Institut Polytechnique de Paris
Telecom SudParis

Projet Cassiopée n°77
Semantic segmentation of brain tumors via Mamba

Encadrant: Nicolas Rougon

Students:
Ayman Orkhis
Gustavo Melo Scheidt Paulino
Mariana Simões Penello Meirelles

06/2025

Contents

1. Introduction

1.1 Context

Semantic segmentation of brain tumors in Magnetic Resonance Imaging (MRI) is crucial for medical diagnosis and treatment planning. In recent years, U-shaped neural architectures have become the standard in this field.

The first generation of these models, represented by traditional U-Net [13], relies on convolutional layers but has limitations in modeling long-range dependencies. The second generation introduced Transformer-based models, such as ViT [2] and SwinUNETR [14], which enhanced global context modeling through self-attention mechanisms. However, these models faced challenges due to their quadratic computational complexity and high data requirements, which limited their scalability, particularly for 3D medical images.

Recently, a third generation of models based on State Space Models (SSMs) has emerged. Among these, Mamba [3] presents a new approach to modeling long-range dependencies with quasi-linear complexity, effectively combining memory efficiency with strong performance. These qualities make Mamba architectures particularly promising for 3D medical image segmentation.

1.2 Project Objectives

This project aims to evaluate and compare Mamba-based neural network architectures for brain tumor segmentation in multiparametric MRI using the BraTS2021 dataset [(9), (10)]. We benchmark three models — SegMamba [(4)], U-Mamba [(5)], and VM-UNet [(7)] — against established baselines, including a CNN-based model (MedNeXt) and a Transformer-based model (SwinUNETR). All implementations are integrated into the nnUNet framework [12] to ensure consistency in preprocessing, training, inference, and evaluation.

1.3 Contributions

This project makes several contributions to the field of medical image segmentation. Firstly, tumor regions were carefully **classified using 3D Slicer**, in which we completed a Google Sheets to ensure that all data's analysis is easily found, and also thoroughly identified critical anatomical structures. The original training scripts were replaced with a unified **nnUNetTrainer** framework from nnU-Net, which facilitates the training process for enhanced efficiency.

Additionally, customized trainers for the **SegMamba**, **U-Mamba**, and **VM-Unet** 3D models were developed, applying uniform training parameters across all models to facilitate

direct comparisons. To improve training stability, **deep supervision** was integrated into the models.

It was also necessary to launch a thorough re-evaluation of previously omitted components from the original repositories, ensuring the inclusion of all relevant elements. Furthermore, a novel **3D architecture inspired by VM-Unet** specifically for MRI segmentation was designed, contributing a new perspective to existing methodologies.

To assess the effectiveness of our models, we accomplished a comprehensive **comparison of the Mamba architectures** against CNN- and Transformer-based baselines, providing valuable insights into their performance. Finally, a user-friendly web-based interface for model deployment was developed.

2. Literature Review and Theoretical Foundations

2.1. Literature review

The study began with the U-Net architecture, which became the basis for medical image segmentation due to their encoder-decoder structure with skip connections. To understand long-range dependencies, RNNs, LSTMs, and GRUs were studied, which introduced the concept of recurrent hidden states (which would be useful in SSMs).

The next important mark was understanding Word Embedding methods such as Word2Vec to understand how discrete inputs (image patches) can be mapped to vector spaces. After that, Seq2Seq Encoder-Decoder models and the Attention mechanism, due to their introduction of dynamic context weighting to improve performance in sequence modeling.

Then, the basis of almost all of AI's work today, the Transformers [1], which revolutionized sequence modeling by replacing recurrence with self-attention, and inspired architectures like SwinUNETR, used as the project's baseline.

The last pillar was State Space Models (SSMs), used in control theory but recently adapted for deep learning. With a focus on Mamba, which improves SSMs with selective scan and hardware-aware algorithms, achieving efficient long-range modeling.

To apply Mamba to vision, Vision Mamba (ViM) [8] and VMamba [6] were studied, which addresses spatial structure and context in image data through bidirectional SSMs and positional embeddings.The topics mentioned above will be explained in further detail in the following sections.

