Satakunnan ammattikorkeakoulu
Satakunta University of Applied Sciences

RISHIKA UMESH KALYANAPPA

Early Fusion-based Self-Supervised Learning (SSL) with Cross-Modality Prediction for Brain Tumour Segmentation

DEGREE PROGRAMME IN 2025

ABSTRACT

Umesh Kalyanappa, Rishika: Early Fusion-based Self-Supervised Learning (SSL) with Cross-Modality Prediction for Brain Tumour Segmentation
Bachelor's thesis

Data Engineering

June 2025

Number of pages: 40

The study investigated the impact of self-supervised learning (SSL) pretraining on brain tumour segmentation performance using the BraTS2020 dataset from Kaggle. A two-stage methodology was performed. The first stage involved training an SSL model to predict FLAR MRI modality from other modalities (T1 and T1Gd) without relying on labels. Reconstruction quality was assessed using Mean Squared Error (MSE), Peak Signal-to-Noise ratio (PSNR), and Structural similarity index (SSIM). In the second stage, a ResUNet-based segmentation model was trained to perform multi-class segmentation of tumour subregions, which were encoded into a single-channel label map.

The model was trained both with SSL-initialized weights and with random initialization to compare performance. Early fusion was performed by stacking the four MRI modalities (T1, T1Gd, T2, FLAIR) as input channels, allowing the model to learn combined anatomical and pathological features from the initial layers.

A patch-based sampling strategy was employed, along with various loss functions, including Dice combined cross-entropy, Dice combined focal loss, and focal loss only. The learning strategy was based on 5-fold cross-validation. The segmentation model was evaluated on a hold-out test set. Performance for each tumour subregion was measured using the Dice coefficient and Average Surface Distance (ASD) metrics. The same evaluation methodology was applied for the SSL encoder. The results demonstrated that using SSL for weight initialization led to slightly higher Dice scores and lower ASD values across most configurations.

The findings indicate that implementing SSL can enhance the model accuracy and robustness in medical image segmentation tasks, especially when training with a limited amount of labelled data.

Keywords: Self-supervised learning, Brain tumour segmentation, Deep learning, BraTS 2020, Medical imaging, ResUNet, Dice Coefficient, Average surface distance, patch-based training, early fusion

PREFACE

This bachelor's thesis was completed as part of my final study requirements for the Bachelor of Engineering degree in Data Engineering at Satakunta University of Applied Sciences. This work has been carried out independently from December 2024 to June 2025.

This topic was chosen based on my strong interest in Artificial intelligence, deep learning, and medical imaging. The objective was to explore how AI can contribute to the field of medical imaging, particularly in situations where annotated data is scarce. Therefore, this study focuses on evaluating the impact of SSL on enhancing segmentation performance in medical image analysis.

I would like to express my sincere gratitude to my thesis supervisor, Mitra Daneshmand, for her valuable guidance, support, and feedback throughout this project. I am also grateful to the Faculty of Engineering at Satakunta University of Applied Sciences for providing me with the academic environment and resources necessary to conduct this research.

CONTENTS

LIST OF SYMBOLS AND TERMS

MRI – Magnetic Resonance Imaging, a medical imaging method used to visualize internal brain structures through various contrast sequences.

T1 – T1-weighted MRI scan captures the anatomical details, which are useful for identifying normal brain structures.

T1Gd – T1-weighted MRI with Gadolinium, which highlights actively enhancing tumour regions using contrast agent.

T2 – T2 weighted MRI scan shows fluid content and edema with bright pathological regions.

FLAIR – Fluid Attenuated Inversion Recovery, which suppresses fluid (CSF), signals to highlight lesions and edema.

NCR/NET (Class 1) – Necrotic and non-enhancing tumour core, a central part of the tumour that does not enhance with contrast.

ED (Class 2) – Peritumoral edema, swelling surrounding the tumour, seen in T2 and FLAIR.

ET (Class 3) – Enhancing tumour, a region that actively enhances on T1Gd due to the blood-brain barrier disruption.

1 CHAPTER ONE/ INTRODUCTION

Recently, the domain of medical image segmentation has garnered significant attention in the research community, as brain tumour segmentation has become a challenging task due to its importance in clinical diagnosis and patient management. Accurate delineation of tumour regions such as enhancing tumours (ET), edema (ED), and necrotic or non-enhancing tumour cores (NCR/NET) is essential for effective treatment planning. However, distinguishing between pathological and healthy brain tissue requires expert knowledge, making the process both time-consuming and prone to inter-observer variability. Manual segmentation by clinicians is both resource-intensive and subject to inconsistencies. To address these challenges, automated segmentation methods using deep neural networks have emerged as promising solutions for tumour delineation (Bauer et al., 2013; Havaei et al., 2017; Ronneberger et al., 2015).

Magnetic Resonance Imaging (MRI) is a non-invasive imaging modality that provides detailed anatomical information for brain tumour segmentation (Sudre et al., 2017; National Institute of Biomedical Imaging and Bioengineering, n.d.). MRI can capture multiple imaging modalities, including T1-weighted (T1), contrast-enhanced T1-weighted (T1Gd), T2-weighted (T2), and Fluid-Attenuated Inversion Recovery (FLAIR). Each of these modalities provides different information about tumour characterization (Pan et al., 2024). As shown in Figure 1, these modalities highlight different aspects of brain tissue:

- T1 (Figure 1a) provides high-contrast images, where healthy tissues appear bright and pathological tissues appear darker, making it useful for visualizing anatomical structures and assessing brain morphology (MRI basics, 2016; Pan et al., 2024).
- T1Gd (Figure 1b) is a specialized type of T1-weighted MRI that utilizes a contrast agent called Gadolinium, which highlights tumour

regions, enabling the differentiation between tumorous and non-tumorous tissues (MRI basics, 2016; Pan et al., 2024).

- T2 (Figure 1c) offers contrast opposite to T1, where the pathological tissues appear brighter. This is useful in detecting edema (brain swelling) (MRI basics, 2016; Pan et al., 2024).
- FLAIR (Figure 1d) is a specialized T2-weighted modality that suppresses cerebrospinal fluid signals to enhance the visibility of lesions and edematous regions (MRI basics, 2016; Pan et al., 2024).

Previous research has shown that combining these modalities leads to improved segmentation performance, due to the complementary nature of the information they provide (Chen et al., 2018; Oktay et al., 2018; Khan et al., 2023).

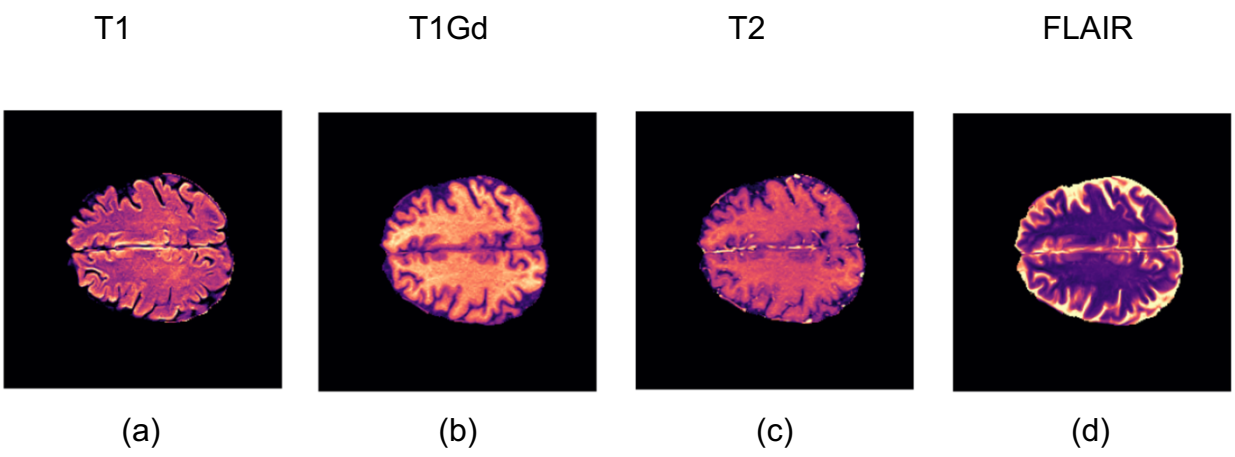

Figure 1. Representative axial slices of the four MRI modalities used in this study: (a)T1-weighted(T1), (b) contrast-enhanced T1-weighted (T1Gd), (c) T2-weighted(T2), and (d) Fluid-Attenuated Inversion Recovery (FLAIR).

Most existing research methods are based on the conventional U-Net architecture, which typically integrates multimodal MRI data by concatenating the different modalities as separate input channels (Myronenko, 2018; Ronneberger et al., 2015). While effective, this approach often overlooks the relationships between different modalities, limiting the potential for more holistic feature learning (Ronneberger et al., 2015; Taleb et al., 2019). Recently, self-supervised learning (SSL) has gained traction in medical analysis (Azizi et

al., 2021; Taleb et al., 2021; Taleb et al., 2019). SSL utilizes unlabelled data to learn useful representations through pretext tasks, offering a promising solution to the challenges posed by supervised segmentation methods (Huang et al., 2023).

