MIM-HeMIS: Self-Supervised Masked Image Modeling with Heterogeneous Modality Fusion for Brain Tumor Segmentation under Missing MRI Modalities

by

Md. Jahidul Hasan
22301146
Raduan Ahmed Opy
21301604
Md Alif Khan Shafi
21301341
Ahnaf Ashraf Jameel
21301234
Md. Jobayer Alvi
21301310

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
June 2026.

Declaration

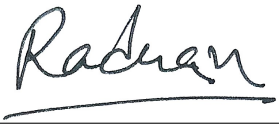
It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Jahidul Hasan

22301146



Raduan Ahmed Opy

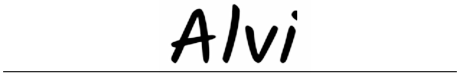
2130604

Md Alif Khan Shafi

21301341

Ahnaf Ashraf Jameel

21301234



Md. Jobayer Alvi

21301310

Approval

The thesis titled "MIM-HeMIS: Self-Supervised Masked Image Modeling with Heterogeneous Modality Fusion for Brain Tumor Segmentation under Missing MRI Modalities" submitted by

1. Md. Jahidul Hasan (22301146)
2. Raduan Ahmed Opy (21301604)
3. Md Alif Khan Shafi (21301341)
4. Ahnaf Ashraf Jameel (21301234)
5. Md. Jobayer Alvi (21301310)

of Spring, 2026 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in 17th June, 2026.

Examining Committee:

Supervisor:
(Member)

Md. Sabbir Ahmed

Senior Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor, Thesis Coordinator
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor & Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

Accurate segmentation of regions of interest from MRI to diagnose brain tumors is crucial. MRI is the standard technique used in imaging the brain, but manual segmentation of tumor regions is time-consuming, manual and prone to human error. Although deep learning has been able to significantly enhance segmentation performance, it requires big data sets of labeled images which is difficult in medical applications. In this study, we introduce a hybrid CNN-Transformer framework that integrates the Masked Image Modeling (MIM) based self-supervised pretraining and the HeMIS statistical abstraction mechanism. A hybrid model is pretrained on the BraTS 2021 dataset and finetuned for 4-class brain tumor segmentation for all 4 MRI modalities following the proposed framework. A major advantage is that the modalities are robust, meaning that there is no need to train a new model for each of the 15 possible combinations of modalities. When modalities are missing, the variance computed by the HeMIS abstraction layer captures feature disagreement across available inputs, providing useful information for identifying cases where predictions may be less reliable. Though this variance has not been calibrated as a clinical confidence score. The model was tested on BraTS 2021 with all 15 modality combinations with excellent and clinically applicable results: Whole Tumor Dice of 90.2%, Tumor Core Dice of 85.9%, and Enhancing Tumor Dice of 79.1% for all four modalities.

Keywords: Brain Tumor Segmentation, Self-Supervised Learning, Masked Image Modeling, Magnetic Resonance Imaging, Missing Modality Robustness, HeMIS, CNN-Transformer Hybrid, Deep Learning, Machine Learning, Vision Transformers, BraTS 2021, Multimodal MRI, Medical Image Analysis, HeMIS Variance Analysis, 3D Segmentation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Brain tumor segmentation in MRI is an important task in medical image analysis. It helps identify the tumor regions in the brain. It also helps measure how large the tumor is. This information is useful for planning proper treatment. In clinical practice, radiologists usually examine different MRI sequences. These sequences help them understand different features of the tumor. T1-weighted MRI, contrast-enhanced T1-weighted MRI (T1ce), T2-weighted MRI, and Fluid-Attenuated Inversion Recovery (FLAIR) are the common MRI modalities used in brain tumor analysis. Different information about brain structure and tumor region is provided by each modality. For instance, FLAIR can be used to detect edema, T1ce can be used to visualize enhancing regions of the tumor, T2 can be used to visualize fluid-sensitive abnormalities, and T1 can be used to visualize anatomical structure.

Manual segmentation of brain tumors is time consuming, labour intensive and requires expert radiologists. The process may also differ among the experts, since the edges of the tumor are not always smooth and well-defined. This has led to a significant research focus in medical artificial intelligence, namely in segmenting brain tumors automatically. The potential for the analysis of 3D MRI volumes using deep learning models, particularly encoder-decoder based segmentation networks is very promising. Most of these models, however, necessitate a large number of labeled data and full multimodal MRI inputs, which may not be always available in real clinical settings.

In a real world scenario, a hospital might not always have all modalities for all patients in its MRI. Limited scanning time, patient mobility, time constraints, cost, machine availability, and/or differences in imaging protocols across hospitals can cause some modalities to be missing. This is particularly relevant in clinical settings with low resources where full imaging sequences and expert annotations might be unavailable. Hence, a useful brain tumor segmentation system should not only be effective on a complete research set of data, but also be useful in the case of incomplete modality availability. In order to address these challenges, this research aims at a self-supervised and modality-aware brain tumor segmentation framework. The proposed work uses self-supervised learning to learn meaningful 3D MRI features by reconstructing original MRI patches before supervised segmentation training. It

also integrates a HeMIS-inspired fusion idea inside an encoder-decoder framework to improve robustness under missing-modality conditions. The aim is to develop a more practical segmentation system that can support brain tumor analysis using BraTS 2021 multimodal MRI data. This real-world motivation is approximated through simulated missing-modality experiments on BraTS 2021, not through real hospital missing-data cases.

1.2 Rational of the Study or Motivation

This study is motivated by two main shortcomings of the existing medical image segmentation systems, namely, their reliance on labelled training sets and their requirement of full MRI modalities. In medical imaging, it takes a lot of time and expertise to collect annotations at the level of the individual pixel or voxel in the tumor. This is costly and time-consuming to prepare large datasets with annotations. While there are many high-quality datasets that can be used as benchmarks like BraTS, real hospitals typically have many MRI scans that may not be fully annotated. Thus, supervised learning cannot be sufficient for the application of deep learning systems in medical settings.

One approach to alleviate this reliance on labeled data is self-supervised learning. The model first learns from a pretext task to learn useful image representations from MRI volumes, rather than learning only from manually annotated segmentation masks. This research concentrates on the SSL model that learns to reconstruct the original MRI patches. This way, the encoder can learn structural and contextual patterns of a 3D brain MRI prior to the last training stage in segmentation. This leads to better start-up of the model for supervised learning with less labeled medical data.

The missing-modality issue is another motivation. The majority of brain tumor segmentation models that are used for training and testing are trained on all four MRI modalities: T1, T1ce, T2, and FLAIR. But the actual clinical information is often limited. Some models will not operate or become unreliable without all modalities. This decreases the model's usefulness in real hospitals. To address this issue, a modality-aware fusion approach inspired by HeMIS is introduced to enable the flexibility of the framework to fuse information from available modalities.

The motivation is to study a more modality-flexible research framework under benchmark-based missing-modality simulation. In the proposed HeMIS-SSL Fusion Hybrid Framework, the self-supervised representation learning and missing-modality-aware fusion are integrated. This not only has implications for the technical aspects of the study but has a real-world healthcare context, particularly within hospitals with limited imaging resources.

1.3 Problem Statement

Segmentation of brain tumors in 3D multimodal MRI is a challenging computing and medical image analysis task. The current deep learning-based segmentation

models perform well when the data set is ideal, but they rely on two assumptions: large labeled data sets and complete MRI modalities. These assumptions are not necessarily true in clinical practice.

There are three problem areas that define the core research bottlenecks:

1. **Scarcity of Labeled Data:** The limited availability of labeled medical imaging data. Supervised model training is costly and challenging due to the need for expert radiologists to annotate voxels for each tumor.

2. **Incomplete Modality Acquisitions:** The absence of any of the MRI modalities. While the general characteristics of BraTS-type research datasets include (T1, T1ce, T2, FLAIR), all four modalities are not captured and stored for all patients in hospitals. Consequently, models that need all the input channels are less useful in actual situations.

3. **Anatomical and Structural Complexity:** The complexity of the 3D tumor segmentation itself. Brain tumors can be quite variable in shape, size, location, intensity and subregion structure, making the segmentation of the whole tumor, tumor core and enhancing tumor very challenging.

Therefore, the underlying research problem can be stated as follows:

How can a robust 3D brain tumor segmentation framework be developed that uses self-supervised learning to improve feature representation and HeMIS-inspired modality fusion to support segmentation under missing-modality conditions?

To address this problem, this research proposes a HeMIS-SSL Fusion Hybrid Framework using BraTS 2021 multimodal MRI data. The framework uses all four MRI modalities during the main training process and evaluates missing-modality robustness after training. The model is designed using an encoder-decoder structure, where self-supervised pretraining reconstructs original MRI patches and supervised fine-tuning performs multi-class tumor segmentation.

Therefore, a robust and data-efficient 3D brain tumor segmentation framework is needed that can learn useful MRI representations through self-supervised learning and remain applicable when the availability of MRI modality is incomplete in remain robust under simulated missing-modality conditions using BraTS 2021.

1.4 Objective

The main objective of this research is to develop a robust 3D brain tumor segmentation framework that combines self-supervised learning and HeMIS-inspired modality fusion to improve segmentation performance and missing-modality robustness using BraTS 2021 MRI data.

To achieve this overarching goal, the study is divided into the following specific, technical objectives:

- **Architectural Development:** To develop an encoder-decoder-based 3D segmentation model for brain tumor segmentation using the BraTS 2021 dataset.

- **Multi-Modal Input Integration:** To use all four MRI modalities (T1, T1ce, T2, and FLAIR) as input sources for learning complementary and multi-scale pathological tumor information.
- **Representation Learning via SSL:** To apply self-supervised learning through masked MRI patch reconstruction so that the model can extract and learn useful volumetric representations before supervised segmentation training.
- **Robust Fusion Layer Modeling:** To integrate a HeMIS-inspired modality fusion mechanism within the proposed framework for handling modality variation and evaluating model robustness when sequences are missing after training.
- **Hierarchical Subregion Delineation:** To segment clinically relevant tumor regions, specifically including the whole tumor (WT), tumor core (TC), and enhancing tumor (ET) structures.
- **Quantitative Benchmarking:** To evaluate the overall segmentation performance using the Dice score and quantitatively compare the effectiveness of the proposed framework under both complete and missing-modality conditions.

1.5 Methodology in Brief

The BraTS 2021 dataset is mainly used in this research for the experiments. Multimodal brain MRI scans are available in four modalities: T1, T1ce, T2, and FLAIR. These modalities are employed since each offers complementary anatomical and pathological information for the identification of regions in the brain tumor. The whole tumor (WT), tumor core (TC), and enhancing tumor (ET) are the segmentation targets.

1.5.1 Data pre-processing and Input Patch Extraction

The methodology starts with pre-processing of the 3D MRI volumes. The MRI modalities are organized, aligned, normalized, and prepared for training. The 3D MRI volumes have high computational memory requirements, therefore the processing approach is patch-based to make the training process more manageable. The model takes in multimodal MRI patches and learns to generate voxel-wise tumor segmentation masks.

1.5.2 Two-Stage Training Strategy

The training pipeline is divided into two distinct sequential stages:

1. **Stage 1 – Self-Supervised Pretraining:** Self-supervised learning (SSL) is used for initial pretraining. In this step, some of the MRI input is masked and the model is trained to recover the original MRI patches using a Masked Image Modeling (MIM) objective. This enables the encoder to gain valuable structural and contextual information from the MRI data prior to the segmentation labels. For the experiment under consideration, the SSL pretraining loss

is given by the reconstruction behavior learned by the model was meaningful and the reconstruction error was around 0.4874.

2. **Stage 2 – Supervised Fine-Tuning:** The pretrained weights of the encoder are transferred to the segmentation model for supervised fine-tuning. The segmentation model adopts encoder-decoder structure and incorporates a layer based on HeMIS in the model for modality-aware feature fusion. All four MRI modalities are used for training the model, which is then evaluated completely in the absence of modalities after training. This enables the study to investigate the ability of the learned representations and fusion strategy to enable robust segmentation in the absence of some of the input modalities.