2.2. Traditional CNN-based Models

Convolutional Neural Networks (CNNs) have been the foundation of medical image segmentation due to their efficiency and ability to capture local spatial patterns. U-Net architectures introduced a U-shaped design, which combines convolution and pooling in the encoder with upsampling in the decoder, along with skip-connections to improve the capacity of capturing these local spatial patterns. However, CNNs have naturally limited receptive fields, which restricts their ability to capture long-range dependencies, which is a crucial aspect of 3D medical segmentation tasks. Although extensions such as UX-Net and MedNeXt aim to expand the receptive field by incorporating larger kernels or modern architectural blocks, they still struggle to effectively model global context.

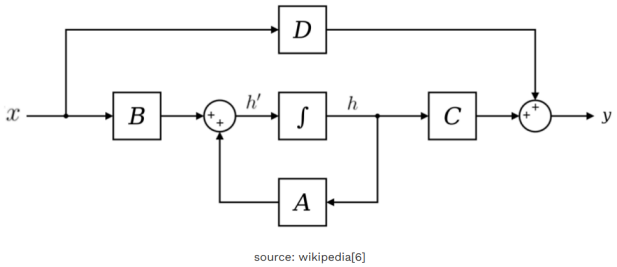
2.3. Transformer-based Models

Transformers revolutionized the field of artificial intelligence due to the self-attention mechanism, which captures long-range dependencies, addressing one of the key challenges in CNNs-based architectures. They have been effectively adapted for vision tasks through the introduction of the Vision Transformer (ViT), which processes images as sequences of patches. This approach allows for the modeling of global dependencies via self-attention mechanisms. Architectures such as UNETR and SwinUNETR have successfully integrated Transformer-based encoders into the U-Net framework, enhancing the modeling of global context and achieving state-of-the-art performance across various segmentation benchmarks. However, a significant limitation of Transformers in 3D medical imaging is their high computational and memory complexity, which scales quadratically with input size. This scaling makes them less suitable for processing high-resolution volumetric data.

2.4. State Space Models (SSMs)

State Space Models (SSMs) have their foundation in control theory and are used to describe and predict the behavior of dynamic systems over time. In their classical formulation, the system is described through a hidden state that evolves based on previous states and external inputs, producing observable outputs, enabling precise modeling of time-dependent processes. Figure 1 illustrates this classical continuous-time SSM representation.

$$h'(t) = Ah(t) + Bx(t)\,;\; h_{new}(t) = \int h'(t)dt$$
$$y(t) = Ch_{new}(t) + Dx(t)$$



source: wikipedia[6]

Figure 1: Classic SSM

Since its introduction, SSMs have been reformulated to serve as powerful sequence models. By translating their recurrent nature into efficient parallelizable operations, modern neural SSMs can capture long-term dependencies while benefiting from scalable training on GPUs, making them ideal for volumetric medical image segmentation, where understanding global structure across multiple slices is essential.

2.5. Mamba

Mamba is a recent sequence modeling architecture based on State Space Models, designed to efficiently capture long-range dependencies with linear computational complexity. Unlike self-attention mechanisms, Mamba uses structured state dynamics combined with a selective scan operation that enables fast and parallelizable processing of very long sequences. This design allows it to maintain temporal memory and context over extended inputs without the computational burden associated with Transformers.

Mamba proposes a selective, hardware-aware version of SSMs that further improve computational efficiency, as illustrated in Figure 2.

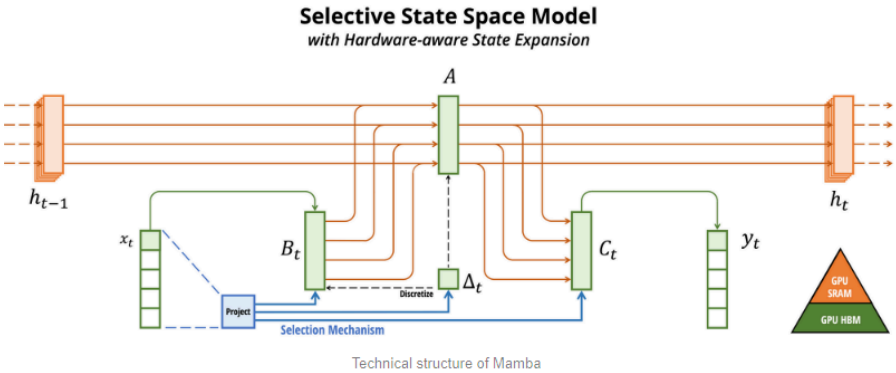


Figure 2: Selective SSM

These models are typically implemented as blocks composed of projections, activations, and the selective SSM core, as illustrated in Figure 3.

