantial risk of overfitting due to the intrinsic knowledge of the tested patient. To note that the model was trained with only three MRI modalities, whereas many state-of-the-art architectures utilise all available modalities. The model was evaluated using the model coming from the 'best' iteration obtained during each training session, specifically referring to the iteration derived from the epoch with the highest Dice score in validation. To evaluate the performance of the overall results, we also reported the results obtained with the final model (after 1200 epochs).

B. Results

To clarify the rationale behind the selected architecture, a series of studies were conducted. The network structure comprises a baseline backbone with subsequent modifications and we will analyse the outcomes of its variants. Specifically, tests involved the removal or addition of specific components within the network to better understand their impact on both individual modules and the overall performance of the model. The aim is to quantitatively assess the impact of each component on the overall model. The results of these studies are shown in Table III and described below.

1) Baseline: Our analysis started with the baseline architecture of [18], evaluating performance on the ISBI2015 dataset. We trained the model using the five folds outlined in Table II. These configurations were meticulously chosen to have the most possible generalization in training occurrences. The results of the chosen evaluation metrics obtained from the five folds are shown as 'Baseline' in Table III. Our baseline model achieved a mean Dice score of 0.7149, demonstrating competitive performance compared to state-of-the-art methods that do not incorporate patient-specific information [34] [35].

2) Baseline with attention: Following this, marking a significant improvement in this work, we integrated the custom-designed squeeze and attention 3D blocks into both pathways of the Attention U-Net. This augmentation significantly bolstered the architecture responsible for executing the reverse process. We named this model 'MS-SegDiff' and its architectural representation is depicted in Fig. 3. From the results obtained in Table III (shown as 'MS-SegDiff'), it can be stated that these additions significantly augment the segmentation process of MS lesions, providing a notable enhancement in performance compared to the baseline approach. In comparison to state-of-the-art methods, our approach achieved a mean Dice score of 0.74 over the 20 configurations (0.7588 considering the five configurations in Table III). Furthermore, it's worth noting that with further refinements (some of which are discussed in subsequent paragraphs), Diffusion Models could potentially be the future of this task.

3) Impact of the encoder on network performance: Subsequently, we assessed the performance of the 'Baseline' and 'MS-SegDiff' models by excluding the encoder block and retraining it on the five folds, retaining only the forward and reverse processes. The results of this additional tests are reported in Table III as 'Baseline-Encoder' and 'Baseline-Encoder+Attention', respectively. Based on the conducted study and the outcomes presented in the table, it is evident that the encoder module only improves performance when combined with the denoising UNet modified with attention (DAUNeT). This test underline the importance of the encoder and the embedding it produces. Its capability to extract discriminative features from the whole volume is crucial for effective segmentation. This further supports the notion that the encoder module significantly contributes to the model's ability to discern and utilize relevant features for segmentation tasks.

4) MS-SegDiff results: We opted to train the 20 configurations described in Table I using the best-performing model identified in the ablation tests. This approach provides a comprehensive perspective by considering all possible combinations of patients across the various folds during the training phase. Table IV shows the results obtained in the 20 configurations. We present the outcomes achieved with both the optimal model selected during training, referred to as the 'Best,' and the final model obtained after the completion of training, denoted as the 'Final'. The results also validate the findings obtained in the 5 folds. It is notable that, in several cases, the results of the 'Final' model surpass those of the 'Best' model. This phenomenon is likely attributed to the limited heterogeneity in the dataset employed. When the validation set includes a patient clearly distinct from those in the training set, while the test patient is more similar, the 'Final' model tends to perform better.