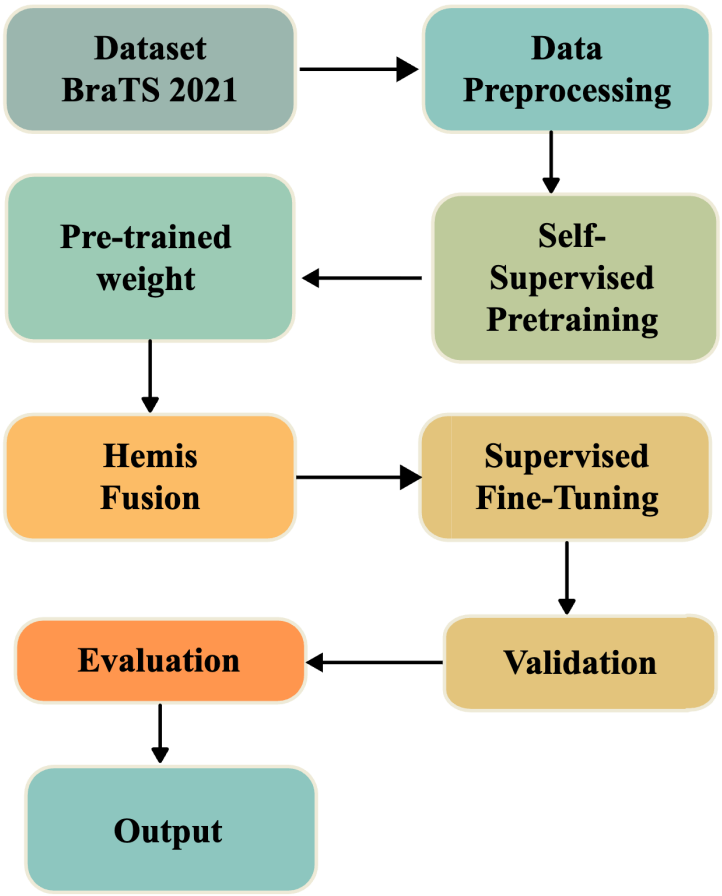


Figure 1.1: Methodology

1.5.3 Performance Evaluation Metrics

The standard Dice similarity score is used to assess the final segmentation performance. This metric measures the volumetric overlap between the tumor masks predicted by the model and the ground truth tumor regions drawn by experts, evaluating overall model accuracy.

The methodology thus integrates self-supervised MRI patch reconstruction, encoder-decoder segmentation and modality fusion based on HeMIS (modalities inspired

by the human eye) to overcome real-world challenges of limited labeled data and incomplete clinical MRI modalities.

1.6 Scopes and Challenges

This study is limited to the segmentation of brain tumors in 3D from the BraTS 2021 dataset. All four MRI modalities (T1, T1ce, T2, and FLAIR) are used in the study. Target segmentation regions are whole tumor (WT), tumor core (TC), and enhancing tumor (ET). The proposed framework is a self-supervised learning (SSL) and modality-aware fusion (HeMIS) hybrid framework, which is built within an encoder-decoder structure. The study investigates the segmentation performance of the dice score and the behavior of the models when trained in the absence of certain modalities.

1.6.1 Practical Scope and Research Limitations

The clinical relevance of this study is important due to the clinical limitation that is being addressed. In real hospitals, particularly in low- and middle-income countries, MRI data may be incomplete, for a variety of reasons including cost, time, machine limitations, patient condition and protocol variants. A segmentation model with some useful results even in the case of incomplete modalities can help radiologists and decrease the reliance on ideal imaging protocols. Thus, the framework is a research prototype evaluated on BraTS 2021. Any clinical use would require external validation, safety testing, calibration, and expert review.

Moreover, this work can be further extended in the future to other multimodal or multi-sequence 3D medical image segmentation applications. The key ideas of self-supervised representation learning and modality-aware fusion can be transferred to other organs, diseases or imaging scenarios, as long as adequate datasets and expert validation are available.

1.6.2 Key Challenges and Technical Bottlenecks

Although there are clinical benefits, there are a number of practical and technical issues:

- **Computational Demands:** The 3D MRI segmentation models are trained with high computational costs and require a large amount of GPU memory as the inputs are volumetric.
- **Severe Class Imbalance:** Tumor subregions are very imbalanced, with certain regions of the image volume (e.g., enhancing tumor) accounting for a small proportion of the data volume.
- **Asymmetric Modality Variance:** The model may be different when the missing modality is different, due to the fact that the information provided by the various modalities is vastly different and asymmetric.External validation on real-world hospital data or other datasets is needed to make definite clinical claims, as the current work is performed using the BraTS 2021 dataset only.

- **Clinical Accountability:** The model can be used to automate segmentation, but a real clinical application would need to be further tested and undergo strict safety validation by the radiologist.

Overall, this research problem presents a novel and intricate computing problem by addressing the shortcomings of the labeled data dependency and missing-modality sensitivity in 3D brain tumor segmentation. The proposed HeMIS-SSL Fusion Hybrid Framework aims to enhance the robustness, data-efficiency, and real-world applicability of brain tumor segmentation in medical imaging.

Chapter 2

Literature Review

2.1 Preliminaries

First, it is important to understand some fundamental concepts of deep learning, medical imaging and representation learning to comprehend automated brain tumor segmentation using self-supervised learning (SSL). In this section some of the terminology and parts that are fundamental to the methods examined and proposed in this piece of work are presented.

2.1.1 Clinical Background and Task Definition

Brain tumor segmentation is a key medical image analysis problem that requires the automatic detection, classification and segmentation of tumorous tissues from multi-modal Magnetic Resonance Imaging (MRI) images. In clinical practice, this process is crucial for diagnosis, surgical planning, radiotherapy targeting and treatment monitoring.

Although manual segmentation by expert radiologists remains the clinical gold standard, it is very time consuming (typically 1-2 hours per case), and is prone to interrater variability. Therefore, the general aim of automated segmentation research is to find methods which are not only accurate, but efficient, objective and repeatable.

The task is typically structured around segmenting three hierarchically related subregions, as standardized by benchmarks like the Brain Tumor Segmentation (BraTS) challenge:

- **Enhancing Tumor (ET):** The central, actively growing part of the tumor that exhibits uptake of gadolinium-based contrast agent in T1ce scans.
- **Tumor Core (TC):** Comprises the enhancing tumor plus surrounding necrotic (dead) and non-enhancing tumor tissues. It is formally defined as:

$$\text{TC} = \text{ET} + \text{NCR/NET (necrosis/non-enhancing tumor)}$$

- **Whole Tumor (WT):** Encompasses all abnormal tissues visible on MRI, including the tumor core and the surrounding peritumoral edema (swelling). It is formally defined as:

$$\text{WT} = \text{TC} + \text{Edema}$$

2.1.2 Magnetic Resonance Imaging (MRI) Modalities

The effectiveness of brain tumor segmentation relies on the complementary information provided by different MRI sequences, each highlighting specific pathological and anatomical features:

- **T1-weighted (T1):** Provides detailed anatomical structure of the brain, offering a clear baseline view of gray matter, white matter, and cerebrospinal fluid.
- **T1-weighted with Contrast Enhancement (T1ce):** After injecting a gadolinium-based contrast agent, this sequence highlights areas with a compromised blood-brain barrier, making it essential for identifying the enhancing tumor (ET).
- **T2-weighted (T2):** Excellent for visualizing fluid-filled areas, making it highly sensitive to edema and cystic components of the tumor.
- **Fluid-Attenuated Inversion Recovery (FLAIR):** Suppresses the cerebrospinal fluid (CSF) signal, allowing for clearer visualization of hyperintense pathological tissues, particularly the edema surrounding the tumor core. This is crucial for defining the whole tumor (WT) boundary.

A significant challenge in clinical practice is the frequent occurrence of missing modalities, where one or more of these sequences are absent due to acquisition errors, protocol variations, or patient contraindications (e.g., renal failure preventing contrast use).

2.1.3 Deep Learning Architectures for Segmentation

Modern automatic segmentation systems are based on Fully Convolutional Networks (FCNs). The most influential is the U-Net, which has a symmetric encoder-decoder architecture with skip connections that retain spatial context for accurate localization. There are many variants that have been created to improve its capabilities:

- **3D Convolutional Networks:** These models process volumetric data directly, such as 3D U-Net and 3D-UX-Net, which help capture the important 3D spatial context.
- **Hybrid Architectures:** The combination of local feature extraction power of CNNs and the global contextual understanding of Transformers is very effective, known as hybrid architectures. This is a powerful synergy that is exemplified by architectures such as ResViT and LVM-Med.
- **Attention Mechanisms:** To enhance feature maps, modules such as channel attention, spatial attention, and wavelet-based attention are incorporated into the model to enable it to focus on the most relevant features for identifying the tumor boundaries.

2.1.4 Self-Supervised Learning (SSL) Paradigms

The availability of expert-labeled data being costly and limited, the paradigm of Self-Supervised Learning has grown in popularity. The idea behind SSL is to learn powerful representations from a large amount of unlabeled data, before fine-tuning on the small amount of labeled data for the downstream task (segmentation). Some of the key SSL strategies are:

- **Contrastive Learning:** Methods such as SimCLR, MoCo, and SSCLNet learn representations by maximizing similarity between two views of the same image and minimizing similarity between views of the same image and views of other images.
- **Generative & Predictive Tasks:** This includes inpainting masked image modeling (MIM) as in MAE and E-MIM, predicting image rotations, solving jigsaw puzzles, and using diffusion models (BrainMRDiff) to generate missing modalities or learn underlying data distributions.

The main advantage of SSL is that it can alleviate the annotation workload to a great extent while enhancing the generalization and robustness of the model.

2.1.5 Evaluation Metrics

The performance of the proposed segmentation model is evaluated using the Dice Similarity Coefficient (DSC), which is a widely used evaluation metric in medical image segmentation. The DSC measures the volumetric overlap between the predicted segmentation mask and the ground truth mask, ranging from 0 (indicating no overlap) to 1 (indicating perfect overlap). In this work, the Dice score is used to evaluate the segmentation performance for three specific regions:

- **Whole Tumor (WT)**
- **Tumor Core (TC)**
- **Enhancing Tumor (ET)**

2.2 Review of Existing Research

There have been several studies over the past years on how to enhance the accurateness, efficiency, and generalizability of brain tumor segmentation by adopting self-supervised learning (SSL), deep learning structures and huge-scale medical imaging datasets.

Self-supervised learning (SSL) has become one of the most promising approaches in medical image analysis, particularly in brain MRI, because it enables models to learn useful feature representations from large unlabeled datasets. Existing studies have developed two main families of SSL methods: contrastive learning, which pulls representations of different views of the same sample closer while pushing apart unrelated samples and reconstruction or predictive methods such as masked image modeling, which learn to reconstruct missing or corrupted image patches (Mondal

et al., 2022; Lin et al., 2024) [2][12]. Research has consistently shown that SSL improves downstream tasks such as classification and segmentation, especially when labeled data are limited (Wang et al., 2025; Zhao et al., 2024) [19][15]. To address this challenge, there is a growing trend to develop adaptations of SSL methods, such as SimCLR, MoCo, and masked autoencoders, which have proven to be highly transferable to 3D medical imaging, further highlighting the importance of pretraining on unlabeled MRI volumes (Singh et al., 2024; Li et al., 2025) [14][4]. The findings are the reason for us to choose SSL as the basis of our proposed framework.

The new architecture has also changed the design of segmentation networks, besides SSL. Convolutional neural networks (CNNs) are good at learning local textures, edges and intensity variations which are crucial for tumor boundary identification. The encoder-decoder architectures such as 3D U-Net like architectures are popular due to their ability to provide volumetric context and reconstruct detailed segmentation masks. Models like ResViT and attention-enhanced U-Net variants combine the advantages of both paradigms, with a more precise tumor segmentation by considering both fine-grained boundaries and long-range dependencies (Karagoz & Fox, 2024; Havaei et al., 2019) [11][1]. Other works also go further to add boundary delineation with edge-aware modules or specialized loss functions to improve the quality of segmentation of complex tumor regions (Karagoz & Fox, 2024; Zhang et al., 2023) [11][8]. However, the present work does not adopt a CNN-Transformer pipeline. Instead, it focuses on a custom encoder-decoder framework combined with self-supervised MRI patch reconstruction and HeMIS-inspired modality-aware fusion.

A consistent issue in brain MRI research is missing modalities. In practice, some MRI sequences such as T1ce or FLAIR may not be available, which lowers segmentation performance. Existing methods address this problem mainly in two ways: by designing models that can work with variable modality inputs or by generating missing sequences with synthesis-based approaches. Although modality synthesis can be useful in some settings, the present research focuses on robustness through HeMIS-inspired modality-aware fusion rather than generating missing MRI modalities. This direction is more aligned with the current framework, where available modalities are represented and fused through modality-specific features and statistical abstraction. Recent progress, including diffusion-based models and anatomy-aware GANs, has shown that reconstructing absent modalities can recover much of the lost accuracy (Bhattacharya et al., 2025; Li et al., 2025; Xu et al., 2025) [16][4][20].

Most previous works are benchmarked on the BraTS dataset, which is the de facto standard dataset for brain tumor segmentation tasks (Alam et al., 2024) [9]. BraTS offers uniform evaluation procedures for whole tumor, tumor core and enhancing tumor segmentation. It also provides an environment for research that is curated, and in which full modalities are generally available, which may not be representative of actual clinical settings. Thus, the current research employs the BraTS 2021 dataset as an experimental dataset as well as evaluates missing-modality conditions, following training, to assess practical robustness. The evaluation is kept focused on Dice-based segmentation performance. In recent years, there has been a focus on evaluating generalization on alternative datasets like METS, ISLES, or ATLAS to

assess robustness in different clinical contexts (Nguyen et al., 2023; Alam et al., 2024) [6][9].