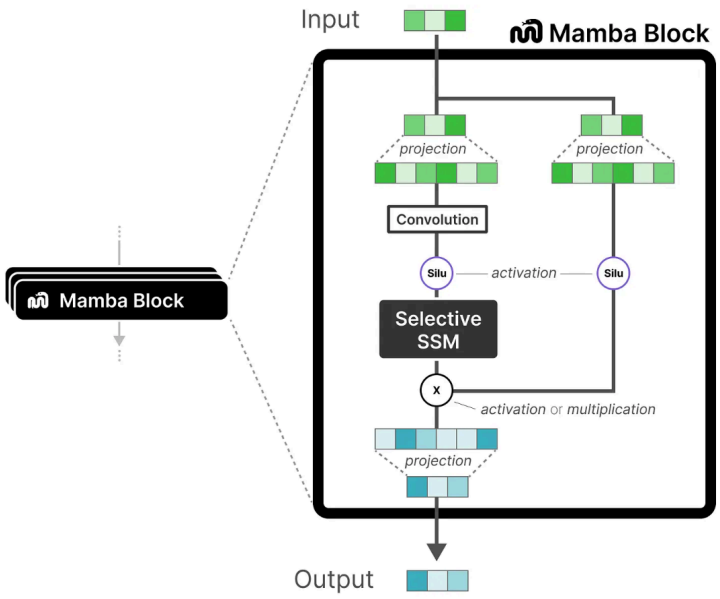


Figure 3: Mamba Block

To extend Mamba's applicability to vision tasks, researchers have adapted its internal operations to handle spatially structured data such as 2D images and 3D volumetric scans. These adaptations involve reinterpreting sequence dimensions as spatial axes and redesigning Mamba blocks to operate over feature maps, preserving spatial hierarchies important for segmentation tasks. This gave rise to architectures like Vision Mamba and VMamba, which apply Mamba blocks as visual encoders while benefiting from the model's efficient long-range modeling.

2.6. Mamba-Based Models for Medical Image Segmentation

The integration of Mamba into medical image segmentation led to the development of several architectures adapted to 2D and 3D volumetric data. One of the first efforts in this direction was U-Mamba, which replaces the encoder of a U-Net with Mamba blocks while retaining the decoder structure from nnUNet. This hybrid design leverages the efficient long-range modeling of Mamba in the encoding stage while preserving the strong localization abilities of CNN-based decoders, showing its potential as a lightweight and effective alternative to both CNNs and Transformers.

VM-UNet model proposes a fully Mamba-based U-Net architecture that integrates VMamba blocks in both encoder and decoder. Although limited to 2D slices, VM-UNet demonstrates that Mamba-based blocks can effectively replace convolutional layers across an

entire segmentation. In parallel, SegMamba introduced a more advanced 3D feature, which introduced a Tri-orientated Spatial Mamba (TS-Mamba) block, that models spatial dependencies along three anatomical planes. In addition, it incorporates modules such as Gated Spatial Convolution (GSC) and Feature-level Uncertainty Estimation (FUE) to improve spatial precision and robustness in volumetric segmentation.

3. Dataset and Preprocessing

3.1 The BraTS 2021 Dataset

The BraTS 2021 [9] dataset is a benchmark for brain tumor segmentation in multiparametric magnetic resonance imaging (mpMRI). The dataset includes 1,265 cases, each consisting of four aligned MRI modalities: T1-weighted (T1), T1-weighted post-contrast (T1-CE), T2-weighted (T2), and Fluid-Attenuated Inversion Recovery (FLAIR). All scans are co-registered to the same anatomical template, resampled to an isotropic resolution of $1mm^3$, and skull-stripped [10]. The accompanying ground truth segmentations were manually annotated by expert neuroradiologists and include voxel-level labels identifying tumor subregions.

It is particularly suitable for evaluating 3D segmentation models, as it offers diverse tumor shapes, sizes, and locations across multiple imaging modalities. It also provides a consistent preprocessing pipeline, making it compatible with automated frameworks such as nnUNet. Each case includes a segmentation mask delineating key tumor regions, enabling a detailed analysis of model performance on specific substructures.