This behavior is highlighted by the curves illustrating the Dice score progression in both training and validation in Fig 4. While the training curve continues to ascend, the validation curve reaches a peak, stabilizes, and even declines. This insight sheds light on the challenges inherent in evaluating methods with such a limited number of patients. Despite the employment of transformations and data augmentation

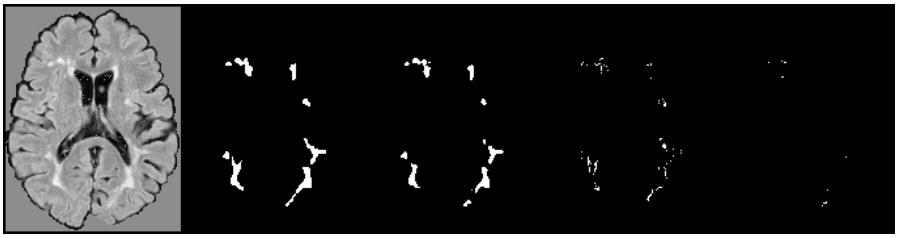


Fig. 5. Example of a segmentation mask obtained with our approach for fold: *Fold13*. The image shows: the original slice, the ground-truth mask, the prediction mask, false-positive pixels and false-negative pixels.

TABLE V
RESULTS OBTAINED FROM THE STUDIES PERFORMED CONSIDERING DIFFERENT CONFIGURATIONS OF S.

S	**4**	**5**	**6**
DSC	0.7391	0.7412	**0.7416**

TABLE VI
RESULTS OBTAINED FROM THE STUDIES PERFORMED CONSIDERING DIFFERENT PATCH SIZES.

Patch-Size	**DSC**	**TPR**	**PPV**
96x96x96	0.7149	0.7772	0.7060
64x64x64	**0.7388**	0.7063	0.8385

techniques, these models, including state-of-the-art ones, tend to exhibit overfitting to the patient data.

This observation further reinforces the efficacy of our previous approach, where we utilized all available patient scans for training the network, reserving only one time-point for testing. This strategy yielded superior results, highlighting the importance of a comprehensive dataset in achieving robust model performance.

The efficacy of the proposed approach can be observed in the results presented in Fig. 5, specifically on a single slice extracted from a patient within the *Fold13* test set, which achieved the best Dice score. Fig. 5 sequentially displays: the slice itself, the corresponding ground truth mask, the predicted mask, as well as the locations of false positive and false negative pixels.

C. Ablation study

The authors in [18] introduced an uncertainty evaluation of the predicted mask. It is based on a parameter *S* which represents the number of uncertainty steps conducted during the test phase. The final segmentation output is obtained by combining these steps with their corresponding uncertainties. We carried-out test to evaluate the model at varying of $S \in \{4, 5, 6\}$: the Dice scores corresponding to these combinations are reported in Table V and demonstrate that the optimal configuration is $S = 6$, therefore we selected this setup to train our 'MS-SegDiff' model. It is important to note that increasing the value of *S* may improve segmentation accuracy, but it would also require higher computational power.

We conducted an additional test to assess the effect of patch size on segmentation reliability. Ad described before, we evaluated performance using a patch size of 96x96x96, selected randomly. However, we performed a subsequent test using a smaller patch size of 64x64x64, so we trained the baseline model using the five configurations outlined in Table II to understand the impact of patch size (we used the baseline architecture because MS-SegDiff architecture required significantly more training time and consumes more computational resources). As observed from the results obtained in Table VI, the model with smaller patch sizes demonstrated a slight improvement, indicating that opting for a smaller patch may enhance classification accuracy. Given that a 64x64x64 patch size appears to enhance performance, we will consider further reducing the patch size, keeping in mind that the patch size should ideally be larger than the lesion size for reliable segmentation, and considering that lesions typically span more than 30 pixels. However, it's worth noting that a smaller patch size resulted in a notable increase in training time. Due to time constraints, we intend to carry out a comprehensive set of future tests across all configurations to demonstrate the actual performance improvement gained by selecting a smaller patch size.

V. CONCLUSION REMARKS

In this study, we delved into the application of Diffusion models for multiple sclerosis lesion segmentation, a task of utmost importance for clinical diagnosis and treatment planning. Our findings shed light on the challenges associated with the limited dataset size, which affected the performance of Diffusion models. Nevertheless, the experiments we conducted, including variations in parameters, such as *S* and patch size, revealed promising avenues for potential improvement in this approach. Looking ahead, our primary objective is to surpass the current state-of-the-art results in lesion segmentation. This will entail further refining the model architecture, exploring additional data augmentation strategies, and investigating advanced training techniques aimed at enhancing the model's robustness and generalization capabilities.