Another important area of the literature that has been lacking is clinical reliability and integration. While many SSL and segmentation models achieve high accuracy, they may not be easily deployed in clinical settings without expert validation. Real world reliability can be reduced due to incomplete MRI data, limited computational resources, and variations in hospital imaging protocols. These issues argue for the need of models that are accurate and robust under incomplete data conditions. In this study, explainability and clinical interface design are not contributions, but rather future directions. Explainable AI tools, such as attention maps and uncertainty visualizations, are still rarely applied in brain MRI segmentation (Rahman et al., 2025; Alam et al., 2024) [17][9]. Moreover, few studies have focused on human-in-the-loop validation, in which radiologists can validate or modify the model's output. Recent studies show that using large language models (LLMs) to help with report summarization or label verification could help close this gap (Rahman et al., 2025) [17].

Chen et al. (2024) [10] conducted one of the earliest large-scale explorations of self-supervised foundation models for multicontrast MRI in brain tumor diagnosis. Their study introduced a cross-contrast masked autoencoder (MAE) for representation learning from over 57,000 unlabeled MRI volumes, followed by fine-tuning for tumor detection, classification, and molecular prediction. The pretrained foundation model achieved up to 94.9% accuracy and AUC 0.981 in tumor detection, and produced high-quality interpretability saliency maps targeting tumor regions, and well-performing convolutional baselines. The work demonstrated the scalability of self-supervised pretraining for 3D MRI of the whole brain and the ability to decrease the need for annotated data. The method, however, was very compute intensive and did not evaluate in the missing-modality conditions, which limits real-world clinical adoption.

Robinet et al. (2025) [18] introduced a multimodal masked autoencoder (BM-MAE) framework that allows for the missing modalities of MRI to be either missing during pretraining or during fine-tuning. They trained their model on the BraTS 2021 dataset using a shared Vision Transformer (ViT) encoder that was trained with cross-modal dependencies using random modality masking, enabling seamless adaptation to different MRI sequences (T1, T1c, T2, FLAIR). The pretrained encoder was fine-tuned for segmentation and subtyping tasks, and achieved better performance than traditional MAE and contrastive learning baselines like SimCLR. BM-MAE showed an improvement of 3–4% Dice score on all tumor subregions, compared to models trained from scratch. This way, the problem of incomplete clinical data was solved successfully, and it was very flexible and generalized. However, it was a complex tuning and a longer pretraining process that made it less accessible to small research institutes.

Wang et al. (2023) [7] proposed Anatomical Prior-Informed Masked Autoencoder (API-MAE) to leverage brain structure priors during self-supervised pretraining to boost downstream brain segmentation. API-MAE employed probabilistic anatom-

ical maps based on tumor occurrence distributions to guide patch selection during pre-training, so that the model is focused on the brain regions that are semantically relevant. Trained with 6,415 T1-weighed MRIs and fine-tuned on BraTS21, it outperforms state-of-the-art SSL frameworks like Swin UNETR and MAE, delivering high Dice and accuracy scores with data efficiency. The inclusion of a reconstruction discriminator enhanced feature learn-ing and visual fidelity. The model's success was offset, however, by the need for hand-crafted anatomical priors to make this model flexible to other imaging modalities or pathologies, and the model's computational complexity, which limited its clinical scalability.

Shurrab et al. (2022) [3] proposed an Enhanced Self-Supervised Learning (SSL) algorithm for 3D MRI Segmentation using a dual-decoder U-Net with SimSiam pre-training and edge detection. They proposed a framework that jointly predicted the segmentation and edge masks for enhancing the accuracy of segmentation boundary and minimizing reliance on annotated data. The model was trained with BraTS 2020 and fine-tuned with BraTS 2018 and it obtained Dice scores of 0.884 (ET), 0.837 (TC) and 0.846 (WT) which outperforms several contemporary models. This was an integration of edge information that helped with spatial awareness and helped to better delineate tumor borders. However, the method was still a very expensive process from the computational perspective and was not externally tested or validated for AI integration, making it unusable for real-world clinical applications.

The BraTS-METS Challenge (2023) [5] report discussed the segmentation of brain metastases in multi-institutional MRI data, noting the issues of small lesions, false positives, and data heterogeneity. The competition offers standardized annotations from the SRI-24 brain atlas so that the segmentation algorithms can be fairly compared. Several machine learning and deep learning models performed well, but many models were not robust in their ability to detect small lesions and were not generalizable across centers. The research highlighted the long-standing issue of consistency in annotation, standardization of modalities, and explainable evaluation. These results are directly relevant to the current work, as they highlight the disconnect between research performance and clinical reliability, and give rise to a desire for the creation of strong, generalizable and interpretable segmentation frameworks based on SSL which can cope with real-world MRI variability.

Furthermore, computational efficiency is now increasingly acknowledged in the literature. Most SSL methods and hybrid 3D architectures are resource consuming, with large memory and long training time. To tackle this, researchers have started to create lightweight variants of U-Nets, efficient hybrid encoders and knowledge distillation methods for achieving accuracy while guaranteeing deployability (Shah & Rathi, 2024; Karagoz & Fox, 2024) [13][11]. In this case, these works demonstrate the importance of resource-aware design in real-world applications, which we also use in our objectives, taking into account inference time, memory usage, and model size.

Overall, the reviewed literature demonstrates the benefits of self-supervised learning, masked image reconstruction, encoder-decoder segmentation, and missing-modality-aware modeling for brain tumor segmentation. However, many existing approaches either focus only on SSL representation learning, depend on complete MRI modal-

ities, require heavy architectures, or address missing modalities through synthesis-based methods. The present research addresses this gap by combining SSL pretraining through original MRI patch reconstruction with HeMIS-inspired modality-aware fusion inside a custom encoder-decoder framework. The work is evaluated on BraTS 2021 for Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET) segmentation using Dice score, with additional missing-modality testing after training.

2.3 Summary of Key Findings

The reviewed research activities reveal several important patterns, opportunities, and remaining gaps in the field of brain tumor segmentation using deep learning and self-supervised learning. Self-supervised learning is a promising solution in medical imaging because it enables models to learn from unlabeled data. Masked image modeling and other reconstruction-based methods are especially relevant when labeled data are limited. This supports the use of SSL pretraining in the current work, where the model learns by reconstructing original MRI patches before segmentation fine-tuning.

Encoder-decoder architectures are still vital for volumetric medical image segmentation due to their ability to effectively leverage 3D spatial context and maintain localization information by means of decoder-based reconstruction and skip connections. This supports the current choice of a custom encoder-decoder framework instead of a more robust CNN-Transformer design.

In the literature, it has also been reported that the lack of the MRI modalities still poses a practical challenge for clinical deployment. There are a large number of segmentation models that work well when all modalities are present but in real clinical data, some modalities may be missing. This will facilitate the use of modality-aware fusion based on HeMIS in the proposed framework and motivate the need for missing-modality evaluation following training.

The BraTS is a popular benchmark dataset for brain tumor segmentation, however, it remains a curated research dataset. Consequently, results from BraTS 2021 should be considered as research validation and not as approval. In the current work, the Dice-based evaluation metric is applied to the evaluation of WT, TC, and ET segmentation and external clinical deployment is not claimed. Last but not least, many of the advanced SSL and 3D segmentation models are very expensive in terms of computation.

This creates a need for approaches that balance robustness, segmentation performance, and practical feasibility. The proposed HeMIS-SSL Fusion Hybrid Framework follows this direction by combining SSL MRI patch reconstruction, modality-aware fusion, and encoder-decoder segmentation within a focused BraTS 2021 experimental setting.

2.4 Summary

Key Findings	Description	Impact
Self-Supervised Learning (SSL)	Uses unlabeled MRI data through reconstruction-based learning, including masked image modeling.	Reduces dependence on manual annotation and improves feature representation.
Encoder-Decoder Segmentation	Processes 3D MRI volumes through encoding, fusion, and decoding stages.	Supports voxel-level brain tumor segmentation with spatial context.
Missing-Modality Handling	Uses modality-aware learning ideas such as HeMIS-style fusion to combine available MRI modalities.	Improves practical relevance when complete MRI protocols are unavailable.
BraTS 2021 Benchmarking	Provides four modalities and standard WT, TC, and ET segmentation targets.	Enables focused Dice-based evaluation under a standard research setting.
HeMIS + SSL Fusion Framework	Combines self-supervised masked image modeling pretraining with HeMIS-inspired modality-aware fusion inside a custom encoder-decoder architecture. The SSL encoder learns volumetric MRI representations from unlabeled data before supervised training, while HeMIS computes mean and variance statistics across available modalities at each encoder scale to handle any combination of present or missing inputs.	Achieves a mean Dice score of 0.766 with all four modalities and maintains usable segmentation performance across all 15 non-empty modality combinations, outperforming both SimCLR and Med3D baselines. Demonstrates that a single trained model can handle varying modality availability without architectural modification, improving both practical robustness and data efficiency in simulated incomplete MRI settings.

Table 2.1: Summary of Key Findings

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

The final specification of this project is to develop a hybrid deep learning model for 3D brain tumor MRI segmentation using self-supervised learning and HeMIS-based fusion. The main objective is to design a model that can segment tumor regions from MRI scans while addressing two practical limitations of current medical image segmentation systems: the dependency on fully labelled datasets and the inability to work effectively when MRI modalities are missing. Therefore, the proposed solution is designed to learn useful image features from MRI volumes without using segmentation labels during the SSL pretraining stage and perform segmentation even when only one or some of the MRI modalities are available.

The project uses the BraTS 2021 dataset as the main experimental dataset. The model uses four BraTS MRI modalities: FLAIR, T1ce, T1, and T2. The required output is a 4-class segmentation mask consisting of background, necrotic/non-enhancing tumor core, edema, and enhancing tumor regions. Instead of performing only binary tumor and non-tumor classification, the model is expected to support multi-class tumor segmentation. This makes the output more useful for detailed medical image analysis and comparison with existing brain tumor segmentation methods.

3.1.1 Functional Requirements

The functional requirements define what the proposed system must be able to perform:

- **Input Processing:** The system must accept 3D MRI volumes as input.
- **Multi-Modal Segmentation:** It must process multiple MRI modalities and generate a tumor segmentation mask.
- **Robustness to Missing Modalities:** The model must be able to operate under missing-modality conditions, including cases where only one modality is available or where different combinations of modalities are present. The model

is evaluated under 15 possible modality combinations using the evaluation script to assess robustness under incomplete input conditions.

- **Self-Supervised Pretraining:** The system must support self-supervised pretraining so that unlabelled MRI data can be used for feature learning before supervised segmentation training.
- **Evaluation and Benchmarking:** Finally, the system must produce evaluation results that can be compared with existing models such as U-Net, Med3D, SimCLR, and Swin U-Net.

3.1.2 Non-Functional Requirements

The non-functional requirements are those which outline the desired quality and operating conditions of the system. The model performance is assessed by Dice-based metrics such as NCR, ED, ET, mean Dice and BraTs-style clinical region scores (ET, TC, and WT). Furthermore, HeMIS variance is considered when assessing to see how variation changes with different combinations of modalities. It must be train-invariant, reproducible, with documented settings, and working in available hardware and software frameworks. The implementation will be implemented in the following languages and frameworks: Python, PyTorch, CUDA and a desktop GPU system. Practical deep learning methods like mixed precision, checkpoint saving, and hyperparameter tuning should also be supported in the training process to deal with long training time and the computational cost.

3.1.3 Constraints

Several constraints are also considered in the final specification:

- The model is limited to research-based evaluation using the BraTS 2021 dataset and does not include direct clinical deployment or hospital-based validation.
- The model performance may depend on dataset quality, pre-processing steps, available GPU memory, training time, and the selected modality combinations. All missing-modality cases in this thesis are simulated from complete BraTS 2021 subjects.
- Since 3D MRI segmentation is computationally expensive, training time and GPU resources are important practical constraints.

3.1.4 Professional Responsibilities and Codes of Practice

Applicable professional responsibilities and codes of practice are carefully considered. Even if the medical imaging data set is publicly available and anonymised, the research should adhere to responsible data-use principles.

The model should not be used as a substitute for a physician or a radiologist. Expert validation and further testing and approval would be needed for any clinical use. Moreover, there are duties to uphold in the academic aspect, such as presenting results in an honest manner, comparison with a baseline model, and citing previous

studies, as well as documenting the model design, training process, limitations, and model evaluation procedure.

3.2 Societal Impact

The proposed hybrid medical imaging model has a number of important societal impacts, particularly due to its ability to deal with both unlabelled data and missing imaging modalities. In practice, in healthcare systems such as in developing countries and low resource hospitals, medical data is frequently incomplete, inconsistent or improperly labelled. A medical AI model that can learn from such incomplete or unlabelled research data can become more practical, accessible and useful in real-world settings away from tightly controlled research settings.

This study may therefore not just result in technical advancement in brain tumor segmentation but also to the general advancement of accessibility to healthcare, efficiency in the clinical setting and patient care.

3.2.1 Healthcare Equity and Global Accessibility

A major impact of this research in society is the avoidance of expensive labelled medical data. In medical imaging, labelled data typically refers to the process of manually identifying and marking tumor regions in MRI scans that requires the expertise of a radiologist or clinician. This can be very time consuming, costly, and hard to scale. Creating large labeled datasets is a significant obstacle when building a reliable AI system, as medical professionals already have a lot to handle.