In this study, an early fusion-based SSL framework is proposed for brain tumour segmentation. The SSL model is first trained on a cross-modality reconstruction task, where it learns to predict FLAIR modality from T1 and T1Gd scans. The encoder weights from this task are then partially transferred into a four-channel segmentation model that uses all MRI modalities as input (Chaitanya et al., 2020; Chang et al., 2019).

The segmentation model adopts a ResUNet-style architecture with skip connections to retain spatial information (Ronneberger et al., 2015). To address the class imbalance, patch-based sampling, and appropriate loss functions, such as Focal Loss, Dice-Cross Entropy, and Dice+Focal Loss, were employed (Lin et al., 2017; Sudre et al., 2017).

In this thesis, the core objectives are twofold:

1) To assess the capability of early fusion combined with SSL in extracting meaningful multimodal representations during the pretraining phase. This is achieved through a cross-modality reconstruction task, where the model learns to predict one MRI modality from others without using manual labels.
2) To evaluate how well these pretrained representations transfer to the segmentation phase. The pretrained encoder, which captures cross-modality relationships, is fine-tuned with a UNet-style decoder to perform multi-class segmentation of brain tumour subregions.

2 RELATED WORKS

Related works on image segmentation, multimodal learning, and self-supervised learning are presented below.

2.1 Medical image segmentation

Medical image segmentation refers to the process of dividing an image into anatomically or pathologically meaningful regions, such as tumours or healthy tissues. Early approaches relied heavily on handcrafted features, such as Gaussian kernels, Fourier transforms, and wavelet filtering, combined with traditional classifiers, including support vector machines (SVMs) (Azizi et al., 2021). Over time, the emergence of deep learning, particularly Convolutional Neural Networks (CNNs), has transformed the field by enabling automatic end-to-end learning. CNN-based models have reduced the dependency on manually engineered features and large amounts of labelled data (Chen et al., 2018).

With the need for accurate boundary predictions and robust performance on limited datasets, the U-Net architecture was introduced, which became one of the most influential models in biomedical segmentation tasks (Ronneberger et al., 2015). It employs an encoder-decoder structure with skip connections, enabling the model to retain fine-grained spatial information (Havaei et al., 2017; Oktay et al., 2018). Over time, various extensions of U-Net have been proposed, incorporating enhancements such as dilated convolutions and ASPP modules (Chen et al., 2017).

However, in the context of brain tumour segmentation, many methods continue to rely on naïve modality stacking or require full supervision (Myronenko, 2018; Havaei et al., 2017). These limitations have opened the door for more flexible SSL-driven approaches, which this thesis aims to build upon.

2.2 Muti-modal learning

Each MRI modality (T1, T1Gd, T2, and FLAIR) captures different tissue characteristics (Pan et al., 2024). Multi-modal learning leverages this by enabling the fusion of these diverse representations to enhance segmentation accuracy (Ronneberger et al., 2015). Typically, multi-modal methodologies either learn a unified representation space by combining all modalities or maintain separate modality-specific encodings that are later fused (Huang, 2023). In the context of brain tumour segmentation, studies have shown that by combining multiple MRI modalities, segmentation of regions such as enhancing tumour (ET), or edema (ED) has been significantly improved (Khan et al., 2023).

Recent research has explored building independent encoders for each modality or fusing multi-modal features through mechanisms such as gating or attention (Taleb et al., 2019). However, these techniques typically rely on fully supervised training and assume the availability of all modalities, which may not always be the case in real-world clinical settings.

2.3 Self-supervised Learning

Self-supervised learning (SSL) involves a “pretext task” designed to learn rich feature representations from unlabelled data, thereby reducing the need for large-scale manual annotations. Common pretext tasks studied in the literature include jigsaw puzzle solving, inpainting, and rotation prediction. In terms of medical imaging, SSL has proven beneficial when labelled data are sparse, as it encourages the network to learn representations of structure, shape, or texture from unlabelled scans (Chaitanya et al., 2020; Zhang et al., 2019).

Prior SSL work in medical imaging has typically focused on strategies such as masking random patches within the same modality or synthesizing one modality from another. These approaches have shown that SSL can be effective in learning cross-modality or contextual features that benefit downstream segmentation tasks (Taleb et al., 2019).

3 METHOD

This chapter presents the two-stage framework developed for multimodal brain tumour segmentation. The approach consists of (1) an SSL pretext task for cross-modality reconstruction and (2) a supervised segmentation model. The overall network design combines a ResNet-based encoder with a U-Net-style decoder. Supporting components such as patch-based sampling, data augmentation, and training configurations are also described.

3.1 Cross-Modality Self-Supervised Learning

The SSL phase implemented in this framework is designed to enrich feature representations before the fully supervised tumour segmentation stage. As illustrated in Figure 2, the SSL model shares a similar architecture with the main segmentation network but is configured to accept two input channels—T1 and T1Gd and tasked with reconstructing the FLAIR modality. This is achieved by slicing the input tensor as $["image"][:, :2, :, :]$ for the input and $["image"][:, 3:4, :, :]$ for the target within the training loop.

Let $x \in \mathbb{R}^{2*H*W}$ denote the concatenated T1 and T1Gd slices and let $y \in \mathbb{R}^{1*H*W}$ represent the corresponding FLAIR ground truth slice, where H and W denote the height and width of the input slices. The SSL model $F_{SSL}(\cdot)$ maps x to a reconstructed FLAIR prediction $\hat{y} \in \mathbb{R}^{1*H*W}$, defined as:

$$\hat{y} = F_{SSL}(x) = D\big(E(x)\big)$$

Here, $E(\cdot)$ represents a ResNet-50 encoder modified to handle two-channel input instead of the standard three, and $D(\cdot)$ denotes a transposed convolutional decoder that up-samples the feature maps from size $\frac{H}{32} \times \frac{W}{32}$ back to the original resolution $H \times W$. No segmentation masks are used during this phase, ensuring that the model learns in a self-supervised manner. The training objective minimizes the mean squared error (MSE) between the predicted and actual FLAIR images:

$$L_{SSL} = \sum_{i=1}^{N} \| \hat{y}^{(i)} - y^{(i)} \|^2$$

By minimizing L_{SSL} , the network is encouraged to learn the underlying fluid-attenuation patterns in the FLAIR image based solely on T1 and T1Gd inputs. This allows the encoder to capture cross-modality representations that are informative for tumour-related signal features, which are helpful in later segmentation.

Acknowledging that more challenging and diverse pretext tasks yield stronger representations, this cross-modality reconstruction task is more clinically meaningful and better aligned with the downstream objective of mapping from T1 and T1Gd to FLAIR. Prior SSL approaches often applied transformations such as independent modality flipping or patch masking. However, such techniques can disrupt the spatial alignment of tumour regions across modalities, potentially weakening the quality of the learned representations.

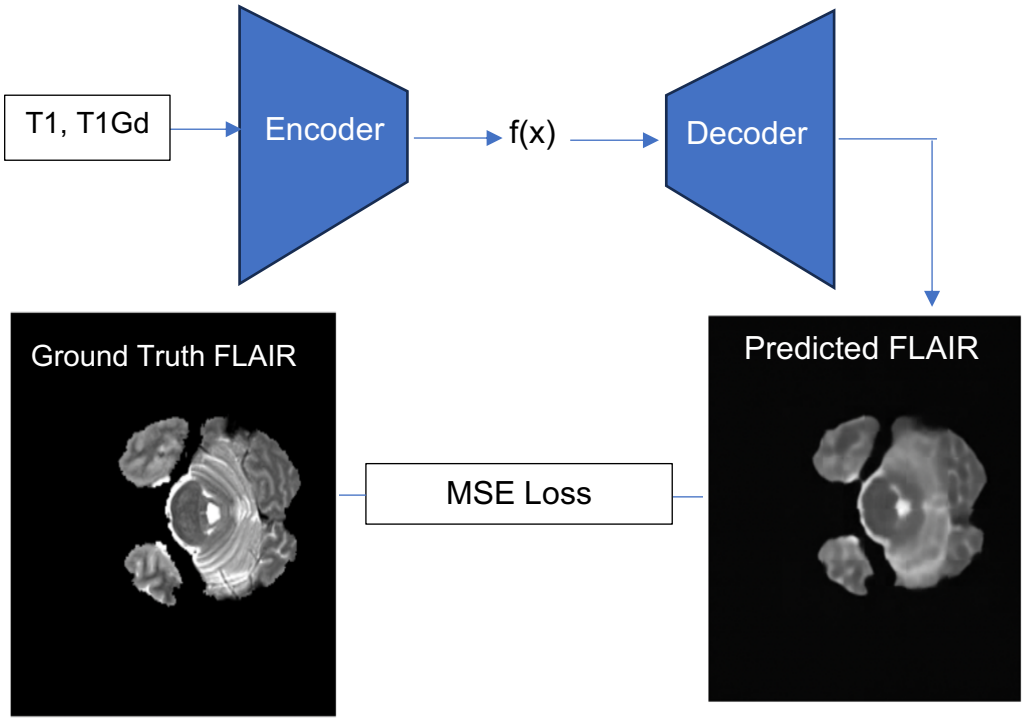

Figure 2. Overview of the SSL learning architecture for cross-modality reconstruction. The input consists of two MRI modalities (T1 and T1Gd), which are passed through a ResNet encoder and a decoder to reconstruct the FLAIR modality. The model is trained using the Mean Squared Error (MSE) loss between the predicted and ground-truth FLAIR images.