3.2 Modalities and Ground Truth Labels

Each of the four MRI modalities in BraTS 2021 serves a distinct diagnostic purpose and provides information for tumor segmentation. T1-weighted images offer high-resolution anatomical detail and are used to delineate brain structures. T1-CE (T1 with contrast enhancement) highlights regions of active tumor where the blood-brain barrier is disrupted, which is crucial for identifying enhancing tumor tissue. T2-weighted images are sensitive to fluid accumulation, making them suitable for visualizing edema and broader lesion boundaries. FLAIR suppresses cerebrospinal fluid (CSF) signals and is particularly effective in detecting infiltrative tumor regions and peritumoral edema.

The segmentation ground truth is encoded using four label classes. Label 0 corresponds to background (non-tumorous tissue). Label 1 represents the Tumor Core (TC), including the necrotic core, non-enhancing tumor, and enhancing regions. Label 2 designates the Whole Tumor (WT), encompassing all tumor-related abnormalities, including edema. Finally, label 3 denotes the Enhancing Tumor (ET), which typically corresponds to high-grade tumor regions with active contrast uptake.

3.3 Preprocessing Pipeline

The preprocessing pipeline was developed using the nnU-Net framework, which automates key steps such as data normalization, resampling, cropping irrelevant regions, and converting images into manageable patches. Each nnU-Net task - identified by a unique Task ID like Task100_BrainTumor - organizes all components required for segmentation, **including raw** and **preprocessed data**, **trained models**, and **evaluation results**, following a strict directory structure (Fig 4) that ensures standardization, automation, and reproducibility. The raw data is stored in **nnUNet_raw/**, and preprocessing - triggered by **nnUNet_plan_and_preprocess** - handles modality-specific normalization (z-score for MRI), resampling to a consistent voxel spacing, and background removal. The outputs are saved under **nnUNet_preprocessed/**.

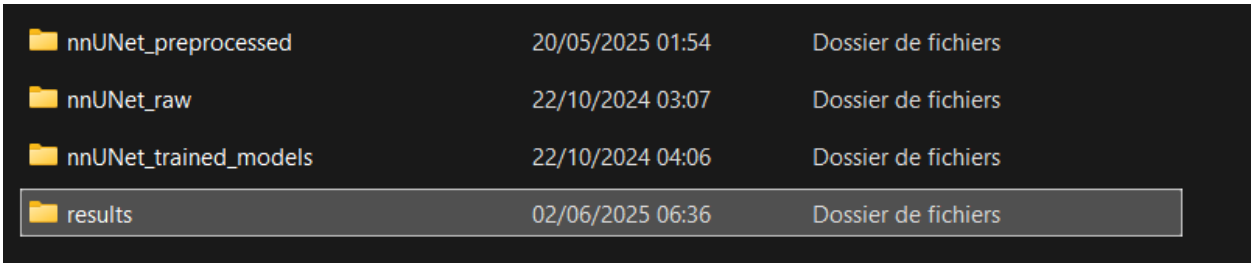


Figure 4: Directory structure in **nnUnet_V1**

For model training, **nnUNet_train** allows configurations such as 2D, 3D full resolution, and low resolution. Inference is performed using **nnUNet_predict**, and evaluation is done with **nnUNet_evaluate_folder**, providing metrics like **Dice Score** and **Hausdorff Distance**. In our case, SegMamba was integrated into **nnU-Net v1**.

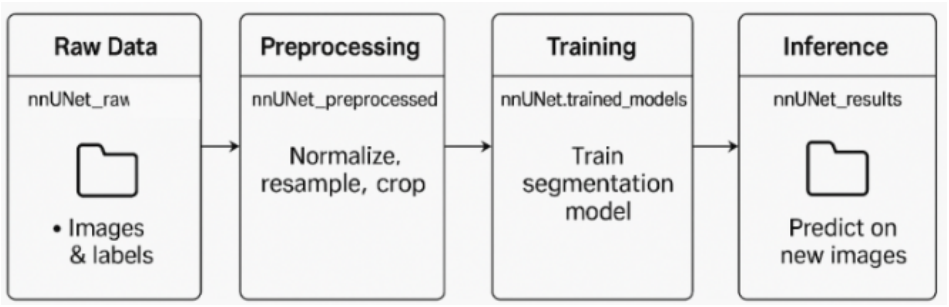


Figure 5: nnUnet Pipeline

In contrast, **U-Mamba** and **VM-UNet** used **nnU-Net v2**, an improved and simplified version of the framework that makes customization and integration of new architectures much easier.