Thus, the utility of the proposed model to learn from unlabelled data is of great importance. There are a lot of large volumes of MRI data stored in hospitals that are not manually annotated. Unfortunately, in a lot of situations such information cannot be leveraged for AI development due to the lack of expert labels. A model that is able to learn useful knowledge from unlabelled data can transform the existing hospital records into useful training resources. This can decrease the data preparation costs and effort, making AI development more viable for institutions that have limited resources.

This capacity is particularly critical in countries with less expertise and medical research facilities. In this context, the collection of a large number of high quality labelled MRI images might not be feasible. The current reliance of AI models on fully labeled datasets would limit meaningful involvement in the medical AI development process to only large hospitals, advanced research centers, or wealthier nations.

This study could pave the way for the adoption of medical AI that is more accessible. It makes the use of available unlabelled imaging data possible in smaller hospitals and in developing regions. This could help create more localised AI tools and models that are trained on data from patients in various populations, not just high-quality international data made up mainly of funded patients. This is socially important because AI systems need to be effective in a variety of populations, clinical settings, and healthcare systems.

3.2.2 Clinical Health Outcomes and Diagnostic Efficiency

A significant societal impact is the potential to help diagnose brain tumors more quickly. Timely diagnosis of brain tumors is crucial since treatment decisions rely on the accurate identification of tumor regions, such as enhancing tumor, tumor core, and whole tumor area regions. Brain tumor MRI segmentation is a time-consuming process, especially when the tumor is complicated in shape and a large number of patients are to be examined.

Use of an automatic or semi-automatic version may help to generate tumor segmentation results more rapidly. The model can be used as an aid to clinical decision making, giving a preliminary segmentation map, but the final decision must be made by qualified doctors. If proven to be true on a clinical level, this might enable the clinician to review cases in a quicker time, spend more time on interpretation, treatment planning and patient communication.

A faster segmentation of tumors can be especially beneficial in busy hospitals, where radiologists need to review numerous scans daily. These settings can lead to delays in diagnosis and treatment planning due to the time it takes to analyze images. An AI model that can provide a preliminary analysis may help to shorten the waiting period and enhance the efficiency of the diagnostic workflow. This can be helpful for patients as they will be able to receive treatment decisions sooner based on the characteristics of the tumor.

In a case of serious illness, like brain tumors, a mere 10 minutes or so of diagnostic speed makes a big difference for people and their situations. Delay in diagnosis and diagnosis-related procedures can be a source of stress for patients and families, as can be the anxiety that ensues during a diagnostic process.

The study could also help with treatment planning. Tumor segmentation plays an essential role in surgery, radiotherapy, chemotherapy monitoring and follow-up evaluation. If the model can accurately and consistently provide tumor-boundaries, it can help doctors understand the location and size of tumors as well as the progression of them. This can assist with the planning of treatments and tracking changes over time. In follow-up, automated segmentation can assist with the comparison of tumor volume across various scans, for example. This can help to make more objective judgments about a tumor's growth, shrinkage or response to treatment.

3.2.3 Research Robustness under Simulated Missing Modalities

Besides the missing modalities, the model can be used for many of the societal and safety concerns. Typically, one would expect all the MRI modalities to be available in ideal research datasets, including T1, T1ce, T2, FLAIR, etc. But not all real hospitals have a full set of imaging information. Limited time, patient movement, technical problems, costs, and variations in hospital scanning protocols may be responsible for the absence of some MRI sequences.

For many low resource countries, full MRI scans will not be available consistently. Traditional models can be inadequate or even be ill-adapted if one or more of the modalities is missing. It is therefore more appropriate to use a model suitable for dealing with incomplete modalities in real-world clinical application. It can help prevent patients from being denied the benefits of AI analysis for incomplete imaging data.

Parameter	**Ideal Research Setting**	**Low-Resource/ Real-World Setting**
Available Modalities	Full Protocol (FLAIR, T1, T1ce, T2)	Incomplete or missing sequences
Data Annotation	Fully and expert-annotated	Predominantly unlabelled historical records
System Response	Low-throughput demand	High-throughput, resource-constrained
Equitability	Restricted to advanced facilities	Adaptable to rural and smaller clinics

Table 3.1: Comparison of AI Operational Constraints across Healthcare Environments

This feature can also work toward promoting health equity. In advanced hospitals, patients may be able to get full high-quality MRI scans, whereas patients in smaller or in rural hospitals may not get as many imaging sequences. While AI is beneficial, it may exacerbate the disparity between high-resourced and low-resourced healthcare systems if the technology is only effective in best-case scenarios. This can be bridged by making AI assistance available to users when the imaging protocol is not complete, through a missing-modality-tolerant model. Consequently, the research can help in the more practical and equitable use of medical AI.

3.2.4 Professional and Legal Aspects of Human-AI Collaboration

Another major social benefit is the burden on radiologists is reduced. A radiologist is a doctor who specializes in reading and understanding intricate medical images, preparing reports, and assisting in treatment choices. In most healthcare systems, there is less number of trained radiologists as the demand for imaging services is growing. Brain tumor segmentation is a time-consuming and meticulous task and hence highly repetitive. The proposed model can ease the burden of manual work by aiding in segmentation. It is not the replacement of a radiologist, but the assistance of a radiologist using an automated tool to do routine work, freeing up his or her time for more advanced clinical decisions. Having a load lightened can also help in getting consistency. Manual segmentation can differ among experts, due to the fact that different radiologists define tumor boundaries somewhat differently. Other factors that can impact consistency include fatigue, workload pressure and time constraints. AI-powered segmentation can give a consistent initial output,
These can then be reviewed and corrected by the radiologist(s). This partnership between human and AI can enhance efficiency, with simultaneous medical-legal ac-

countability and clinical supervision. In the future, this may enable hospitals to serve more patients without adding to the workload of healthcare workers.

3.2.5 Cultural Shifts and Data Sustainability

Another added value is the more efficient and sustainable utilization of the medical data already existing, representing a positive evolution in the hospital data culture. While the amount of imaging data generated in hospitals is significant each year, a substantial portion of that data is not utilized for research or developing AI applications due to being unlabelled or incomplete.

These imperfect datasets can be made more useful by using the proposed model. This will promote a more sustainable use of medical AI, as it leverages existing data rather than collecting new data, which is costly. It can also spur hospitals to take better care of their imaging data collection and management, even if the information is not labeled, since it too can play a role in advancing AI in the future.

The overall societal impact of this research is bringing brain tumor segmentation technology to a more practical, inclusive and clinically supportive environment. It can learn from unlabeled data, which helps to save the cost of expensive expert annotations. It can enhance patient care and workflow in the clinical setting, thanks to its ability to facilitate faster diagnosis. It can lower radiologists' workload, which helps healthcare systems to cope efficiently with rising imaging demand. Concurrently, it can process the data from only one modality, which makes it more practical for use in hospitals that do not always have full MRI data sets. Thus, this study can help to create more user-friendly, scalable, and socially impactful medical AI tools, especially in resource-constrained healthcare settings.

3.3 Environmental Impact

The proposed hybrid medical imaging model can create several meaningful environmental impacts by making medical image analysis more data-efficient and reducing unnecessary use of clinical resources. Although the research is mainly focused on brain tumor segmentation and medical AI performance, its ability to handle unlabelled data and missing MRI modalities has indirect environmental benefits.

In modern healthcare systems, medical imaging services require large amounts of electricity, storage capacity, expert time, and repeated clinical procedures. Therefore, any model that can work effectively with imperfect real-world data can contribute to a more sustainable healthcare workflow. These individual benefits together can support a broader reduction in the environmental footprint of medical imaging services.

The study could also slightly cut down patient travel for rescanning. Depending on the amount of data that is missing or incomplete from the MRI, some patients may be able to not go to the hospital again for another MRI. In a few instances, this can decrease transport-related emissions. But this effect will depend on the hospital practice and patient circumstances and should therefore not be stated as

an absolute benefit, but as a possible indirect benefit.

3.3.1 Energy Optimization via Reduced Scan Repetition

One of the most relevant environmental impacts of this research is the possible reduction in repeated MRI scans. However, in real clinical applications, not all MRI modalities are available. The sequence may not be present due to technical problems, patient motion, time constraints, cost considerations, or variations in the way hospitals acquired the images.

Most of the traditional AI models need all the types of MRI modalities to generate any meaningful outcome. When one modality is not available, the scan might not be as helpful for the AI analysis and sometimes more scans of the patient are required. This dependence on full imaging procedures can be minimized with a model that is able to process missing modalities. Consequently, even if some sequences were not acquired, pre-existing MRI scans may be used for tumor segmentation as well.

This is significant for the environment because MRI machines use a lot of electricity when they are switched on. The more scans a machine performs, the more power is required, cooling systems need to be in place, tech support must be on hand, space must be prepared, and data must be processed. The model can indirectly save electricity when used in hospitals if it can decrease the number of repeat scans or additional scans that need to be done. This does not imply that it is not necessary to undertake scans that are clinically required. But it does not mean that if the data is not complete, then redundant data may be avoided. Thus, the model ensures the efficient use of imaging facilities and helps to reduce the energy consumption in medical imaging departments.

3.3.2 Resource Sustainability in Low-Resource Infrastructure

Being able to work with missing modalities can also enable hospitals to make the best use of available imaging resources responsibly. Full MR protocols are not always feasible in many low resource hospitals, due to machine availability, patient volume, cost or lack of trained personnel.

Hospitals could be tempted to order unnecessary scans to meet the demands of the AI model, as the systems need ideal and complete data.Hospitals may be tempted to order unnecessary scans just to ensure the technical needs of the model are met since AI systems require ideal and complete data. A missing-modality-tolerant model lessens this pressure. It would enable hospitals to leverage clinical data they already have on an MRI, not create specific data just for AI purposes. This makes the model more environmentally viable for deployment in the real world, particularly in healthcare settings that have limited resources.

3.3.3 Computational and Human-Resource Efficiency

Another environmental advantage is related to the decrease in the computational and human-resource waste of medical data labelling. Medical image labelling requires not only a lot of time but also a significant amount of resources. Labelled datasets require expert radiologists or trained annotators to work on dedicated workstations, software, data storage, and the repeated review process.

The proposed model can learn from unlabelled data, which can help to reduce the requirements for fully labelled data in all phases of the model development. This can reduce the need for experts' time and digital infrastructure for annotation. A significant environmental benefit may still exist for the preparation of large, perfectly annotated datasets, even though computation will be needed for model training, if better use of unlabelled data can be made.

3.3.4 Data Lifecycle and Digital Storage Optimization

Medical imaging data are big, particularly in the case of 3D MRI, which comes with several modalities. The increased demand for digital storage is caused by repeated scans, duplicate imaging and unnecessary complete imaging protocols. You need to have the servers, backup infrastructure, electricity and cooling for data storage systems. The model may help to decrease unnecessary duplicate scans and make the incomplete data more useful, which will help to mitigate unnecessary imaging storage growth. This will help in the promotion of better data management in hospitals and research institutions in a more sustainable way.

3.3.5 Summary of Environmental Sustainability Factors

To provide a structured assessment of the project's sustainability footprint, the core environmental vectors are summarized in Table 3.2.

Sustainability Factor	**Description**	**Environmental Impact**
Modality Tolerance	Eliminates technical mandates for missing sequence scanning	Lowers hospital electricity consumption and cooling load
Unlabelled Data Utilization	Leverages existing raw records via Self-Supervised Learning	Minimizes the infrastructure carbon footprint of manual curation
Storage Optimization	Reduces duplicate files and redundant high-resolution volumes	Checks growing demand for data center energy and backup servers
Patient Workflow Efficiency	Mitigates need for clinical recalls due to faulty/incomplete initial protocols	Indirectly curtails transport-related emissions from diagnostic travel

Table 3.2: Summary of Project Sustainability and Environmental Impact

Overall, the environmental impact of this research comes mainly from improving efficiency in medical imaging workflows. By handling missing modalities, the model

may reduce unnecessary repeated MRI scans, lower energy consumption, and reduce pressure on low-resource hospitals to perform extra imaging. By learning from unlabelled data, it can make better use of existing medical datasets and reduce some of the resource burden associated with manual labelling and repeated data preparation.

Therefore, the broader environmental value of this research is its potential to contribute indirectly to reducing the overall carbon footprint of healthcare imaging by promoting better data use, fewer unnecessary procedures, lower energy demand, and more efficient clinical AI deployment.

3.4 Ethical Issues

Ethical considerations and professional responsibilities have been taken into consideration as part of this research as the proposed model is for medical image analysis in particular brain tumor segmentation. In spite of its technical nature, the possible use of the model has a relationship with patient care, clinical decision making, and medical research.

Hence, ethical responsibility should be upheld throughout the design, training, assessment and reporting phases of the work. The primary ethical issue is not so much the performance of the model, but the development and presentation of the model in a safe, honest and responsible manner.