ACKNOWLEDGMENTS

Alessia Rondinella is a PhD candidate enrolled in the National PhD in Artificial Intelligence, XXXVII cycle, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma.

Acknowledge financial support from: PNRR MUR project PE0000013-FAIR

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 MUR – M4C2 – Investimento 1.4 - Avviso “Centri Nazionali” - D.D. n. 3138 del 16 dicembre 2021

REFERENCES

[1] Hans Lassmann “Multiple sclerosis pathology.” Cold Spring Harbor perspectives in medicine 8.3 (2018).

[2] Ruth Dobson, Giovannoni Gavin “Multiple sclerosis–a review.” European journal of neurology 26.1 (2019): 27-40.

[3] M. J. Hohol, E. J. Orav, H. L. Weiner “Disease steps in multiple sclerosis: a simple approach to evaluate disease progression.” Neurology 45.2 (1995): 251-255.

[4] M. Filippi, M. A. Horsfield, S. Bressi, V. Martinelli, C. Baratti, P. Reganati, A. Campi, D. H. Miller, G. Comi “Intra-and inter-observer agreement of brain MRI lesion volume measurements in multiple sclerosis: a comparison of techniques.” Brain 118.6 (1995): 1593-1600.

[5] Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L. Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H. Sudre and others “Longitudinal multiple sclerosis lesion segmentation: resource and challenge.” NeuroImage 148 (2017): 77-102.

[6] Francesco Guarnera, Alessia Rondinella, Oliver Giudice, Alessandro Ortis, Sebastiano Battiato, Francesco Rundo, Giorgio Fallica, Francesco Traina, Sabrina Conoci ”Early Detection of Hip Periprosthetic Joint Infections Through CNN on Computed Tomography Images.” In: Foresti, G.L., Fusiello, A., Hancock, E. (eds) Image Analysis and Processing – ICIAP 2023. ICIAP 2023. Lecture Notes in Computer Science, vol 14234. Springer, Cham. https://doi.org/10.1007/978-3-031-43153-1_12

[7] Alba Meça “Applications of Deep Learning to Magnetic Resonance Imaging (MRI),” 2023 International Conference on Computing, Electronics & Communications Engineering (iCCECE), Swansea, United Kingdom, 2023, pp. 113-120, doi: 10.1109/iCCECE59400.2023.10238598.

[8] Afshin Shoeibi, Marjane Khodatars, Mahboobeh Jafari, Parisa Moridian, Mitra Rezaei, Roohallah Alizadehsani, Fahime Khozeimeh, Juan Manuel Gorriz, Jónathan Heras, Maryam Panahiazar and others “Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review” Computers in Biology and Medicine 136 (2021): 104697.

[9] Amrita Kaur, Lakhwinder Kaur, Ashima Singh “DeepCONN: patch-wise deep convolutional neural networks for the segmentation of multiple sclerosis brain lesions” Multimedia Tools and Applications (2023): 1-33.

[10] Roba Gamal, Hoda Barka, Mayada Hadhoud “GAU U-Net for multiple sclerosis segmentation.” Alexandria Engineering Journal 73 (2023): 625-634.

[11] Alessia Rondinella, Elena Crispino, Francesco Guarnera, Oliver Giudice, Alessandro Ortis, Giulia Russo, Clara Di Lorenzo, Davide Maimone, Francesco Pappalardo, Sebastiano Battiato “Boosting multiple sclerosis lesion segmentation through attention mechanism.” Computers in Biology and Medicine 161 (2023): 107021.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin “Attention is all you need.” Advances in neural information processing systems 30 (2017).

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly and others “An image is worth 16x16 words: Transformers for image recognition at scale.” arXiv preprint arXiv:2010.11929 (2020).

[14] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, Jiangyun Li ”Transbts: Multimodal brain tumor segmentation using transformer.” Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer International Publishing, 2021.