4. Methodology

4.1 Baseline Architectures

4.1.1 MedNeXt

The primary baseline is MedNext, which utilizes a U-shaped architecture composed of convolutional blocks with progressive downsampling and upsampling. This design is efficient and scalable for 3D medical image segmentation.

4.1.2 SwinUNETR

SwinUNETR serves as our Transformer-based baseline and follows a U-shaped architecture. It utilizes the Swin Transformer in the encoder and incorporates skip connections with a CNN decoder. This design offers strong performance in volumetric segmentation and acts as a benchmark for attention-based architectures.

4.2 Mamba-based Architectures

4.2.1 SegMamba

SegMamba is 3D architecture that combines convolutional and Mamba-based modules. It uses the Tri-orientated Spatial Mamba (TS-Mamba) block in the encoder. As in the Figure 6, Gated Spatial Convolution (GSC) and Feature-level Uncertainty Estimation (FUE) modules are integrated to improve spatial precision and robustness in scales.

To improve gradient flow and enhance training stability, deep supervision was integrated. This technique allows intermediate outputs from multiple decoder stages to contribute to the final loss, thereby providing more direct gradient flow to earlier layers in the network.

Additionally, a custom trainer named nnUNetTrainer_SegMamba was implemented. This trainer inherits from the base nnUNetTrainer class and has been tailored specifically to handle the architectural and training nuances of SegMamba. It ensures that the deep supervision mechanism is properly applied during training and integrates seamlessly with the nnU-Net framework.

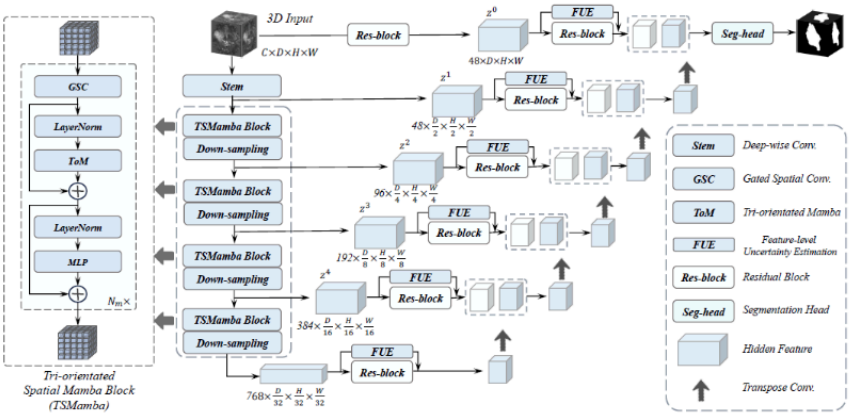


Figure 6: Segmamba architecture

4.2.2 U-Mamba

U-Mamba integrates Mamba blocks into the encoder of a U-Net architecture while preserving a CNN decoder. Two training configurations were explored: one features a single Mamba block at the bottleneck, and the other employs Mamba blocks throughout the entire encoder, as shown in Figure 7. This hybrid design leverages the long-range modeling capabilities of Mamba while retaining the spatial resolution benefits provided by convolutional decoding.