3.2 Multi-Modal Segmentation Network

To effectively leverage the rich, complementary information provided by multiple MRI modalities (T1, T1Gd, T2, FLAIR), this framework adopts a hierarchical fusion strategy based on a ResNet-50 encoder paired with a U-Net-style decoder. Each MRI modality is treated as a separate input channel, resulting in a four-channel input tensor passed to the encoder. This ensures that the network can effectively learn modality-specific details and integrate them to increase overall performance. The overall segmentation framework is illustrated in Figure 3.

The encoder is a modified ResNet 50 architecture pretrained on ImageNet, modified to accept four input channels $\mathbb{x} \in \mathbb{R}^{4*H*W}$. It extracts multi-level

feature representations at various spatial resolutions, denoted as $(c0, c1, c2, c3, c4)$, where each c_i corresponds to a progressively deeper stage in the network. This can be formally denoted as:

$$\left(c_0, c_1, c_2, c_3, c_4\right) = E(\mathbb{x})$$

Here, $E(\cdot)$ refers to the ResNet-50 encoder, which outputs features from different depth levels. These hierarchical feature maps are critical for capturing both high-level semantic features and low-level spatial details, which are essential for accurate tumour segmentation.

A key innovation of this segmentation pipeline is the *partial loading* of weights pretrained through SSL. As mentioned, during the pretext task, a ResNet-based encoder-decoder model is trained to reconstruct FLAIR modality using T1 and T1Gd as input modalities:

$$F_{SSL}\left(x^{(T1,T1Gd)}\right) \rightarrow \hat{y}^{\,(FLAIR)}$$

Once trained, the learned SSL encoder weights, which are specific to channels 0 and 1 (T1 and T1Gd), are selectively transferred to the corresponding channels of the 4-channel ResNet encoder used for segmentation. The remaining channels (T2 and FLAIR) are randomly initialized. This partial weight transfer can be described as:

$$new_{weight}[:, 0:2, :, :] \leftarrow old_{weight}$$

This strategy enables the segmentation model to leverage representations that capture cross-modality interactions between T1 and T1Gd, which are learned without supervision. The encoders become sensitive to contrast and structural patterns useful for tumour recognition, particularly in cases with limited labelled data.

The decoder follows a standard U-Net-style architecture, consisting of up-sampling operations and skip connections. At each stage of decoding, the feature map is up-sampled (by a factor of 2) and concatenated with the corresponding encoder feature map c_k :

$$u_{k-1} = concat(updample(u_k), c_k)$$

This process preserves spatial details lost during down-sampling and helps refine segmentation boundaries.

Given the multi-class nature of the segmentation task and the imbalance among tumour subregions, a combination of dice loss and focal loss is used to optimize the model. The focal loss is defined as:

$$\mathcal{L}_{focal} = -\sum_{c=1}^{C} \alpha_c \, (1 - p_t)^{\gamma} \log(p_t)$$

where p_t is the predicted probability of the true class c, γ are the focusing parameters and α_c is the class weight.

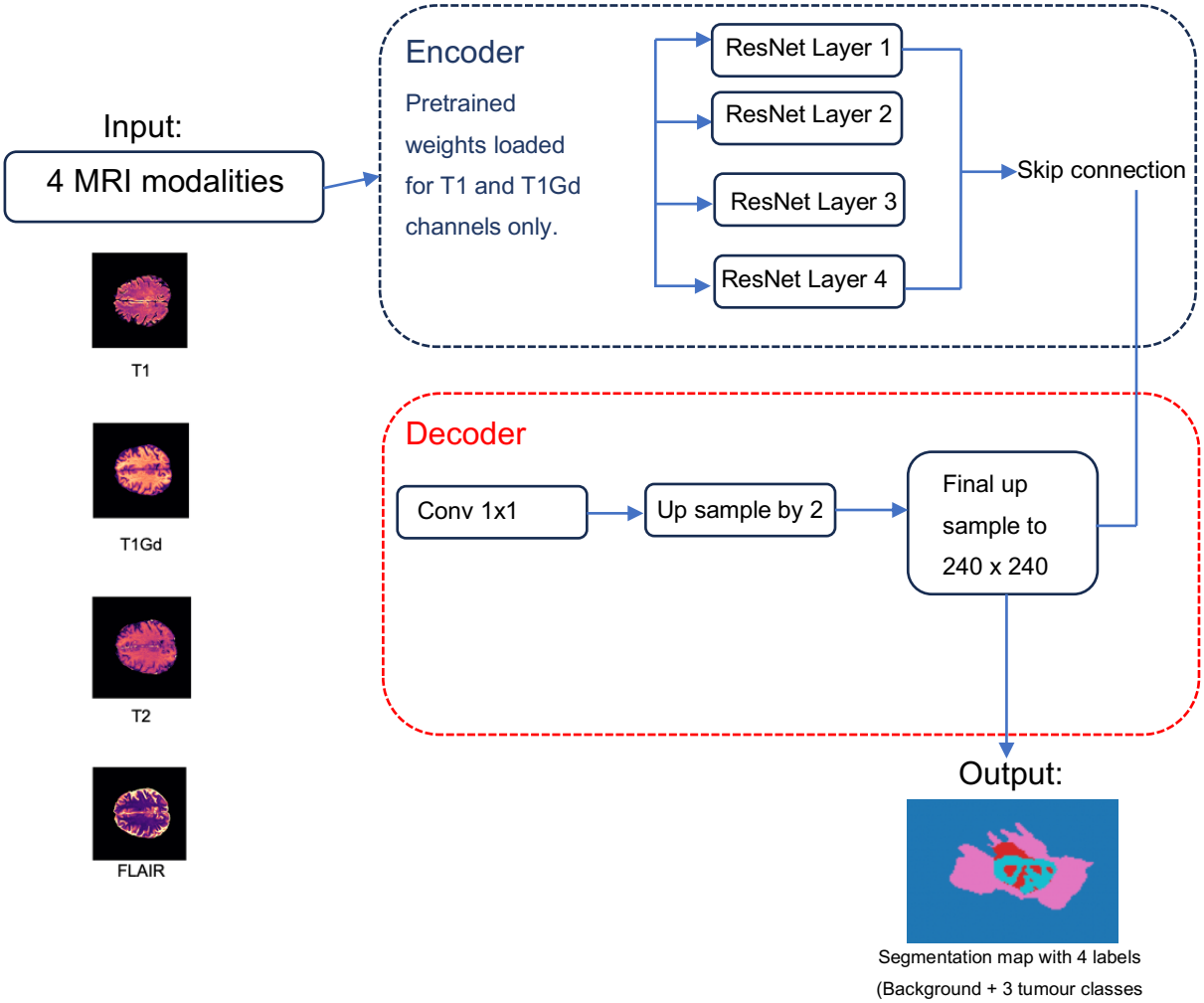

Figure 3. Overview of the segmentation architecture using a ResNet-50 encoder and U-Net-style decoder. The input consists of four MRI modalities, and the output is a multi-class segmentation map.

4 EXPERIMENTS

4.1 DATASET

The experiments in this study were conducted using the BraTS 2020 multimodal brain tumour segmentation dataset (Multimodal Brain Tumor Segmentation Challenge 2020, n.d.). It comprises 3D MRI scans from multiple institutions, with each patient volume consisting of 155 axial slices and four modalities: T1, T1Gd, T2, and FLAIR. All scans have been interpolated to a common resolution of $240 \times 240 \times 155$ voxels. Ground truth segmentation for each slice is provided, annotating three tumour subregions: necrotic and non-enhancing core (NCR/NET), peritumoral edema (ED), and enhancing tumour (ET).

The version of the dataset used in this work was sourced from Kaggle (Brain Tumor Segmentation (Brats2020), n.d.), where each subject is represented by a single 2D axial slice containing visible tumour regions. This filtering strategy was applied to reduce class imbalance, lower computational load, and ensure that training samples had relevant tumour information. Only slices with non-zero segmentation labels were retained, as reflected in the dataset metadata (volume_no, slice_no, target) (Brain Tumor Segmentation (Brats2020), n.d.).