3.4.1 Clinical Risk Mitigation and Human-in-the-Loop Supervision

The potential for misdiagnosis is one large ethical problem. The model could generate wrong segmentation results, particularly in challenging cases, if the MRI series was not of high quality, or in cases where there are missing MRI modalities. The segmentation process could lead to an over or under estimation of the tumor area if not supervised accordingly in the clinical interpretation.

So, the model should not be shown as a stand-alone diagnostic system. It should be described as an assistive tool to aid medical practitioners not replace medical practitioners. The distinction is ethically significant since it could place patients in serious danger when trusting an AI system too much in medical practice.

Therefore, human supervision should play an integral role in the use of the model for this reason. Doctors and radiologists can be qualified to finalize the diagnosis, treatment plan, surgery, radiotherapy or patient follow-up. The model can give an initial segmentation result, save manual labour, and aid in more rapid review; but the decision for clinical judgment is to be left to medical experts. This should be explicitly addressed when designing the solution as part of the solution design process, as the model should be designed as a *human-in-the-loop* system, but not as a full-fledged clinical decision-maker.

3.4.2 Data Governance and Ethics of Unlabelled Data

Another ethical responsibility is the proper use of unlabelled data. The strength of the proposed model is that it can learn from unlabelled MRI data, but unlabelled data are still medical data. Even without segmentation labels, MRI scans may contain sensitive patient information.

Therefore, the use of unlabelled data must follow ethical research standards, dataset rules, institutional approval, and data protection requirements. Researchers must ensure that the data are handled securely and used only for the intended academic purpose. The availability of unlabelled data should not be treated as permission for unrestricted use.

3.4.3 Academic Transparency, Reproducibility, and Fair Comparison

Other academic duties include the ability to produce reproducible work and to document it appropriately. The model architecture, dataset description, pre-processing procedure, training process, missing modality strategy, evaluation metrics and limitations should be well documented. This enables other researchers to comprehend how the results have been generated, and whether the comparison is valid. Good documentation will also help to reduce the chances of confusion, missed errors, or misinterpretation. This is a hybrid research and needs to be properly explained to each component to maintain academic transparency.

Another significant duty is to compare with other models that have existed. To compare the proposed model, it is desirable to use similar datasets, evaluation settings, metrics, and pre-processing conditions. If the model is tested in other situations, it should be stated. Comparing one model to another and taking a superior stance without a balance would be unethical.

Last but not least, responsibility for the mistakes needs to be taken into account. To ensure that the segmentation errors do not affect the clinical interpretation, it is important to establish a clear definition of the role of the model, researcher and medical expert. The research should recognize that the information generated by AI is not necessarily accurate and that it requires the review and confirmation of experts before it can be used in clinical practice.

The overall ethical responsibility of this research is to develop and publish the model in an honest, transparent, repeatable and cautious clinical fashion. In doing so, the research can be used to help develop medical AI without compromising academic integrity and professional responsibility.

3.5 Project Management Plan

The main goal of the project is to develop a deep learning-based brain tumor segmentation model that can work with both unlabelled MRI data and missing MRI

modalities. In simple terms, the project aims to make tumor segmentation possible even when:

- Labelled data are limited;
- One or more MRI modalities are missing;
- Real-world hospital data are incomplete.

The model should segment the tumor area from brain MRI scans. The output should identify tumor-related regions from the 3D MRI volume. The main purpose is to produce a segmentation map that highlights tumor areas. The pre-processing pipeline includes non-zero voxel normalization, percentile clipping, label remapping from BraTS label 4 to label 3, central cropping, and optional caching of processed volumes.

3.5.1 Project Resources

To achieve these development milestones, the computational framework is established on the following hardware and software resources:

- **Hardware Infrastructure:** NVIDIA GeForce RTX 4080 Super desktop GPU system.
- **Software Environment:** Python, PyTorch deep learning framework, CUDA toolkit, and VS Code IDE.

3.5.2 Core Architecture Architecture Features

The proposed hybrid model incorporates two critical technical innovations to ensure robustness under clinical constraints:

1. **Self-Supervised Learning (SSL) Capability:** The network can extract and learn useful image features from vast pools of unlabelled MRI data before entering supervised fine-tuning.
2. **HeMIS-Based Missing Modality Handling:** The architecture can perform reliable multi-class segmentation even when one or more essential MRI modalities are missing from the patient profile.

This makes the model more practical for research-level experiments that reflect real-world incomplete clinical data, where complete and labelled data are not always available.

3.6 Risk Management

The main goal of this section is to identify, analyze, and provide structured mitigation strategies for the primary challenges within this project. The overall execution risks are categorized into technical, performance, and training time domains:

- **Computational Barriers:** Low GPU utilization, prolonged training time, and dataset loading or pre-processing bottlenecks.

- **Optimization Challenges:** High loss, unstable training behaviors, gradient explosion, and slow model convergence needing excessive epochs.
- **Performance Concerns:** Lower Dice score, poor segmentation accuracy, or a final architecture that fails to outperform baseline benchmarks.
- **Operational Constraints:** Code implementation bugs and critical time management issues near the thesis submission deadline.

3.6.1 Technical Risks

The main risks are technical since it is an advanced project on deep learning techniques. Since the implementation is based on patch-based 3D training, which is limited by GPU memory, the evaluation results need to be considered in light of the evaluation setting used (patch/crop-based). For instance, the model might not be trained as well as expected, the loss function may become unstable, the GPU consumption may not rise to the expected level, or the Dice score may not increase as it should. Such technical constraints can have a significant impact on the quality of the final product and can also cause project milestones to be postponed.

3.6.2 Performance Risks

The model may show lower segmentation metrics than initially expected. Key factors behind potential performance degradation include:

- Structural complexities within the challenging BraTS 2021 dataset;
- Severe class imbalance across different tumor regions (e.g., enhancing tumor vs. large background regions);
- Insufficient training epochs due to timeline pressures;
- Selection of an unsuitable loss function or poor initial hyperparameter settings;
- Incomplete optimization or feature convergence issues across the hybrid model architecture.

3.6.3 Training Time Risks

Training might be longer than expected due to the use of large 3D MRI volumes and intensive self-supervised pretraining (SSL) but also due to the multi-class segmentation fine-tuning used in the model pipeline. Long training time is a major realistic obstacle throughout the research cycle, as 3D medical image networks are very computationally costly.

3.6.4 Risk Mitigation Strategy

The project has manageable risk, but the primary operational hurdles are training time, model performance, GPU efficiency, and managing datasets. These risks will be actively addressed by iterative hyperparameter tuning, automatic mixed precision (AMP) training, DataLoader optimization, loss, and compression tuning, as

well as the use of training/validation curve monitoring. Function optimisation, correct modular debugging.

Even though the model did not attain the best overall performance rating, the research contribution can be justified in its architectural capabilities for unlabelled data and missing MRI modalities, which are very significant to real-world medical imaging use cases.

3.6.5 Summary of Risk Assessment and Mitigation

A detailed overview of the core identified risks, their prospective system impacts, and technical mitigation workflows is provided in Table 3.3.

Risk Event	**System/ Project Impact**	**Mitigation Strategy**
Low GPU utilization	Extended training cycles and inefficient resource consumption	Optimize PyTorch DataLoader pipelines, enable pinned memory, and increase worker threads
High loss / gradient explosion	Training instability and model non-convergence	Implement strict gradient clipping thresholds and integrate adaptive learning rate schedulers
Low Dice score	Weak final segmentation mask generation and poor benchmarking	Fine-tune hyperparameters, transition to dynamic hybrid loss functions, and scale training epochs
Dataset loading error	Complete training interruption and pipeline execution bugs	Verify file paths, enforce precise modality parsing orders, and validate pre-processing scripts
Long training time	Project delays and high risk of missing the thesis submission deadline	Leverage automatic mixed precision (AMP), optimize 3D crop patch size, and automate checkpoint saves

Table 3.3: Risk Management Strategy and Technical Mitigations

3.7 Economic Analysis

The proposed hybrid brain tumor segmentation model has important economic relevance when compared with current medical image segmentation methods. Although this research does not include a direct financial cost calculation, a qualitative economic analysis can be made based on the practical advantages of the model.

Current segmentation methods often depend heavily on fully labelled MRI datasets and complete imaging modalities. In real clinical and research environments, preparing such ideal datasets can be expensive, time-consuming, and difficult. The proposed model addresses this issue by using self-supervised learning and HeMIS-based fusion, allowing it to learn from unlabelled data and perform segmentation even when some MRI modalities are missing.

3.7.1 Summary of Qualitative Economic Impacts

A breakdown of the traditional healthcare cost barriers and how the proposed hybrid architecture alleviates them is compiled in Table 3.4.

Economic Dimension	**Traditional Cost Driver/ Waste**	**Proposed Model Advantage**
Dataset Curation	Expensive expert radiologist labor required for manual frame-by-frame labeling	Self-Supervised Learning utilizes free, raw, unlabelled volumetric assets
Operational Equipment Use	Redundant clinical testing and machine hours due to missing/corrupted modalities	HeMIS fusion processes incomplete scans without triggering a re-scan fee
Clinical Labor Efficiency	Repetitive, manual pixel tracing by medical personnel consuming costly hours	Rapid preliminary segmentation output allows rapid expert review and faster throughput
Data Capital Value	Sunk costs in maintaining vast archives of historical unannotated data vaults	Regenerates utility from neglected hospital historical databases

Table 3.4: Qualitative Economic Assessment of the Proposed AI Architecture

Chapter 4

Proposed Methodology

4.1 Proposed Framework

The proposed method is a fusion of robust brain tumor segmentation through the use of Self-Supervised Learning (SSL) and HeMIS statistical fusion under missing modalities in MRI. The full dataset has 1,251 cases, but training used 1,000 cases and Fold 0 was held out with four modalities (FLAIR, T1ce, T1, and T2). Pre-processing consists of percentile clipping (0.05–99.95%) of all volumes, z-score normalization of non-zero voxels and central cropping to $128 \times 128 \times 128$ voxels, which were augmented randomly during training.

There are two phases of the framework. Phase 1 involves Masked Image Modeling on SSL pretraining used training folds only, consisting of four modalities with 75% random masking of 3D patches followed by 75% reconstruction from the Mean Squared Error (MSE) loss by a Transformer encoder. This teaches the model structural MRI patterns without the need for any annotations. Only the weights of the encoder are kept and the weights of the SSL decoder are discarded and used for the downstream feature learning.

The pretrained encoder is loaded, and the entire segmentation model is trained using labeled data in Phase 2. All the modalities go through an independent CNN encoding layer that generates multi-scale feature maps. Random modality dropout ($p = 0.3$) is used to simulate clinical scenarios of missing modalities during training. The mean extracts complementary information and the variance captures disagreement in the feature set across available modalities, a signal that is likely to be stronger as the number of modalities becomes smaller or their information becomes less — a signal that is not yet shown to be a calibrated clinical uncertainty measure, the HeMIS abstraction layer computes mean and variance across the available modality features. The output of SSL features are fused with the HeMIS features and these features are then passed through a segmentation decoder that gives 4 output classes (Background, NCR, ED, ET) optimized using combined Cross-Entropy and Dice loss. SSL features are fused with HeMIS features and passed through a segmentation decoder producing a 4-class output (Background, NCR, ED, ET) optimized via combined Cross-Entropy and Dice loss.