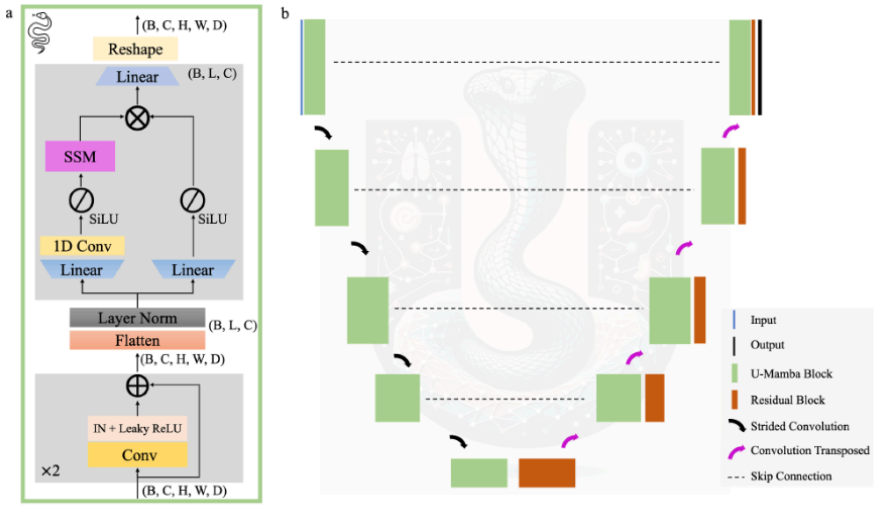


Figure 7: U-Mamba architecture

4.2.3 VM-UNet

VM-UNet is a fully SSM-based architecture that uses VMamba blocks. It replaces both encoder and decoder paths with SSM modules. The architecture uses VSS blocks, which incorporate the Selective SSM operator (SS2D) along with normalization and convolution layers. As shown in Figure 8, VM-UNet uses patch embedding, merging, and expanding operations to

perform all processing with Mamba-based components. To extend its applicability to MRI image segmentation, we developed a 3D-adapted version of VM-UNet.

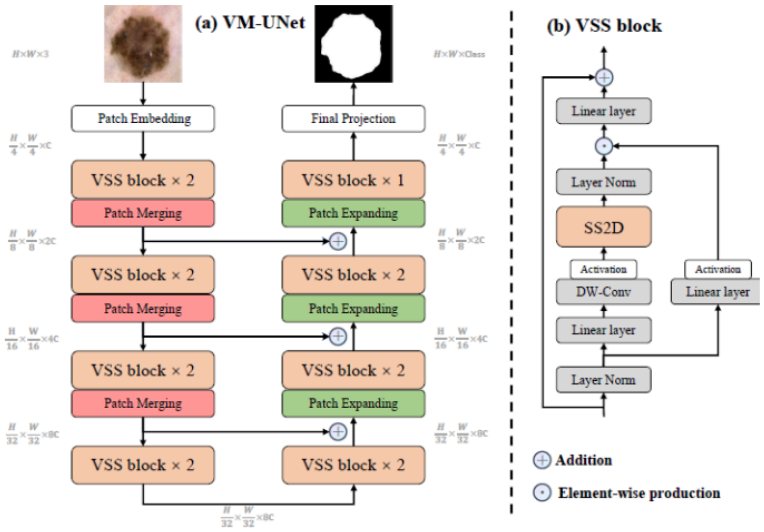


Figure 8: VM-UNet architecture

4.3 Integration with nnUNet Framework

All Mamba-based models were integrated into the nnUNet framework to leverage its standardized pipeline for preprocessing, training, and evaluation. A custom network class was created to support Mamba blocks and properly manage deep supervision during both training and inference.

Each model received a dedicated trainer to initialize architectural parameters such as depth and feature sizes, override pooling configurations, and manage training with the AdamW optimizer using a custom eps value to address gradient stability. This integration ensured full compatibility with nnUNet's automated setup and allowed consistent comparison across models.

4.4 Training Strategy and Challenges

Training of SegMamba began with Tasks 600–610, initially disabling both deep supervision and mixed precision (fp16). Under these conditions, the model exhibited gradient instability and failed to converge. Reintroducing deep supervision significantly improved gradient flow and enabled successful training, indicating a gradient vanishing issue when supervision is absent. To further stabilize training, mixed precision was re-enabled and the eps parameter of the AdamW optimizer was adjusted. This change resolved gradient-related issues, enabling convergence even without deep supervision.

For U-Mamba, the architecture was already compatible with the nnUNet v2 framework, benefiting from its modular configuration, improved logging, and simplified experimentation. During training, several hyperparameters were adjusted: the initial learning rate was set to $1\times10^{-3}1$, and the epsilon parameter of the AdamW optimizer to 1×10^{-4}. These modifications were implemented within the custom trainer to better suit the model's convergence behavior. No significant training issues were observed.