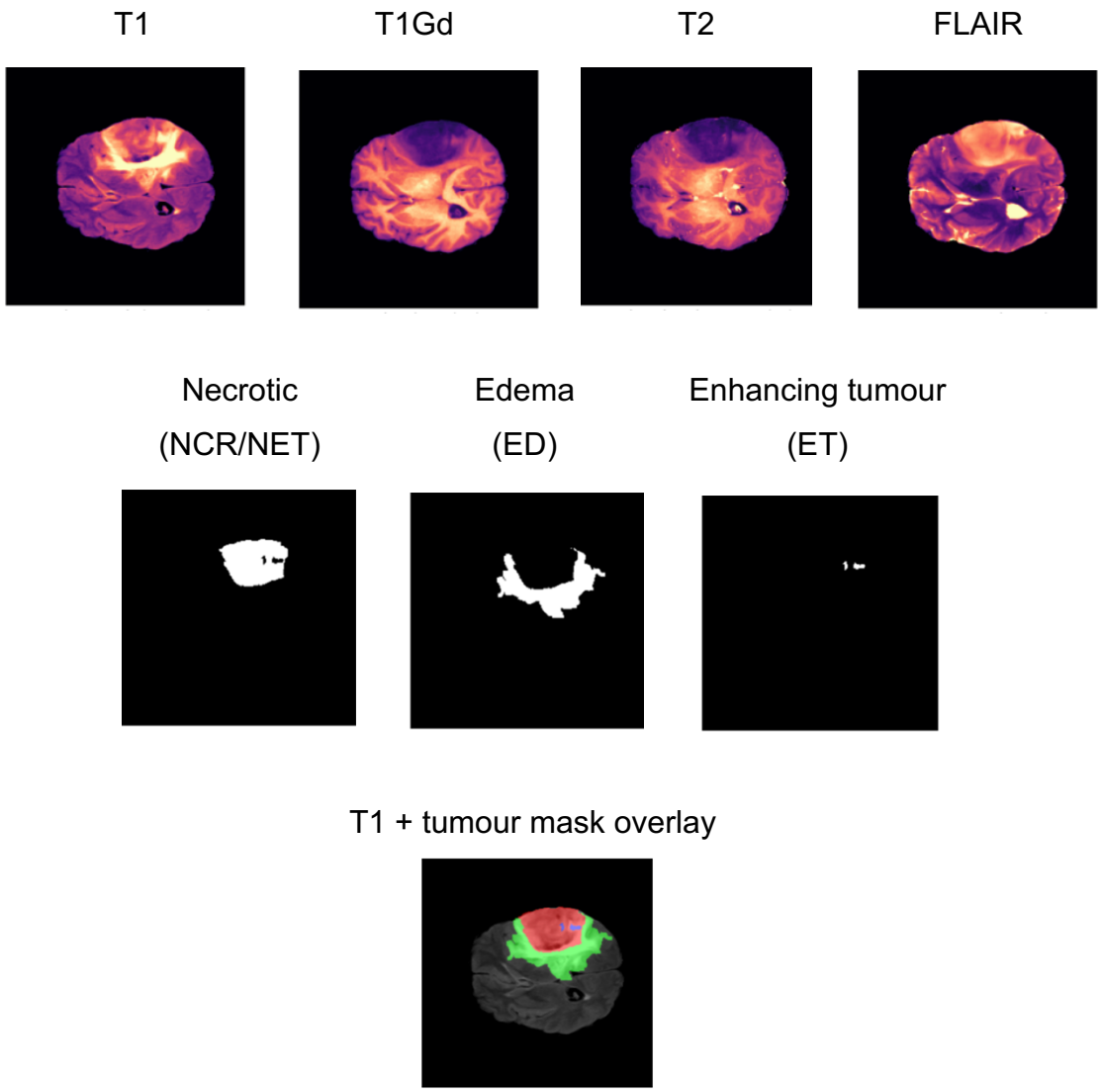

Figure 4. Visualization of an example slice from the BraTS dataset. Top: Four MRI modalities (T1, T1Gd. T2, FLAIR). Middle: Corresponding binary masks for the three tumour subregions—necrotic and non-enhancing tumour core (NCR/NET), peritumoral edema (ED), and enhancing tumour (ET). Bottom: T1 modality with RGB overlay for visualization (NCR/NET= red, ED = green, ET = blue)

The data is stored in HDF5 format, where each image tensor has the shape $[4,240,240]$, representing four modalities (T1, T1Gd, T2, FLAIR), and each label tensor has the shape [3,240,240], representing the three annotated tumour subregions as shown in Figure 4.

A challenge in the dataset is the high class imbalance due to the presence of many slices that do not contain any tumour. Training directly on all slices would bias the model toward background regions. To address this, a patch-based sampling strategy was adopted.

Instead of training on full 240x240 slices, random 2D patches of size 128×128 pixels were extracted during training. This method not only reduces memory requirements but also focuses the model on local regions that are more likely to contain relevant tumour features.

To further mitigate the class imbalance, a tumour-aware sampling strategy was implemented: if a slice contains tumour pixels, a patch including tumour pixels is sampled with probability $p = 0.5$; otherwise, patches are randomly sampled. This improves the model's ability to detect small and sparse regions such as ET and NCR/NET.

All modalities undergo Z-score normalization, which involves clipping the top 0.5% of intensity values (corresponding to the 99.5^{th} percentile) and standardizing the data to have a mean of zero and a variance of one. This helps stabilize training and reduce the influence of outliers.

The dataset has been split into training, validation, and test sets. 15% of the total files are reserved as a hold-out test set, while the remaining 85% are used in a 5-fold cross-validation scheme to ensure reliable training and evaluation splits.

4.2 IMPLEMENTATION DETAILS

All model configurations, including those with and without SSL initialization, were implemented in PyTorch and trained on a system equipped with 47 GB of RAM, 6 CPU cores, and an NVIDIA RTX A5000 (24 GB) GPU.

Both the SSL-initialized and non-SSL models were trained for 30 epochs, with a batch size of 16, using the Adam optimizer with a learning rate of 1×10^{-4}. A StepLR scheduler was applied to decay the learning rate by a factor of 0.5 every 10 epochs. The best-performing model based on validation performance was saved for evaluation.

When partial SSL loading is enabled, the first two encoder channels from the pre-trained checkpoint are initialized, while the rest are initialized randomly.

Lastly, all models adopt a ResNet-50 as the backbone and use a U-Net style decoder with skip connections for segmentation. All experiments are conducted on the BraTs2020 dataset using the train-test splits described earlier.

4.3 EVALUATION AND ANALYSIS OF SSL MODEL

4.3.1 Evaluation of Self-Supervised Learning (SSL) phase

In the SSL phase, the model was trained using a cross-modality reconstruction task, where the objective was to reconstruct the FLAIR modality from T1 and T1Gd inputs. This task was performed without using segmentation labels, allowing the model to learn rich, modality-level representations. The architectural and mathematical details of the self-supervised framework are presented in Section 3.1.

The SSL phase was conducted and evaluated using three standard image quality metrics:

1) Mean Squared Error (MSE): Measures the average squared difference between the predicted pixel values ($\hat{y}_i$) and the true pixel values (y_i) of the FLAIR modality. A lower MSE score indicates better reconstruction accuracy (Wang et al., 2019).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

2) Peak Signal-to-Noise-Ratio (PSNR): Quantifies the ratio between the maximum possible pixel intensity (MAX_I) and the error (measured by MSE), expressed in decibels (dB). A higher PSNR indicates better image quality (Wang et al., 2019).

$$PSNR = 20 \cdot \log_{10}(\frac{MAX_I}{\sqrt{MSE}})$$

3) Structural similarity index (SSIM): To consider changes in structural, information, and contrast, SSIM values range from 0 to 1, with 1 indicating perfect similarity (Wang et al., 2019).It is computed using the means (μ_x, μ_y) of the two images x and y, their variances (σ_x^2, σ_y^2), and their covariance (σ_{xy}). The constants C_1 and C_2 are used to stabilize the division and avoid numerical instabilities when the denominators are close to 0.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}$$

The SSL model was evaluated using a 5-fold cross-validation with 30 epochs each, which is consistent with the segmentation training methodology.

Quality Metric	Average
MSE	0.038
PSNR (dB)	29.770
SSIM	0.916

Table 1. Average performance metrics for the SSL reconstruction task across five folds. Metrics include MSE, PSNR, and SSIM.

4.3.2 Performance of the Self-Supervised learning model

The SSL model was assessed using a 5-fold cross-validation. As shown in Table 1, the model demonstrated strong and consistent performance across all folds. Specifically, the average MSE was 0.0381, indicating minimal reconstruction error between the predicted and the actual FLAIR modality images. The PSNR averaged 29.77 dB, suggesting low noise and high fidelity in

reconstructed images. Lastly, the SSIM score value of 0.916 confirms the model's ability to preserve key structural and contrast features.

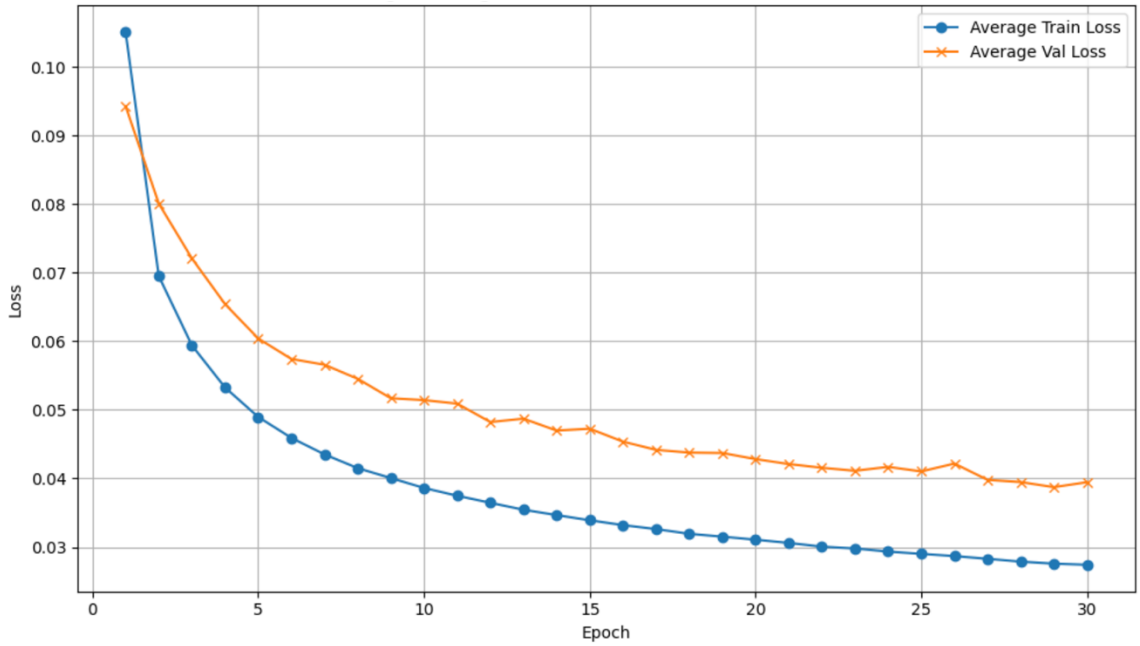

Figure 5. Average training and validation loss curves across five folds during the SSL phase

As illustrated in Figure 5, the training and validation loss curves exhibit a consistent downward trend without significant divergence, indicating stable learning and a low risk of overfitting across 30 epochs. The graph provides evidence that the SSL model has good generalization capability, reinforcing the positive evaluation findings in Table 1.