During evaluation, all 15 modality combinations are tested and HeMIS variance is

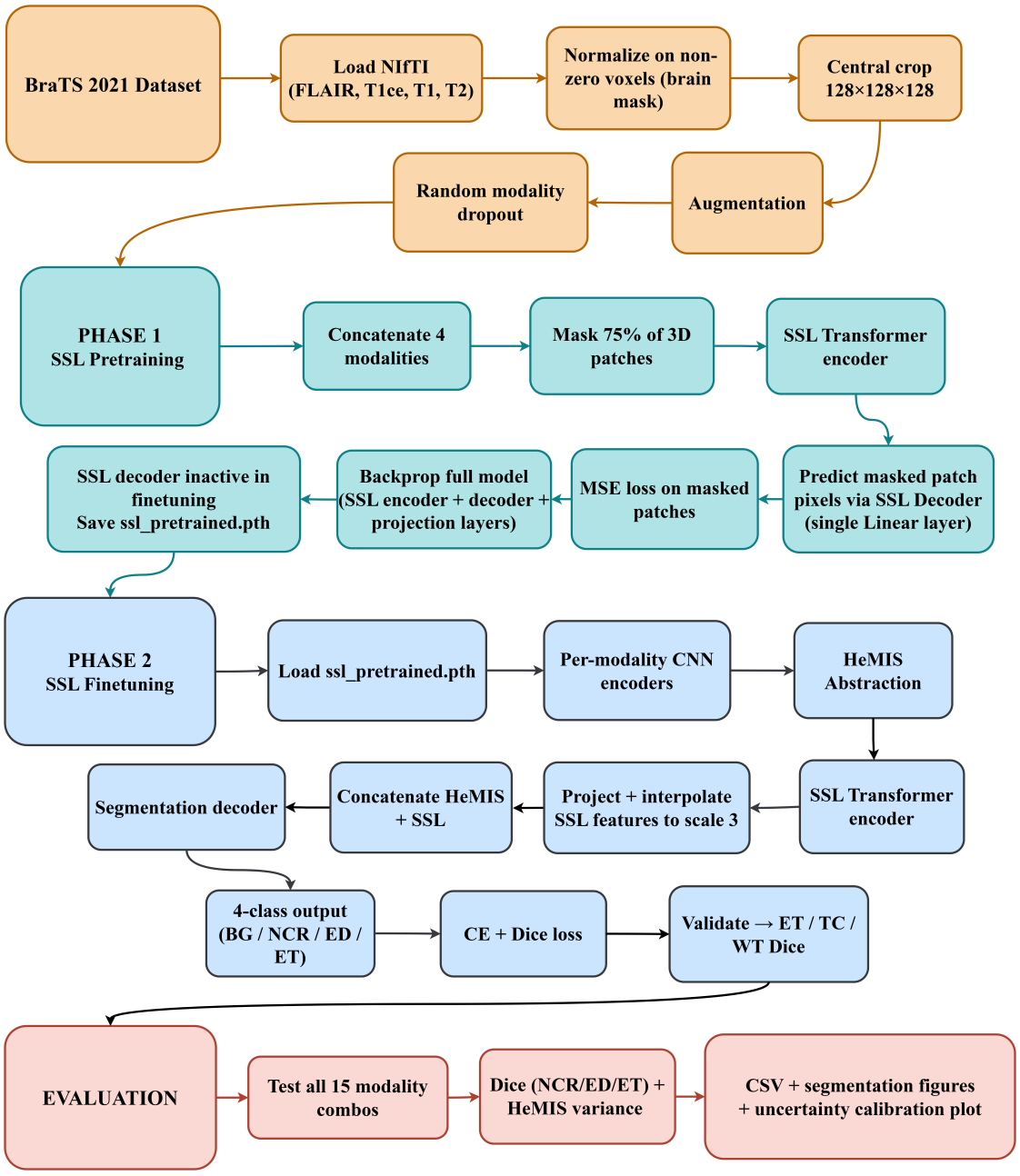


Figure 4.1: Workflow of the proposed Framework

reported as a feature-disagreement signal alongside Dice scores.

As shown in Figure 4.1, summarizes the workflow of the proposed pipeline. The pipeline consists of two phases. In Phase 1 (SSL Pretraining), four MRI modalities are concatenated and 75% of 3D patches are randomly masked. The SSL Transformer encoder processes visible patches and an MLP decoder reconstructs masked patches via MSE loss. The decoder is discarded after pretraining. In Phase 2 (Finetuning), pretrained encoder weights are loaded and the full model is trained on labeled data. Each modality passes through a dedicated CNN encoder producing multi-scale features. The HeMIS abstraction layer computes mean and variance across available modalities, providing explicit uncertainty estimation when modali-

ties are missing. SSL features are projected and fused with HeMIS features before the segmentation decoder produces 4-class output (BG, NCR, ED, ET). During evaluation, all 15 modality combinations are tested and HeMIS variance is reported as a confidence measure alongside Dice scores.

4.2 Design (Model) Specification

Table 4.1 presents a comparative analysis of alternative self-supervised learning architectures evaluated in this study, including 3D SimCLR, Med3D, and 3D DINO. Each model is assessed based on its ability to handle missing MRI modalities, architectural approach, and computational efficiency. The comparison justifies the selection of HeMIS+SSL as the proposed framework for brain tumor segmentation under incomplete modality conditions.

Alternative Design	**Approach**	**Why We Considered It**	**Why We Rejected**
SimCLR (3D)	Contrastive learning (3D); treats augmented views of same image as positive pairs; uses 3D ResNet encoder	Well-established SSL baseline in medical imaging; adapted to 3D volumes	Cannot handle missing MRI modalities; requires all 4 modalities at test time
Med3D	Transfer learning on 3D medical images; pretrained on large CT/MRI datasets; uses 3D ResNet backbone	Specifically designed for 3D medical imaging; leverages large-scale pretraining	No missing modality handling; no uncertainty estimate; requires complete modality set for inference
HeMIS+SSL (Selected)	HeMIS statistical fusion (mean+variance) + MIM pretraining; CNN encoders per modality	Handles missing modalities explicitly; provides uncertainty via variance; lightweight (26.5M parameters)	Slightly more parameters than HeMIS-only, but justified by handling all 15 modality combinations with one model

Table 4.1: Comparative Analysis of Alternative Architectural Designs

4.2.1 SimCLR Model Specification

Input Modalities: The alternative model design was developed for multimodal brain tumor segmentation using the BraTS 2021 dataset. Four MRI modalities were used as input: T1-weighted MRI, contrast-enhanced T1-weighted MRI (T1ce), T2-weighted MRI, and FLAIR. Each modality provides complementary information about the tumor structure and surrounding brain tissues. FLAIR and T2 are useful for identifying edema and abnormal tissue regions, while T1ce helps to highlight enhancing tumor regions.

The four modalities were combined as separate input channels and passed into the 3D segmentation network. In the implementation, the channel order was FLAIR, T1ce, T1, and T2. The four MRI modalities are stacked along the channel dimension to form a single 3D input tensor, formally expressed as:

$$X = \Big[X_{\text{FLAIR}} \parallel X_{\text{T1ce}} \parallel X_{\text{T1}} \parallel X_{\text{T2}} \Big] \in \mathbb{R}^{4\times 128\times 128\times 128}$$

Here, X represents the complete input tensor supplied to the segmentation network, where $\parallel$ denotes the concatenation operator along the channel axis. X_{FLAIR}, X_{T1ce}, X_{T1}, and X_{T2} represent the individual 3D MRI volumes from the FLAIR, T1ce, T1, and T2 modalities respectively, each of shape $\mathbb{R}^{1\times 128\times 128\times 128}$. The first dimension, 4, represents the total number of input channels, while $128 \times 128 \times 128$ represents the cropped 3D patch size used during training and validation. This formula formally demonstrates that the model receives multimodal MRI information as a single, structurally unified, four-channel volumetric input.

The segmentation labels were mapped into four classes: background, NCR/NET, edema, and enhancing tumor. This remapping converted the original BraTS label 4 into class 3. The original BraTS enhancing tumor label 4 is remapped to label 3 so that the target mask has four consecutive classes, formally defined as:

$$\tilde{y}_i = \begin{cases} 3 & \text{if } y_i = 4 \\ y_i & \text{otherwise} \end{cases}$$

Here, y_i denotes the original BraTS label of voxel i, and $\tilde{y}_i$ denotes the remapped label used by the model. The original BraTS label 4, which corresponds to enhancing tumor, is converted to class 3. All other labels remain unchanged. After this remapping, the final classes are 0 for background, 1 for NCR/NET, 2 for edema, and 3 for enhancing tumor. This formula ensures that the segmentation network processes a mathematically continuous index range of $[0, 3]$.

Network Architecture: The alternative model follows a 3D encoder-decoder segmentation architecture. The encoder is based on a 3D ResNet-18-style architecture, which extracts hierarchical volumetric features from the four-channel MRI input. The network uses 3D convolutional operations so that spatial information is learned across depth, height, and width axes.A 3D convolution extracts volumetric features across depth, height, and width, formally expressed as:

$$Y_{o,d,h,w} = \sum_{c=1}^{C_{\text{in}}} \sum_{u=0}^{k_D-1} \sum_{v=0}^{k_H-1} \sum_{r=0}^{k_w-1} W_{o,c,u,v,r} \cdot X_{c,d+u,h+v,w+r} + b_o$$

Here, X is the input feature map and Y is the output feature map produced by the convolution. The index o represents the output channel, c represents the input channel, and d, h, and w represent the depth, height, and width positions. W is the 3D convolution kernel with spatial dimensions K_D, K_H, and K_W, while b_o is the optional bias term for output channel o. This formula explains how the model learns spatial features from 3D MRI patches instead of treating slices independently.

To improve stability during 3D small-batch training, Group Normalization was used instead of Batch Normalization in the final model. Group Normalization normalizes features within channel groups to stabilize small-batch 3D training.

$$\hat{x}_i = \frac{x_i - \mu_G}{\sqrt{\sigma_G^2 + \epsilon}}$$
$$y_i = \gamma \hat{x}_i + \beta$$

Here, x_i is an input activation and $\hat{x}_i$ is its normalized value within group G. μ_G and σ_G^2 are the mean and variance of the activations inside the same normalization group. ϵ is a small constant used for numerical stability. γ and β are learnable scale and shift parameters. This formula is relevant because the final model uses Group Normalization instead of Batch Normalization to reduce instability during small-batch 3D medical image training. This helped reduce validation instability and improved the consistency of the segmentation output.

SSL Framework Used: Self-supervised learning was used before the supervised segmentation stage. The model used a SimCLR-based 3D contrastive learning framework during pretraining. In this stage, two differently augmented 3D views of the same MRI volume were passed through the shared encoder.

A projection head then mapped the encoder features into a lower-dimensional representation space. The contrastive loss encouraged features from the same subject to be closer, while features from different subjects were pushed apart. The NT-Xent loss pulls two augmented views of the same subject closer and pushes different subjects apart.

$$s_{ij} = \frac{\tilde{z}_i^\top \tilde{z}_j}{\tau}$$
$$\mathcal{L}_i = -\log \frac{\exp(s_{i,p(i)})}{\sum_{k \neq i} \exp(s_{ij})}$$
$$\mathcal{L}_{SimCLR} = \frac{1}{2n} \sum_{i=1}^{2n} \mathcal{L}_i$$

Here, $\tilde{z}_i$ and $\tilde{z}_j$ are L_2-normalized projection vectors produced from two augmented views. $s_{i,j}$ is the temperature-scaled similarity score between two projected features, and τ is the temperature parameter, set to 0.07 in the code. $p(i)$ denotes the positive pair index of sample i, meaning the other augmented view from the same MRI subject. N is the batch size before augmentation, so $2N$ projected samples are used in the contrastive batch. This loss encourages representations from the same subject to become similar while separating representations from different subjects. This allowed the encoder to learn useful anatomical and tumor-related representations before supervised fine-tuning.

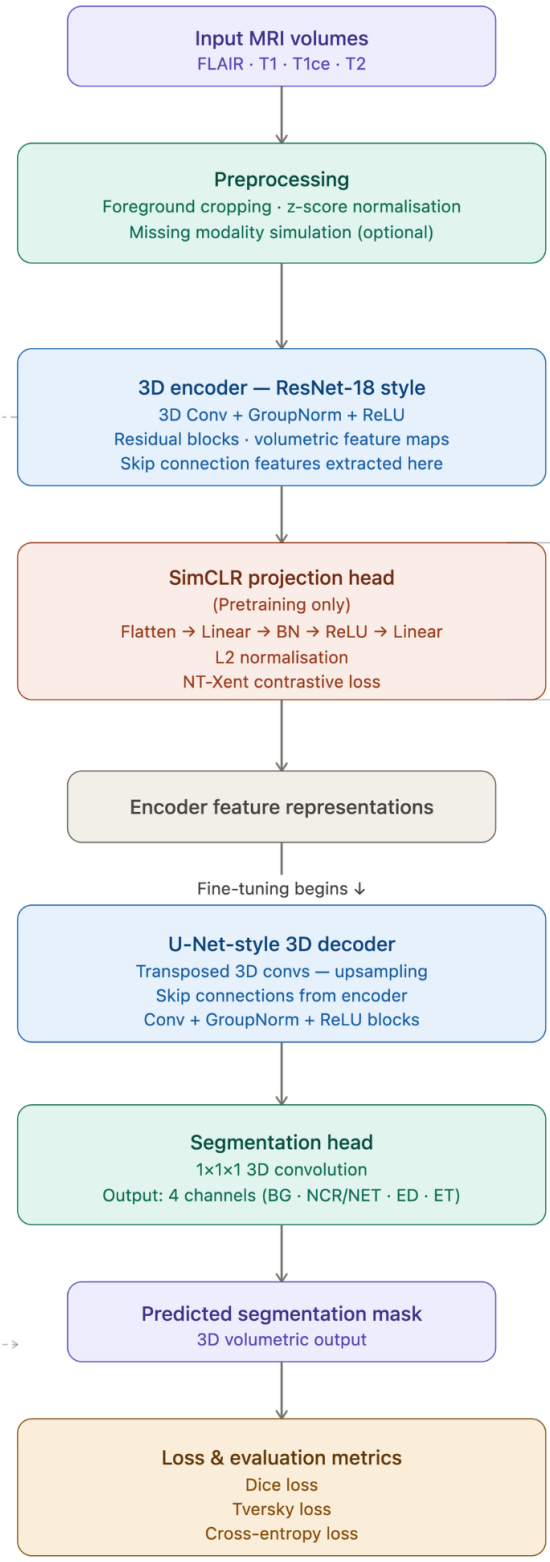


Figure 4.2: SimCLR3D Architecture

Figure 4.2 illustrates the SimCLR3D brain tumor segmentation pipeline. Four MRI modalities (FLAIR, T1, T1ce, and T2) are first preprocessed with foreground cropping, z-score normalization, and optional modality simulation. The preprocessed volumes are passed through a 3D ResNet-style encoder to extract hierarchical volu-

metric features. During pretraining, features are mapped through a SimCLR projection head using contrastive loss, while for supervised fine-tuning, the encoder feeds a U-Net-like 3D decoder with skip connections. The decoder output passes through a 1×1×1 convolutional segmentation head to produce voxel-wise predictions for four classes. Final evaluation is performed using Dice, Tversky, and Cross-Entropy metrics.