VM-Mamba integration is currently under development. The basic idea of the project was to turn the blocks already existing (such as SS2D), into 3D versions of themselves. For that there were two main architectures used as the basis for these new 3D blocks: MedSegMamba [Cao et al., 2024] and Mamba Morph [Guo et al., 2024]. The challenges consisted on adapting each block with the already existing 2D version of VM-UNet, whilst attempting to be the most faithful possible to the original version.

All models were trained using the Dice loss function and optimized with AdamW. Experiments were conducted with training durations of 400 and 800 epochs, depending on the specific task.

5. Experiments and Results

5.1 Evaluation Metrics

The performance of segmentation was assessed using the Dice Similarity Coefficient (DICE), which is a standard metric for evaluating medical image segmentation. The Dice coefficient measures the overlap between the predicted segmentation and the ground truth, with a range from 0 (indicating no overlap) to 1 (indicating perfect overlap). Higher Dice scores reflect better accuracy in the segmentation of tumor subregions.

5.2 Quantitative Comparison

Table I presents the Dice scores for the Mamba-based architectures across three tumor subregions: Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET). The performance metrics are reported for both training and testing phases, along with training loss and epoch duration.

SegMamba (IDs 610–611) achieved the highest Dice score in the ET region, reaching 0.848, and demonstrated consistently strong performance in WT segmentation. In contrast, U-MambaEnc (ID 653) outperformed all other models in the TC region with a score of 0.757, which is recognized as the most challenging subregion across all architectures. Notably, the U-MambaBot variant (ID 652) produced competitive results while requiring significantly less training time, indicating a beneficial trade-off between efficiency and performance.

All models exhibited high consistency in WT segmentation, with Dice scores ranging from approximately 0.83 to 0.84, although overall performance in TC was slightly lower. The segmentation for ET was generally the most stable across the different architectures.

It is important to mention that the evaluation of the VM-UNet trainer was not completed in time, resulting in the exclusion of its metrics from this comparison.

Task ID	Architecture	Number of Epochs	Epoch duration		Final Loss	TRAINING | raw (1)			TESTING (2)		
						DICE (postprocessing/raw)			DICE		
			min	max	(train loss)	WT	TC	ET	WT	TC	ET
610	SegMamba	400	188.99	200.48	0.8603	0.925	0.864	0.924	0.839	0.749	0.848
611	SegMamba	800	188.99	200.48	0.8777	0.933	0.881	0.934	0.839	0.746	0.846
650	UMambaBot	400	180.46	192.31	0.877	0.934	0.878	0.932	0.839	0.747	0.848
651	UMambaEnc	400	177.8	183.98	0.8612	0.916	0.844	0.914	0.831	0.746	0.841
652	UMambaBot	800	95.38	102.34	0.8685	0.934	0.877	0.93	0.836	0.754	0.848
653	UMambaEnc	800	177.78	184.49	0.8787	0.927	0.864	0.926	0.836	0.757	0.843

Table I: Qualitative Results

5.3 Qualitative Analysis and Visual Results

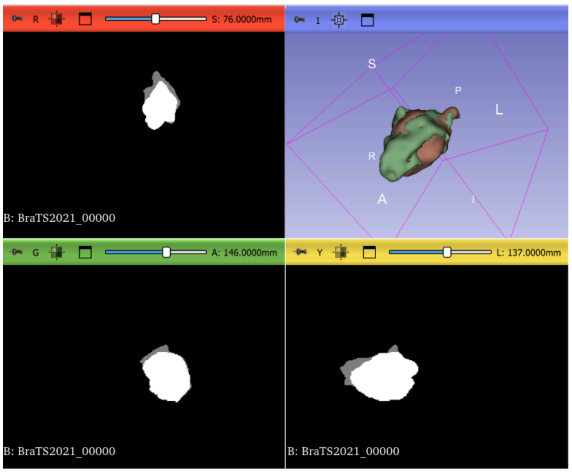


Figure 9: Results of the Segmentation (SegMamba)