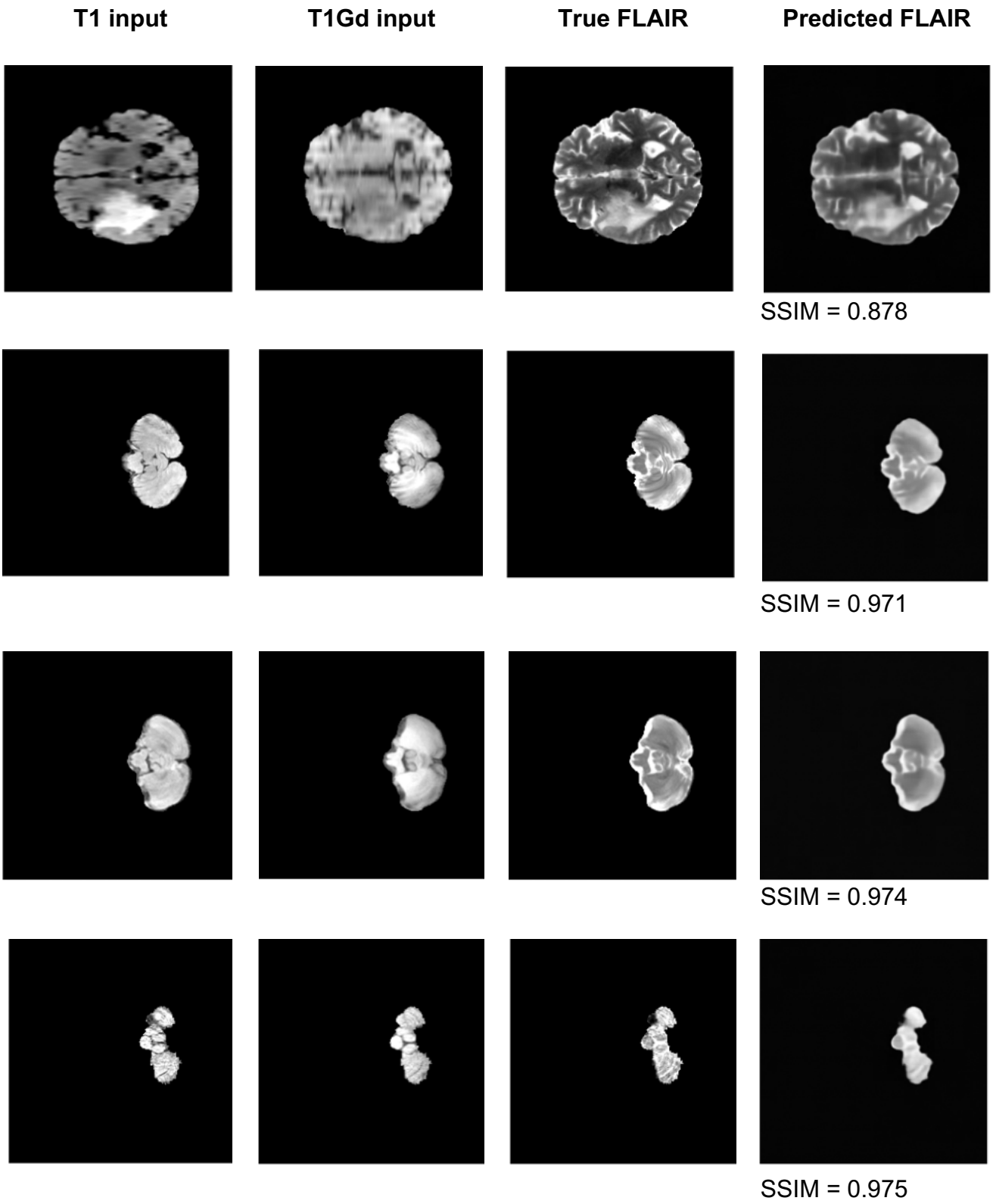

Figure 6. Qualitative visualization of self-supervised FLAIR reconstruction using T1 and T1Gd modalities as input. Each row corresponds to a different subject slice, showing (from left to right): the T1 input, T1Gd input, the ground truth FLAIR image, and the predicted FLAIR image generated by the self-

supervised model. The predicted FLAIR images have their corresponding SSIM scores, indicating the structural similarity to the ground truth.

Figure 6 visually demonstrates the performance of the SSL model in reconstructing the FLAIR modality from T1 and T1Gd inputs. Across the four visual examples, the predicted FLAIR images closely resemble the ground truth, supported by high SSIM values. This validates that the SSL model effectively captures relevant cross-modality information, which is helpful for downstream segmentation tasks.

4.4 EVALUATION AND ANALYSIS OF SEGMENTATION MODEL

4.4.1 Segmentation configurations

Following the SSL pretraining phase, the full segmentation pipeline was trained and evaluated. The experiments aimed to assess the benefits of SSL compared to random (baseline) initialization under various focal loss configurations.

Experiment ID	Self-supervised initialization	Loss Function
A-base	No	Dice+Cross-Entropy loss (1/2+1/2)
A-SSL	Yes	Dice+Cross-Entropy loss (1/2+1/2)
B-base	No	Dice +Focal Loss (1+1, γ=2)
B-SSL	Yes	Dice +Focal Loss (1+1, γ=2)
C-base	No	Focal Loss Only ($\gamma = 2$)
C-SSL	Yes	Focal Loss Only ($\gamma = 2$)

Table 2. Experimental configurations for the segmentation models. Each configuration varies according to the loss function and whether pretraining was used.

In each of these configurations mentioned in Table 2:

- Baseline (No SSL): models are randomly initialized.
- SSL-initialized: Models have partially loaded weights from the SSL phase from the first two channels of T1 and T1Gd.

- A fixed batch size of 16, a total of 30 training epochs, and a learning rate of 1e-4 were used for every configuration.

The structure of these configured experiments allowed a controlled comparison of how SSL pretraining affects segmentation performance across different loss functions.

4.4.2 Loss function selections

Due to the extreme class imbalance between background and tumour regions, careful selection of the loss functions is vital in achieving a reliable performance. Therefore, the following loss functions were evaluated:

1) Dice+Cross-Entropy loss (1/2+1/2)

- This combination leverages Cross-Entropy for stable pixel-wise classification and Dice Loss for improved sensitivity to small tumour regions. As demonstrated by Liu et al. (2021), this pairing mitigates background dominance while enhancing tumour detection.

2) Dice +Focal Loss (1+1, γ=2)

- Focal loss emphasizes hard-to-classify pixels by reducing the influence of easy ones. With γ=2, which particularly helps detect sparse enhancing tumour regions (Lin et al., 2017). Combined with Dice Loss, it further improves localization and overlap accuracy (Sudre et al., 2017).

3) Focal Loss Only ($\gamma = 2$)

- Use to assess the standalone effectiveness of Focal Loss in handling imbalance and focusing on difficult tumour pixels without Dice Loss guidance (Lin et al., 2017).

All loss functions were evaluated with and without SSL initialization. This analysis helped to determine how reliable loss functions are and the extent to which SSL improves segmentation results.

4.4.3 Evaluation metrics

For each class, the Dice coefficient and Average surface distance (ASD) have been computed. The Dice coefficient is defined as:

$$Dice = \frac{2TP}{2TP + FP + FN}$$

where TP, FP, and FN represent the number of true positives, false positives, and false negatives, respectively. It quantifies the overlap between the predicted and ground-truth segmentations.

The ASD between the predicted segmentation P and ground truth G is defined as:

$$ASD(P, G) = \frac{1}{|S_P| + |S_G|}\left(\sum_{p \in S_P} d(p, S_G) + \sum_{g \in S_G} d(g, S_P)\right)$$

here, S_P and S_G denote the set surface points on the predicted and ground truth boundaries, respectively, and $d(a, B)$ is the minimum Euclidean distance from point a to surface B. This metric evaluates the spatial discrepancy between boundaries.