[15] Qiran Jia, Hai Shu ”Bitr-unet: a cnn-transformer combined network for mri brain tumor segmentation.” International MICCAI Brainlesion Workshop. Cham: Springer International Publishing, 2021.

[16] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C. Kitamura, Sarthak Pati and others ”The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification.” arXiv preprint arXiv:2107.02314 (2021).

[17] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, Daguang Xu ”Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images.” International MICCAI Brainlesion Workshop. Cham: Springer International Publishing, 2021.

[18] Zhaohu Xing, Liang Wan, Huazhu Fu, Guang Yang, Lei Zhu ”Diff-UNet: A Diffusion Embedded Network for Volumetric Segmentation.” arXiv preprint arXiv:2303.10326 (2023).

[19] Nils Gessert, Julia Krüger, Roland Opfer, Ann-Christin Ostwaldt, Praveena Manogaran, Hagen H. Kitzler, Sven Schippling, Alexander Schlaefer ”Multiple sclerosis lesion activity segmentation with attention-guided two-path CNNs.” Computerized Medical Imaging and Graphics 84 (2020): 101772.

[20] Maryam Hashemi, Mahsa Akhbari, 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 (2022): 105402.

[21] Beytullah Sarica, Dursun Zafer Seker, Bulent Bayram ”A dense residual U-net for multiple sclerosis lesions segmentation from multi-sequence 3D MR images.” International Journal of Medical Informatics 170 (2023): 104965.

[22] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, Yanwu Xu ”Medsegdiff: Medical image segmentation with diffusion probabilistic model.” arXiv preprint arXiv:2211.00611 (2022).

[23] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yanwu Xu ”Medsegdiff-v2: Diffusion based medical image segmentation with transformer.” arXiv preprint arXiv:2301.11798 (2023).

[24] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, Philippe C. Cattin ”Diffusion models for implicit image segmentation ensembles.” International Conference on Medical Imaging with Deep Learning. PMLR, 2022.

[25] Tao Chen, Chenhui Wang, Hongming Shan ”BerDiff: Conditional Bernoulli Diffusion Model for Medical Image Segmentation.” arXiv preprint arXiv:2304.04429 (2023).

[26] Jonathan Ho, Ajay Jain, Pieter Abbeel ”Denoising diffusion probabilistic models.” Advances in neural information processing systems 33 (2020): 6840-6851.

[27] Jiaming Song, Chenlin Meng, Stefano Ermon ”Denoising diffusion implicit models.” arXiv preprint arXiv:2010.02502 (2020).

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga and others ”Pytorch: An imperative style, high-performance deep learning library.” Advances in neural information processing systems 32 (2019).

[29] M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang and others ”Monai: An open-source framework for deep learning in healthcare.” arXiv preprint arXiv:2211.02701 (2022).

[30] Lee R. Dice ”Measures of the amount of ecologic association between species.” Ecology 26.3 (1945): 297-302.

[31] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald and others ”U-Net: deep learning for cell counting, detection, and morphometry.” Nature methods 16.1 (2019): 67-70.

[32] Zilong Zhong, Zhong Qiu Lin, Rene Bidart, Xiaodan Hu, Ibrahim Ben Daya, Zhifeng Li, Wei-Shi Zheng, Jonathan Li, Alexander Wong ”Squeeze-and-attention networks for semantic segmentation.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[33] Ilya Loshchilov, Frank Hutter ”Sgdr: Stochastic gradient descent with warm restarts.” arXiv preprint arXiv:1608.03983 (2016).

[34] Florian Raab, Simon Wein, Mark Greenlee, Wilhelm Malloni, Elmar Lang ”A multimodal 2D Convolutional Neural Network for Multiple Sclerosis Lesion Detection.” (2022).

[35] Shahab Aslani, Michael Dayan, Loredana Storelli, Massimo Filippi, Vittorio Murino, Maria A. Rocca, Diego Sona ”Multi-branch convolutional neural network for multiple sclerosis lesion segmentation.” NeuroImage 196 (2019): 1-15.