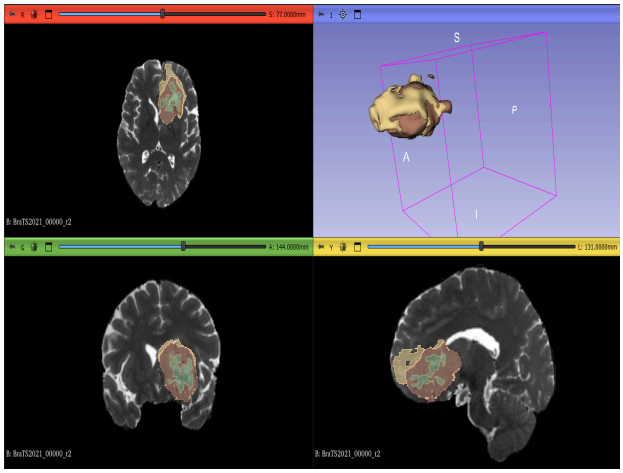


Figure 10: The ground Truth data

6. Conclusion

The purpose of this work was to assess Mamba-based models for semantic segmentation of brain tumors in 3D MRI scans from the BraTS 2021 dataset. These models were compared with conventional convolutional approaches such as MedNeXt and Transformer-based models like SwinUNETR. The results indicate that Mamba architectures offer an extremely efficient alternative for volumetric segmentation tasks by combining long-range modeling with linear computational complexity, although their precision needs refining in order to surpass the standard models in the market.

A key contribution of this work is the standardized integration of the SegMamba, U-Mamba, and VM-UNet models into the nnUNet framework. By developing custom trainers and implementing deep supervision to stabilize the training process, issues like vanishing gradients were targeted and enabled direct comparisons between architectures. SegMamba achieved the best results in the Enhancing Tumor (ET) region with a Dice score of 0.848, while U-MambaEnc reached the highest accuracy in the Tumor Core (TC) region with a Dice score of 0.757, a particularly challenging area. All models demonstrated consistent performance in Whole Tumor segmentation (WT), with Dice scores ranging from 0.83 to 0.84. The U-MambaBot variant resulted in a balanced option, providing a mix of accuracy and computational efficiency.

Despite these advancements, significant limitations were identified. The 3D version of VM-UNet, proposed as an innovative adaptation for volumetric data, could not be fully evaluated due to the difficulties of converting 2D blocks to 3D. Additionally, it is notable that hyperparameter tuning, such as adjusting the epsilon parameter of the AdamW optimizer and using mixed precision, is very important for ensuring training stability.

For future work, next steps involve: (1) finalizing the 3D implementation of VM-UNet; (2) optimizing training strategies to reduce instabilities; and (3) validating the models in heterogeneous clinical settings. In conclusion, Mamba-based models represent a promising advancement in 3D medical segmentation, offering computational efficiency without sacrificing much accuracy, which is especially relevant for real-time applications and resource-constrained environments.

7. References

[1] A. Vaswani et al., "Attention Is All You Need," arXiv preprint arXiv:1706.03762, 2017.

[2] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020.

[3] S. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," arXiv preprint arXiv:2312.00752, 2023.

[4] Z. Xing et al., "SegMamba: Long-range Sequential Modeling Mamba for 3D Medical Image Segmentation," in MICCAI, 2024.

[5] Q. Ma et al., "U-Mamba: A Hybrid SSM-CNN Architecture for General Medical Image Segmentation," arXiv preprint arXiv:2401.13141, 2024.

[6] Y. Liu et al., "VMamba: Visual State Space Models with Selective Scanning," arXiv preprint arXiv:2401.04088, 2024.

[7] Z. Ruan et al., "VM-UNet: Pure State Space U-Net Architecture for Medical Image Segmentation," arXiv preprint arXiv:2403.04572, 2024.

[8] L. Zhu et al., "Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model," arXiv preprint arXiv:2401.09417, 2024.

[9] "BraTS 2021 Dataset," Papers with Code. Available: https://paperswithcode.com/dataset/brats21

[10] "BraTS 2020 Dataset," University of Pennsylvania.. Available: https://www.med.upenn.edu/cbica/brats2020/data.html

[11] Maarten Grootendorst, "A visual guide to Mamba and State Space Models," ML Minded newsletter. Available: https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mamba-and-state

[12] F. Isensee et al., “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” Nature Methods, vol. 18, pp. 203–211, 2021.

[13] Ronneberger et al, “U-Net: Convolutional Networks for Biomedical Image Segmentation”, arXiv preprint arXiv: 1505.04597, 2015.

[14] Hatamizadeh et al, “Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images”, arXiv preprint arXiv: 2201.01266, 2022.