Metric / Class	Baseline	SSL (pretrained)
Dice Class 0	0.999	0.999
Dice Class 1	0.182	0.184
Dice Class 2	0.310	0.314
Dice Class 3	0.194	0.195
ASD Class 0	0.500	0.180
ASD Class 1	2.430	1.760
ASD Class 2	4.280	2.460
ASD Class 3	2.660	1.170

Table 3. Comparison of average Dice coefficient and ASD across four tumour classes between the baseline and SSL pipelines, using **Focal Loss only**.

Metric / Class	Baseline	SSL (pretrained)
Dice Class 0	0.999	0.999
Dice Class 1	0.186	0.186
Dice Class 2	0.322	0.324
Dice Class 3	0.198	0.198
ASD Class 0	0.750	0.540
ASD Class 1	2.750	1.850
ASD Class 2	3.140	2.670
ASD Class 3	1.430	1.310

Table 4. Comparison of average Dice coefficient and ASD across four tumour classes between the baseline and SSL pipelines, using **Dice + Focal Loss**.

Metric / Class	Baseline	SSL (pretrained)
Dice Class 0	0.999	0.999
Dice Class 1	0.187	0.184
Dice Class 2	0.327	0.329
Dice Class 3	0.198	0.199
ASD Class 0	0.180	0.300
ASD Class 1	3.460	8.930
ASD Class 2	2.70	1.910
ASD Class 3	1.560	1.770

Table 5. Comparison of average Dice coefficient and ASD across four tumour classes between the baseline and SSL pipelines, using **Dice + CrossEntropy Loss**.

Method	Supervision	NCR/NET (Class 1)	ED (Class 2)	ET (Class 3)	Average Dice Score
Dice+Focal	Fully Supervised (No SSL)	0.186	0.323	0.198	0.236
Dice+Focal	Self-supervised (SSL)	**0.187**	**0.324**	0.198	0.236
Focal Only	Fully Supervised (No SSL)	0.182	0.310	0.194	0.228
Focal Only	Self-supervised (SSL)	**0.184**	**0.314**	**0.195**	**0.230**
Dice+CE	Fully Supervised (No SSL)	**0.188**	0.330	0.190	**0.238**
Dice+CE	Self-supervised (SSL)	0.184	**0.330**	**0.199**	0.238

Table 6. Comprehensive evaluation of segmentation performance based on average Dice scores of fully supervised learning and Self-supervised learning. Best results are highlighted.

4.4.4 Performance on multi-modal brain tumour segmentation

This section evaluates the proposed multi-modal brain tumour segmentation framework, focusing on its performance when initialized with SSL weights compared to random initialization (no-SSL). To provide a comprehensive assessment, three loss function configurations were tested: **Dice + Focal loss, Focal loss only**, and **Dice + CE Loss**. For each configuration, the model was trained twice, once with SSL and once without SSL, using a 5-fold cross-validation strategy.

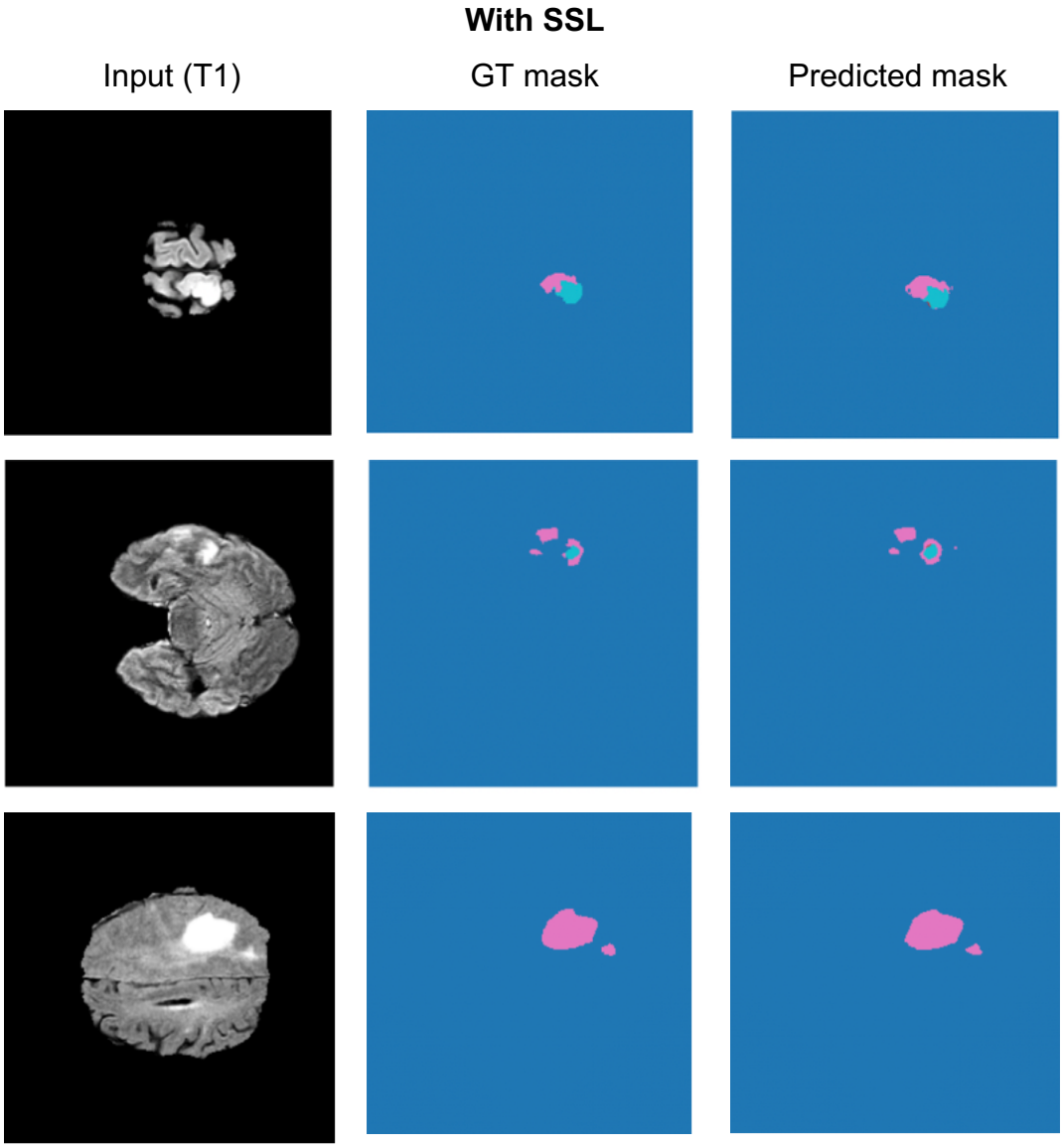

Figure 7. Visual comparison of segmentation predictions using SSL initialization. The first row shows results from a model trained with Focal Loss only, the second with Dice + Focal Loss, and the third with Dice + CE Loss. The columns show: (1) the input T1 MRI, (2) the ground truth mask, and (3) the predicted segmentation mask. Whole tumour (WT) is in pink, tumour core (TC) is in turquoise, and enhancing tumour (ET) is red.