Encoder-Decoder Structure: The encoder consists of an initial 3D convolutional stem followed by multiple residual blocks. A residual block adds the learned transformation to a shortcut connection before activation.

$$Y = \mathrm{ReLU}\big(F(X) + S(X)\big)$$

Here, X is the input feature map to the residual block and Y is the output feature map. $F(X)$ represents the main convolutional transformation learned by the residual branch, while $S(X)$ represents the shortcut path. When the input and output dimensions are the same, $S(X)$ can behave like an identity mapping; when the dimensions differ, a $1 \times 1 \times 1$ convolution is used in the shortcut path. This formula supports the ResNet-style encoder used in the model.

These blocks progressively downsample the input and extract deeper semantic features. The decoder follows a U-Net-like structure and uses transposed 3D convolutions for upsampling. Skip connections from the encoder are connected to the decoder at corresponding spatial levels to recover fine-grained spatial information. Finally, the decoder output is passed through a $1 \times 1 \times 1$ convolutional segmentation head to produce four output channels corresponding to the target segmentation classes.

Training Strategy: The training process was performed in two phases:

1. **Phase 1 – Contrastive Pretraining:** The encoder was pretrained using the SimCLR self-supervised learning framework. This phase did not require segmentation labels and was used to initialize the encoder with useful volumetric feature representations.

2. **Phase 2 – Supervised Fine-Tuning:** The pretrained encoder was attached to the decoder and fine-tuned for supervised brain tumor segmentation. During fine-tuning, the full encoder-decoder network was trained using BraTS segmentation labels.

The data pre-processing pipeline included loading NIfTI MRI volumes, remapping BraTS label 4 to label 3, foreground-based normalization, and cropping patches of size $128 \times 128 \times 128$ voxels. The normalized foreground values were also clipped to the range $[-5, 5]$ in the implementation.

$$\Omega = \{i \mid x_i > 0\}$$

$$\hat{x}_i = \frac{x_i - \mu_\Omega}{\sigma_\Omega + \epsilon}$$

Here, x_i is the original intensity value of voxel i and $\hat{x}_i$ is the normalized intensity value. Ω denotes the foreground voxel set, which contains voxels with intensity greater than zero. μ_Ω and σ_Ω are the mean and standard deviation computed only from the foreground voxels. ϵ is a very small constant, 10^{-8} in the code, added to avoid division by zero. This formula standardizes the intensity distribution of each cropped MRI patch while avoiding background voxels.

During training, tumor-focused foreground cropping was used so that the model received patches containing tumor regions more frequently. Random flipping was used as data augmentation. Validation was performed using deterministic tumor-centered cropping to make the evaluation stable and comparable.

Loss Functions: The fine-tuning stage used a combined segmentation loss to address class imbalance and improve tumor-region overlap. The final loss function combined weighted Cross-Entropy loss, Dice loss, and Tversky loss.

$$\mathcal{L}_{\text{seg}} = 0.2\mathcal{L}_{\text{WCE}} + 0.4\mathcal{L}_{\text{Dice}} + 0.4\mathcal{L}_{\text{Tversky}}$$

Here, $\mathcal{L}_{\text{seg}}$ is the final loss used during supervised fine-tuning. $\mathcal{L}_{\text{WCE}}$ is the Weighted Cross-Entropy loss, $\mathcal{L}_{\text{Dice}}$ is the Dice loss, and $\mathcal{L}_{\text{Tversky}}$ is the Tversky loss. The coefficients 0.2, 0.4, and 0.4 represent the contribution of each loss component. This combined loss allows the model to learn voxel-wise class prediction, improve region overlap, and reduce missed tumor voxels at the same time.

Weighted Cross-Entropy performs voxel-wise classification while giving higher importance to tumor classes.

$$\mathcal{L}_{\text{WCE}} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log(p_{i,y_i})$$

Here, N is the total number of voxels used in the loss calculation. y_i is the ground-truth class label of voxel i, and p_{i,y_i} is the predicted softmax probability for the correct class. w_{y_i} is the class weight assigned to the ground-truth class of voxel i. In the implementation, the class weights are $[0.001, 5.0, 4.0, 6.0]$ for background, NCR/NET, edema, and enhancing tumor respectively. This weighting reduces the dominance of the background class and increases the contribution of tumor regions during training.

Dice loss improved overlap between predicted and ground-truth masks, and Tversky loss gave a stronger penalty to missed tumor voxels. Tversky loss balances false positives and false negatives, with stronger penalty on missed tumor voxels.

$$TI_c = \frac{TP_c + \epsilon}{TP_c + \alpha FP_c + \beta FN_c + \epsilon}$$

$$\mathcal{L}_{\text{Tversky}} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}(1 - TI_c)$$

Here, TI_c is the Tversky index for tumor class c. TP_c, FP_c, and FN_c represent true positives, false positives, and false negatives for class c respectively. α controls

the penalty for false positives and β controls the penalty for false negatives. In the implementation, $\alpha = 0.3$ and $\beta = 0.7$, so missed tumor voxels receive a stronger penalty than extra predicted tumor voxels. The final Tversky loss is computed as one minus the Tversky index and averaged over the tumor classes. Since tumor regions are much smaller than the background, the background class was given a very low weight, while tumor classes were assigned higher weights. This helped prevent the model from collapsing into background-only predictions.

Hyperparameters and Experimental Performance: The baseline model was trained using a patch size of $128 \times 128 \times 128$ voxels across the four input MRI modalities. Fine-tuning was performed for 50 epochs using the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 0.01, scheduled over a warmup and cosine learning rate policy (5 warmup epochs).

$$\text{Mean Dice} = \frac{\text{Dice}_{\text{NCR/NET}} + \text{Dice}_{\text{ED}} + \text{Dice}_{\text{ET}}}{3}$$

Here, $\text{Dice}_{\text{NCR/NET}}$, Dice_{ED}, and Dice_{ET} are the class-wise Dice scores for necrotic/non-enhancing tumor core, edema, and enhancing tumor, respectively. The denominator 3 indicates that the average is computed over the three tumor classes only, excluding the background class. This metric is used to report the overall tumor segmentation performance of the model.

The batch size was adjusted according to GPU memory availability, and safer training environments were maintained using zero DataLoader workers when memory constraints occurred. The final architectural configuration parameters and validation performance metrics are summarized in Table 4.2.

Specification Dimension	**Configured Parameter Variable**	**Validation Metric Value**
Input Context	128 × 128 × 128 patches, 4 modalities	Base channel scale: 32
Optimizer Settings	AdamW, lr $= 1 \times 10^{-4}$, weight decay: 0.01	50 total epochs (5 warmup)
Core Optimization Losses	Combined Cross-Entropy + Dice + Tversky loss functions	Background class down-weighted to limit collapse
Overall Performance	Mean Validation Overlap Metric	Mean Dice Score: 0.7236
Class-Wise Evaluation	Necrotic Core / Non-Enhancing Tumor (NCR/NET)	Class Dice Score: 0.6473
Class-Wise Evaluation	Peritumoral Edema (ED subregion)	Class Dice Score: 0.7590
Class-Wise Evaluation	Enhancing Tumor (ET subregion)	Class Dice Score: 0.7646

Table 4.2: Alternative Design: SimCLR (3D) Hyperparameters and Validation Results

4.2.2 Med3D Model Specification

Input Modalities: The input of the proposed Med3D baseline model consists of four MRI modalities: T1, T1ce, T2, and FLAIR. These modalities provide complementary information for brain tumor segmentation. T1-weighted MRI provides anatomical structure, T1ce highlights contrast-enhancing tumor regions, T2 helps identify tumor-related abnormalities, and FLAIR is especially useful for detecting edema regions. By combining all four modalities, the model receives a more complete representation of the tumor and surrounding brain tissue.

In the implementation, the four modalities are stacked together as input channels. Therefore, each input sample is represented as a four-channel 3D volume with the shape $[4, D, H, W]$, where D, H, and W represent the depth, height, and width of the MRI volume, respectively. During training, the model receives batches in the form $[B, 4, D, H, W]$, where B is the batch size. This formula represents how the four MRI modalities are stacked as one 3D input sample.

$$X_i = \left[X_i^{\mathrm{T1}}, X_i^{\mathrm{T1CE}}, X_i^{\mathrm{T2}}, X_i^{\mathrm{FLAIR}}\right] \in \mathbb{R}^{4\times D\times H\times W}$$

$$\mathbf{X} \in \mathbb{R}^{B\times 4\times D\times H\times W}$$

Here, X_i denotes the input MRI volume for the i-th patient case. X_i^{T1}, X_i^{T1CE}, X_i^{T2}, and X_i^{FLAIR} denote the four MRI modalities used in the Med3D experiment. D, H, and W represent the depth, height, and width of the 3D MRI volume. The value 4 represents the four input channels. B is the batch size, so during training the input batch is represented as $[B, 4, D, H, W]$.

Before training, each modality was normalized using z-score normalization over non-zero brain voxels so that background voxels did not affect the intensity statistics. This formula normalizes each MRI modality using only non-zero brain voxels while preserving the zero-valued background.

$$\Omega_m = \{v \mid X_m(v) > 0\}$$

$$\tilde{X}_m(v) = \frac{X_m(v) - \mu_m}{\sigma_m + \epsilon}$$

$$v \in \Omega_m, \quad \tilde{X}_m(v) = 0, \quad v \notin \Omega_m$$

Here, $X_m(v)$ is the original intensity value of voxel v in modality m, where m can be T1, T1CE, T2, or FLAIR. Ω_m represents the set of non-zero voxels for modality m, which corresponds to the foreground brain region. μ_m and σ_m are the mean and standard deviation calculated only from the non-zero voxels. ϵ is a very small constant used to avoid division by zero. The normalized value is denoted by $X'_m(v)$, and background voxels outside Ω_m are kept as zero.

The original BraTS label 4 was remapped to class 3, producing four consecutive class indices $0, 1, 2$, and 3. This formula converts the original BraTS label 4 into class 3 so that the segmentation classes become consecutive.

$$y'(v) = \begin{cases} 3, & \text{if } y(v) = 4 \\ y(v), & \text{if } y(v) \in \{0, 1, 2\} \end{cases}$$

$$\tilde{y}(v) \in \{0, 1, 2, 3\}$$

Here, $y(v)$ is the original ground-truth BraTS label at voxel v, and $y'(v)$ is the remapped label used by the model. The original BraTS labels are $0, 1, 2$, and 4. Since PyTorch multi-class cross-entropy expects consecutive class indices, label 4 is mapped to class 3. After remapping, the four classes are 0 for background and $1, 2$, and 3 for the three tumor-related classes.

Network Architecture: The proposed model uses a Med3D based 3D convolutional neural network architecture for volumetric segmentation. Unlike 2D CNNs, which process slices independently, the 3D network processes volumetric MRI data and learns spatial information across axial, coronal, and sagittal directions simultaneously. This formula describes how a 3D convolution learns volumetric features across depth, height, and width.

$$Y_{c_o}(d, h, w) = \sum_{c_i=1}^{C_i} \sum_{p=0}^{K_D-1} \sum_{q=0}^{K_H-1} \sum_{r=0}^{K_W-1} W_{c_o,c_i,p,q,r} \, X_{c_i,d+p,h+q,w+r} + b_{c_o}$$

Here, X is the input feature map and Y is the output feature map after applying a 3D convolution. c_i and c_o denote the input and output channels, respectively. d, h, and w are voxel locations along depth, height, and width. W represents the learnable convolution kernel, and b is the bias term. K_D, K_H, and K_W are the kernel sizes along the three spatial dimensions. This operation is important because it allows the model to learn volumetric tumor features rather than processing 2D slices independently. This is important for brain tumor segmentation because tumor boundaries and sub-regions are inherently three-dimensional structures.

The model contains an encoder-decoder design. The encoder extracts hierarchical features from the four-channel input volume, while the decoder reconstructs the segmentation prediction at the voxel level. The first convolutional layer was modified to accept four input channels instead of one, allowing the model to process T1, T1ce, T2, and FLAIR simultaneously. The final output layer predicts four segmentation classes, including the background class and three tumor classes.

SSL Framework Used: No self-supervised learning framework was used in this Med3D experiment. The model was trained using a fully supervised segmentation strategy with paired MRI volumes and ground-truth segmentation masks.

Although SSL methods such as SimCLR can be used for medical image representation learning, this particular Med3D experiment does not include a contrastive pretraining stage. Instead, the network learns directly from the labeled BraTS training data. This distinction is important because the Med3D result should be compared as a supervised baseline rather than as an SSL-based model.