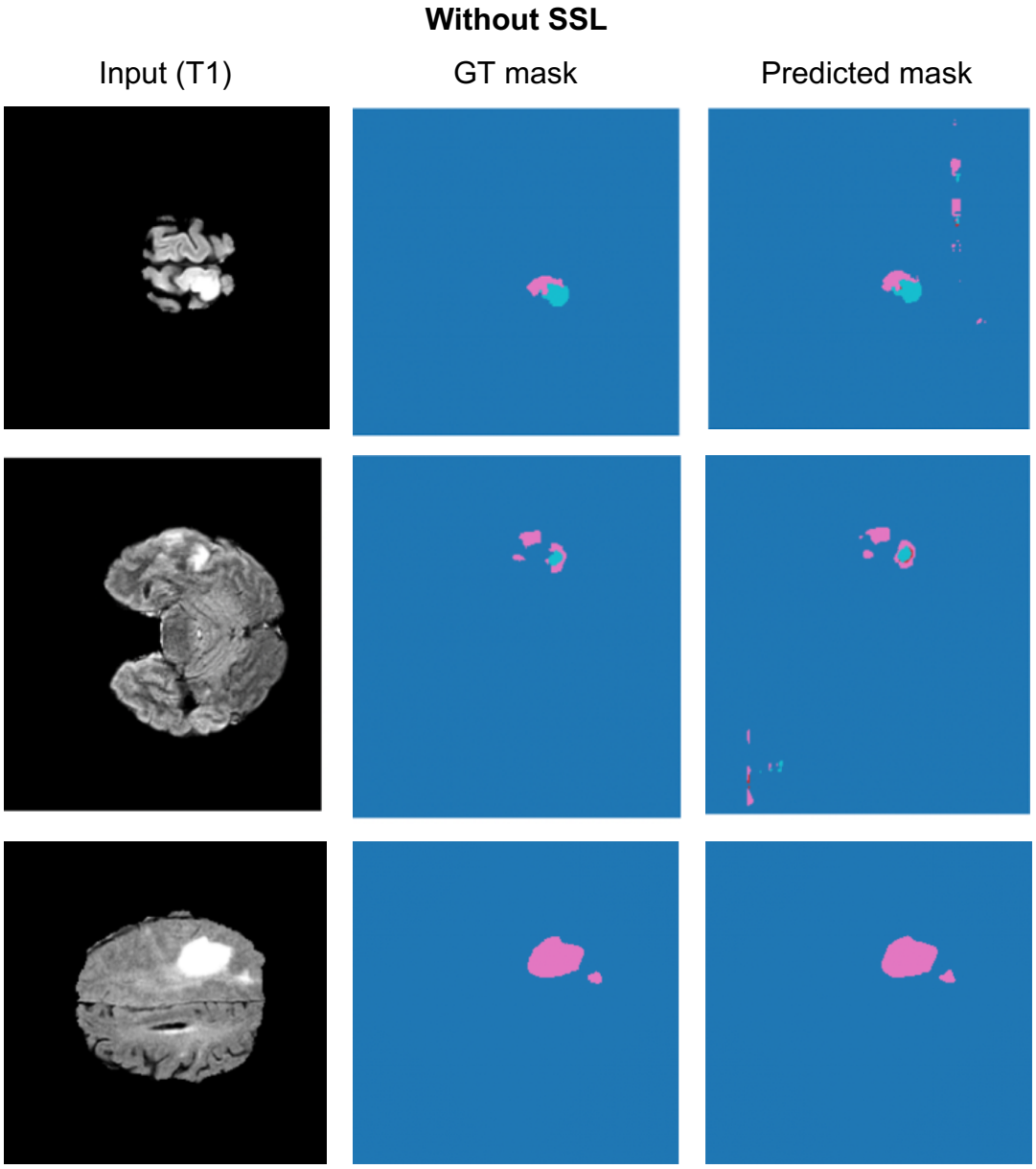

Figure 8. Visual comparison of segmentation predictions without self-supervised initialization. The same inputs and training configurations in Figure 7 are used to highlight the performance difference across loss functions.

The segmentation task aims to differentiate the following tumour subregions: necrotic and non-enhancing tumour core (class 1), peritumoral edema (class 2), and enhancing tumour (class 3). Tables 3-6 summarize the average Dice scores and the ASD results across each tumour class. Figures 7 and 8 present qualitative segmentation results comparing models trained with and without SSL. Each row corresponds to a different loss configuration—Focal

loss only, Dice + Focal Loss, and Dice + CE Loss—applied to the same input slices, presenting a visual comparison of tumour delineation performance.

Referencing tables 3 to 5, the use of SSL weights consistently resulted in improvements in nearly all classes compared to random weight initialization. Specifically:

- As shown in Table 3, in the Focal Loss only scenario, SSL assisted in improving the dice scores slightly for all classes, resulting in a higher Mean Dice score, which increased from 0.2286 to 0.2305.
- In the Dice + Focal loss configuration, the SSL weights showed even more pronounced improvements, notably for class 2 (Edema), where the Dice score increased from 0.3228 to 0.3244, with a slight but consistent improvement.
- With Dice + CE, SSL provided the highest increase in overall segmentation performance (see Table 4, Appendix). Edema (Class 2) again showed improvement.

The results show that SSL-initialized weights provide consistent benefits, particularly when the segmentation task is challenged by an imbalanced dataset or one containing small tumour subregions, such as those found in NCR/NET or ET.

The Average Surface Distance (ASD) measures the boundary error between the predicted and ground-truth segmentations. Lower values indicate more accurate boundaries. Tables 3-5 show that SSL generally reduced ASD across all folds, especially for the more challenging classes:

- Focal Loss only: The average ASD for class 2 dropped from 4.28 mm to 2.46 mm with SSL, indicating a significant improvement in boundary accuracy.
- Dice + Focal loss: ASD for class 2 improved from 3.14 to 2.67 mm with SSL.

However, Table 4 reveals a vital exception (see Table 4, Appendix). In Fold 4 of the SSL pipeline, the ASD for Class 1 spiked significantly to 37.610 mm, a clear outlier compared to the other folds. This anomaly stems from a case where the model failed to predict any or very little of the tumour core, resulting in a large spatial discrepancy from the ground truth. Despite this outlier, the remaining folds show consistent improvements in boundary alignment.

Given that the BraTS 2020 dataset is resampled to an isotropic voxel spacing of 1 mm^3 (Bakas et al., 2017), each voxel corresponds to 1mm in anatomical space. Thus, an ASD of ~2-3 mm translates to a 2–3-pixel average boundary error, which aligns with values commonly reported in the literature and is generally regarded as acceptable for clinical-grade segmentation performance in brain tumour studies (Taha & Hanbury, 2015).

The results demonstrate that SSL initialization not only improves Dice but also yields more precise and smoother tumour boundaries. Its benefits are especially evident in scenarios involving class imbalance and small tumour subregions, such as NCR/NET and ET.

4.5 Limitations

While the use of SSL appears to improve the spatial coherence and alignment of the predicted masks with the ground truth (Figures 7 and 8), particularly in larger tumour regions, several limitations remain. Across both tables, the models occasionally struggle to accurately capture small and disconnected tumour subregions, such as TC and ET, as supported by the lower Dice scores and higher ASD values reported for these classes in tables 6-9 (Havaei et al., 2017; Taha & Hanbury, 2015). These challenges are attributed to the inherent class imbalance in the dataset and the complex boundary characteristics of smaller tumour structures (Sudre et al., 2017).

Furthermore, the SSL pretext task employed in this work focused on cross-modality prediction, mainly to predict FLAIR from T1 and T1Gd images. While

this approach encourages modality interaction, it may have limited the model's capacity to learn modality-specific features, which are crucial for segmenting small or subtle regions (Taleb et al., 2019; Zhou et al., 2019).

Finally, the multi-modal fusion strategy used was a straightforward channel-wise concatenation, which may be suboptimal for capturing complex inter-modal relationships. Integrating more advanced fusion methods, such as attention mechanisms or modality-specific encoders, could potentially improve segmentation accuracy (Khan et al., 2023).

5 CONCLUSION

This study explored the application of Self-Supervised Learning (SSL) to enhance multi-modal brain tumour segmentation using the BraTS 2020 dataset. A simple channel-wise concatenation approach was employed to integrate the four MRI modalities as input to a ResNet-based U-Net segmentation architecture.

The framework was evaluated across multiple configurations: **Dice + Focal loss, Focal Loss only,** and **Dice + CE loss**, with and without SSL-based weight initialization. Across all settings, models initialized with SSL consistently outperformed their randomly initialized counterparts in terms of Dice scores and ASD, particularly for challenging tumour subregions such as ED and ET.

Although the improvements were modest, the results demonstrate the potential of SSL to offer meaningful performance gains, even when used in conjunction with simple fusion strategies. These findings validate the feasibility of using SSL as a pretraining method and provide a solid foundation for future work. Further research may benefit from incorporating more sophisticated fusion techniques, such as attention mechanisms or transformer-based

architectures, and exploring advanced SSL objectives tailored to the unique characteristics of medical imaging data.

APPENDIX

Fold	MSE	PSNR (dB)	SSIM
1	0.038	**30.270**	**0.929**
2	0.039	29.660	0.921
3	0.038	29.290	0.893
4	**0.038**	29.850	0.921
5	0.038	29.790	0.918

Table 1. Fold-wise SSL reconstruction performance using MSE, PSNR, and SSIM metrics. The best results are highlighted.

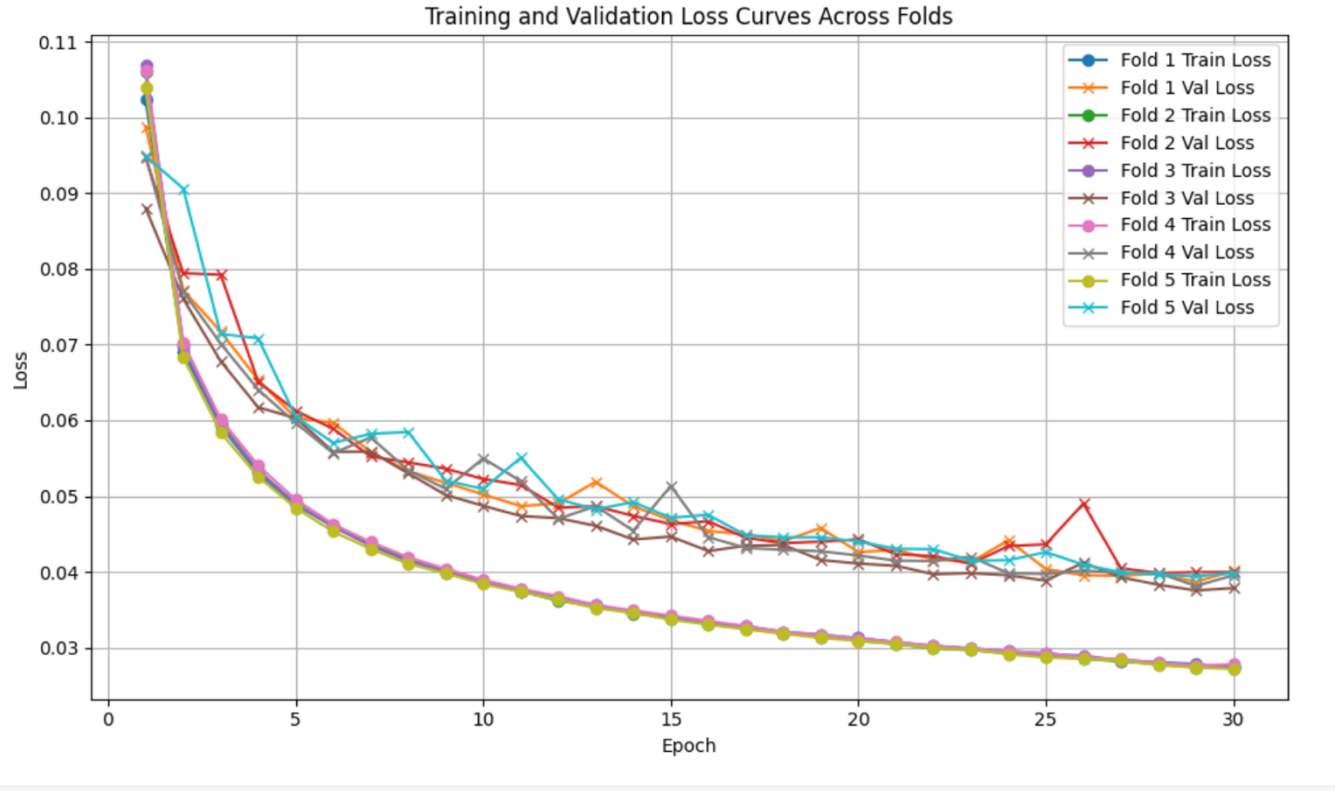

Figure 1. Training and validation loss curves across five folds for the segmentation model. Each line represents the loss trend per epoch for a specific fold.

Baseline Pipeline								
Fold	Dice Class 0	Dice class 1	Dice Class 2	Dice Class 3	ASD Class 0	ASD Class 1	ASD Class 2	ASD Class 3
1	0.998	0.181	0.310	0.194	1.350	2.760	6.010	2.860
2	0.999	0.181	0.310	0.195	0.200	2.070	2.410	1.130
3	0.999	0.184	0.312	0.194	0.190	1.650	2.440	1.270
4	0.999	0.183	0.314	0.194	0.200	1.710	2.320	1.140
5	0.999	0.181	0.304	0.191	0.580	3.950	8.240	6.880
SSL pipeline								
Fold	Dice Class 0	Dice class 1	Dice Class 2	Dice Class 3	ASD Class 0	ASD Class 1	ASD Class 2	ASD Class 3
1	0.999	0.185	0.316	0.194	0.190	1.830	2.740	1.270
2	0.999	0.182	0.311	0.194	0.180	1.710	2.540	1.200
3	0.999	0.185	0.314	0.195	0.170	1.880	2.370	1.120
4	0.999	0.182	0.314	0.195	0.170	1.700	2.240	1.080
5	0.999	0.184	0.312	0.194	0.200	1.690	2.420	1.190

Table 2. Comparison of Dice coefficient and ASD scores per fold across four tumour classes between the baseline and SSL pipelines, using Focal Loss only.

Baseline Pipeline								
Fold	Dice Class 0	Dice class 1	Dice Class 2	Dice Class 3	ASD Class 0	ASD Class 1	ASD Class 2	ASD Class 3
1	0.999	0.185	0.321	0.198	0.410	2.150	4.020	1.830
2	0.999	0.187	0.324	0.198	0.420	3.970	2.900	1.360
3	0.999	0.186	0.324	0.199	0.360	1.850	2.830	1.290
4	0.999	0.186	0.321	0.199	0.420	3.860	3.300	1.330
5	0.998	0.185	0.324	0.199	2.140	1.940	2.640	1.330
SSL pipeline								
Fold	Dice Class 0	Dice class 1	Dice Class 2	Dice Class 3	ASD Class 0	ASD Class 1	ASD Class 2	ASD Class 3
1	0.999	0.186	0.325	0.199	0.330	1.840	2.630	1.320
2	0.999	0.188	0.324	0.198	0.790	1.900	2.710	1.340
3	0.999	0.187	0.325	0.197	0.390	1.840	2.840	1.300
4	0.999	0.187	0.324	0.198	0.600	1.860	2.700	1.300
5	0.999	0.185	0.324	0.198	0.310	1.840	2.450	1.290

Table 3. Comparison of Dice coefficient and ASD scores per fold across four tumour classes between the baseline and SSL pipelines, using Dice + Focal Loss.

Baseline Pipeline								
Fold	Dice Class 0	Dice class 1	Dice Class 2	Dice Class 3	ASD Class 0	ASD Class 1	ASD Class 2	ASD Class 3
1	0.999	0.187	0.329	0.199	0.150	1.830	1.850	1.170
2	0.999	0.190	0.327	0.199	0.140	1.840	2.020	1.140
3	0.999	0.183	0.324	0.197	0.280	9.630	5.250	3.170
4	0.999	0.188	0.328	0.199	0.160	1.820	2.020	1.170
5	0.998	0.187	0.33	0.199	0.190	2.180	2.360	1.150
SSL pipeline								
Fold	Dice Class 0	Dice class 1	Dice Class 2	Dice Class 3	ASD Class 0	ASD Class 1	ASD Class 2	ASD Class 3
1	0.999	0.189	0.330	0.200	0.140	1.750	1.860	1.120
2	0.999	0.189	0.329	0.200	0.140	1.710	1.910	1.140
3	0.999	0.188	0.329	0.198	0.130	1.690	1.620	1.160
4	0.999	0.164	0.330	0.199	0.920	37.610	2.150	4.070
5	0.999	0.188	0.329	0.200	0.170	1.870	1.960	1.350

Table 4. Comparison of Dice coefficient and ASD scores per fold across four tumour classes between the baseline and SSL pipelines, using Dice + CrossEntropy Loss.

REFERENCES

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., ... & Norouzi, M. (2021). Big self-supervised models advance medical image classification. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3478–3488. https://arxiv.org/abs/2101.05224

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., ... & Davatzikos, C. (2017). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4, 170117. https://doi.org/10.1038/sdata.2017.117

Bauer, S., Wiest, R., Nolte, L. P., & Reyes, M. (2013). A survey of MRI-based medical image analysis for brain tumor studies. *Physics in Medicine and Biology*, 58(13), R97–R129. https://doi.org/10.1088/0031-9155/58/13/R97

Case Western Reserve University School of Medicine. (2016). *MRI basics*. https://case.edu/med/neurology/NR/MRI%20Basics.htm

Chaitanya, K., Erdil, E., Karani, N., & Konukoglu, E. (2020). Contrastive learning of global and local features for medical image segmentation with limited annotations. *NeurIPS Workshop on Medical Imaging with Deep Learning*. https://arxiv.org/abs/2006.10511

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

Chen, L.-C. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*. https://arxiv.org/abs/1802.02611

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., ... & Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35, 18–31. https://doi.org/10.1016/j.media.2016.05.004

Huang, S.-C., Pareek, A., Lungren, M. P., Jensen, M., Yeung, S., & Chaudhari, A. S. (2023). Self-supervised learning for medical image classification: A systematic review and implementation guidelines. *NPJ Digital Medicine*, 6(1), 97. https://www.nature.com/articles/s41746-023-00811-0

Khan, A. M., Ashrafee, A., Khan, F. S., Hasan, M. B., & Kabir, M. H. (2023). AttResDU-Net: Medical image segmentation using attention-based residual double U-Net. *arXiv preprint arXiv:2306.14255*. https://arxiv.org/abs/2306.14255

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980–2988. https://doi.org/10.1109/ICCV.2017.324

Liu, F., You, C., Wu, X., Ge, S., Wang, S., & Sun, X. (2021). Auto-encoding knowledge graph for unsupervised medical report generation. *arXiv preprint arXiv:2111.04318*. https://arxiv.org/abs/2111.04318

Myronenko, A. (2018). 3D MRI brain tumor segmentation using autoencoder regularization. *arXiv preprint arXiv:1810.11654*. https://arxiv.org/abs/1810.11654

National Institute of Biomedical Imaging and Bioengineering. (n.d.). *Magnetic resonance imaging (MRI)*. https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri

Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018). Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*. https://arxiv.org/abs/1804.03999

Pan, J., Chen, Q., Sun, C., Liang, R., Bian, J., & Xu, J. (2024). *MRISeqClassifier: A deep learning toolkit for precise MRI sequence classification*. medRxiv. https://doi.org/10.1101/2024.09.19.24313976

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer. https://arxiv.org/abs/1505.04597

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. *MICCAI 2017*. https://arxiv.org/abs/1707.03237

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, 15(1), 29. https://doi.org/10.1186/s12880-015-0068-x

Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., & Lippert, C. (2021). 3D self-supervised methods for medical imaging. *NeurIPS Workshops*. https://arxiv.org/abs/2006.03829

Taleb, A., Lippert, C., Klein, T., & Nabi, M. (2019). Multimodal self-supervised learning for medical image analysis. *arXiv preprint arXiv:1912.05396*. https://arxiv.org/abs/1912.05396

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., & Vercauteren, T. (2019). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338, 34–45. https://doi.org/10.1016/j.neucom.2019.01.103

Zhou, Z., Sodha, V., Pang, J., Shen, W., & Fishman, E. (2019). Models Genesis: Generic autodidactic models for 3D medical image analysis. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020*. https://arxiv.org/abs/1908.06912