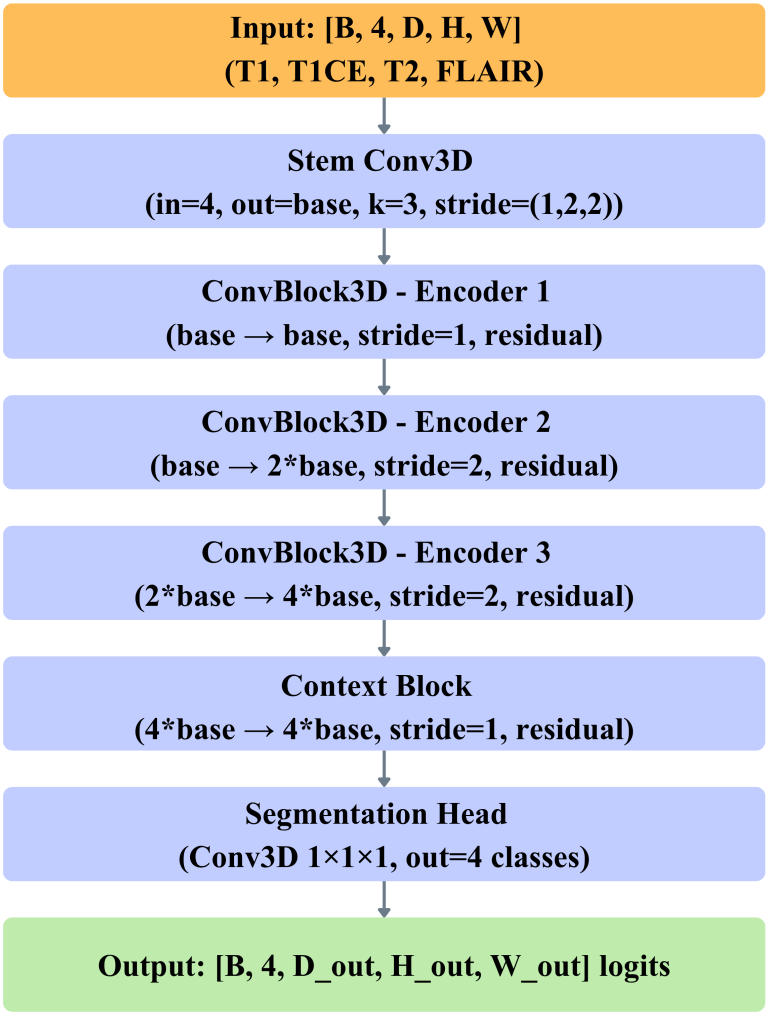


Figure 4.3: Med3D Architechture

Figure 4.3 illustrates the Med3D 3D ResNet-based segmentation network for four-modality BraTS inputs. The model takes a 4-channel 3D MRI volume (T1, T1CE, T2, FLAIR) and processes it through a stem 3D convolution, followed by three residual ConvBlock3D encoders that progressively increase channels and downsample spatial dimensions. A context block further refines features before the segmentation head (1×1×1 Conv3D) outputs 4-class voxel-wise logits. The network is designed to predict tumor subregions while maintaining memory efficiency on 3D volumes.

Encoder-Decoder Structure: In the encoder part of the network, the network extracts progressive high level spatial and semantic features of the input MRI volumes. Learns volumetric features from all four modalities using 3D Convolutional operations. While the data is flowing through the encoder, the network is able to detect the intensity changes, the shape, location and the difference between normal and abnormal tissues within the structure of the tumor.

The decoder portion re-combines the segmentation output of the encoded feature representation. It maps the learned features back to per voxel class predictions. The final layer gives a 4 channel output, one channel per segmentation class. The

class having the maximum likelihood is chosen for each voxel. The encoder-decoder architecture provides the model with global contextual information and local tumor boundaries.

Training Strategy: The model was trained with the training split from the BraTS 2021 with 1,001 cases, while 250 cases were held out for validation and testing. The training was performed on an NVIDIA RTX 4080 SUPER GPU using CUDA acceleration. The final 4-modality experiment was used with a batch size of 2 due to stable training with the available GPU memory.

The model was trained for a maximum of 50 epochs with early stopping. Early stopping was applied to prevent unnecessary training when validation performance stopped improving. The best checkpoint was selected based on the mean tumor Dice score over classes 1, 2, and 3, excluding the background class. This formula averages the Dice scores of the three tumor classes and excludes the background class.

$$\text{Mean Tumor Dice} = \frac{\text{Dice}_1 + \text{Dice}_2 + \text{Dice}_3}{3}$$

Here, Dice_1, Dice_2, and Dice_3 are the Dice scores for the three tumor-related classes after label remapping. The background class, class 0, is excluded from this average. This is important because background voxels dominate medical segmentation volumes and can make performance look better than it actually is. The implemented training script uses mean tumor Dice as the best-checkpoint selection metric.

The best checkpoint was saved at epoch 30 with a validation mean tumor Dice score of 0.7311. Training stopped at epoch 42 because there was no further validation improvement for 12 consecutive epochs (patience = 12).

Loss Function: The model was optimized using weighted cross-entropy loss function suitable for multi-class medical image segmentation. The loss function compares the predicted segmentation mask with the ground-truth label mask and penalizes incorrect voxel-wise predictions. This formula defines the supervised voxel-wise loss used to train the multi-class Med3D segmentation model.

$$p_c(v) = \frac{\exp\left(z_c(v)\right)}{\sum_{c'=1}^{C} \exp\left(z_{c'}(v)\right)}$$

$$\mathcal{L}_{\text{WCE}} = -\frac{1}{|\Omega|} \sum_{v \in \Omega} \sum_{c=0}^{C-1} w_c \cdot \mathbb{I}\left(\tilde{y}(v) = c\right) \cdot \log\left(p_c(v)\right), w = [0.20, 1.0, 1.0, 1.0]$$

Here, $\mathcal{L}_{\text{WCE}}$ is the weighted cross-entropy loss. Ω represents the set of voxels used for training. $z_c(v)$ is the model logit for class c at voxel v, and $p_c(v)$ is the softmax probability for that class. C is the total number of classes, which is 4 in this experiment. $\tilde{y}(v)$ is the ground-truth class label after BraTS label remapping. $\mathbb{I}\left(\tilde{y}(v) = c\right)$ is an indicator function that becomes 1 when the voxel belongs to class c and 0 otherwise. w_c is the class weight. In the implemented code, the background class has weight 0.20, while the three tumor classes have weight 1.0 each.

Since brain tumor segmentation is highly imbalanced, background voxels are much more common than tumor voxels. Therefore, Dice-based evaluation was used as the main performance metric because it is more appropriate than simple accuracy for imbalanced segmentation tasks. This formula measures the overlap between the predicted segmentation region and the ground-truth region for each class.

$$\text{Dice}_c = \frac{2|P_c \cap G_c|}{|P_c| + |G_c|} = \frac{2TP_c}{2TP_c + FP_c + FN_c}$$

Here, Dice_c is the Dice score for class c. P_c is the set of voxels predicted as class c, and G_c is the set of ground-truth voxels belonging to class c. The numerator measures twice the overlap between prediction and ground truth, while the denominator is the total size of the predicted and ground-truth regions. TP_c, FP_c, and FN_c represent true positives, false positives, and false negatives for class c. A higher Dice score indicates better segmentation overlap.

The best model was selected using the mean tumor Dice score. This score was calculated by averaging the Dice scores of the three tumor classes. The background class was not used for best-model selection, because high background accuracy can give a misleading impression of performance in medical segmentation.

Hyperparameters and Evaluation Results: The final 4-modality Med3D model was trained with a model depth of 50 using a ResNet shortcut type B configuration. Eight data-loading workers were used to optimize data throughput speed during training execution.

The backbone follows a residual learning strategy, where the convolutional transformation is added to a skip connection to improve feature propagation. This formula shows how the residual block adds a skip connection to the convolutional transformation.

$$Y = \phi\big(F(X) + S(X)\big)$$
$$S(X) = X \quad or \quad S(X) = Conv_{1\times1\times1}(X)$$

Here, X is the input feature map to the residual block and Y is the output feature map. $F(X)$ represents the main convolutional transformation inside the block. $S(X)$ represents the skip connection. If the input and output dimensions match, $S(X)$ is simply the identity mapping X. If the number of channels or spatial dimensions changes, a $1 \times 1 \times 1$ convolution is used in the skip path. ϕ denotes the activation function applied after adding the transformed feature and the skip connection.
The model was evaluated using corrected evaluation pipelines across two separate modes: *train-crop* (matching training validation styling) and *brain-crop* (deterministic foreground brain cropping optimized). The complete configuration parameters and validation performance metrics are summarized in Table 4.3.

Specification Dimension	Configured Parameter Variable	Validation Metric Value
Input Tensors	Stacked $[B, 4, D, H, W]$ multi-modal volumes	4 channels (T1, T1ce, T2, FLAIR)
Optimizer Settings	Learning rate: 1×10^{-4}, Batch Size: 2	Max 50 epochs (Stopped at 42)
Model Backbone	Med3D ResNet-50 style with Type B shortcuts	8 DataLoader worker processes
Early Stopping Context	Selected at Peak Mean Tumor Validation Overlay	Best Checkpoint: Epoch 30
Baseline Performance	Validation Mean Tumor Dice Score (Epoch 30)	Peak Validation Dice: 0.7311
Pipeline Evaluation	Train-Crop Global Mean Tumor Target	Global Dice Score: 0.7270
Pipeline Evaluation	Brain-Crop Global Mean Tumor Target	Global Dice Score: 0.7043
Pipeline Evaluation	Brain-Crop Per-Case Mean Tumor Target	Per-Case Dice Score: 0.5849

Table 4.3: Alternative Design: Med3D (Supervised) Hyperparameters and Validation Results

Architecture Description: This baseline workflow starts directly from the BraTS 2021 dataset, and it works in a succession of steps. First, for each patient case, the four MRI modalities (T1, T1ce, T2, FLAIR) are loaded. The four modalities will be preprocessed and label remapped and then stacked into a 4 channel 3D input volume.

This input tensor is directly fed into the Med3D encoder that learns hierarchical volumetric features in all the three spatial dimensions. The decoder then converts the localized feature maps back into voxel-wise predictions of classes, providing a unified four-class tumor mask that localizes background and tumor sub-regions.

4.2.3 Hybrid HeMIS-SSL Model Specification

The proposed hybrid model combines Hetero-Modal Image Segmentation (HeMIS) with a Self-Supervised Learning (SSL) framework based on Masked Image Modeling (MIM). The model is designed for 3D brain tumor segmentation using multi-modal MRI data while maintaining robustness when one or more MRI modalities are unavailable. The final system uses HeMIS abstraction to handle missing modalities and an SSL transformer encoder to learn global volumetric MRI representations before supervised segmentation.

Input Modalities: The model uses four MRI modalities from the BraTS 2021 dataset: T1-weighted MRI, T1-weighted contrast-enhanced MRI, T2-weighted MRI, and FLAIR MRI. In the implementation, these modalities are loaded in the order FLAIR, T1ce, T1, and T2. Each modality is represented as a single-channel 3D volume. After pre-processing, the input volume size is standardized to $128 \times 128 \times 128$

voxels. Therefore, when all modalities are available, the SSL branch receives a four-channel input volume of shape $[B, 4, 128, 128, 128]$, where B represents the batch size.

The segmentation output consists of four voxel-wise classes: background or healthy tissue, necrotic and non-enhancing tumor core (NCR/NET), peritumoral edema (ED), and enhancing tumor (ET). The final model produces four-channel segmentation logits, which are converted into a discrete segmentation mask by selecting the class with the highest predicted probability for each voxel.

Data Loading Pipeline: An efficient data loading pipeline was implemented to maximize GPU utilization and minimize idle time between batches. The PyTorch DataLoader was configured with the following settings:

- `pin_memory=True`: enables faster CPU-to-GPU memory transfers by pinning host memory.
- `prefetch_factor=2`: allows worker processes to prefetch two batches ahead, reducing GPU wait time.
- `persistent_workers=True`: keeps worker processes alive across epochs, eliminating the overhead of spawning new processes at each epoch start.
- `non_blocking=True`: enables asynchronous GPU data transfers, allowing computation to overlap with memory transfers.
- A batch size of 4 was used during SSL pretraining. A batch size of 1 was used during segmentation finetuning and evaluation, due to the high memory requirements of 3D volumetric inputs at $128 \times 128 \times 128$ resolution.

Network Architecture: The proposed network follows a hybrid HeMIS + SSL fusion architecture. The HeMIS component is responsible for extracting modality-specific features and fusing information from available modalities, while the SSL component provides global contextual representations learned through masked patch reconstruction.

The architecture consists of five main components:

1. Four independent CNN-based modality encoders extract multi-scale features from the available MRI modalities.
2. HeMIS abstraction layers compute the mean and variance of available modality features at each encoder scale.
3. An SSL transformer encoder processes the concatenated multi-modal MRI input and generates spatial representation features.
4. An SSL projection layer aligns the SSL feature representation with the HeMIS bottleneck features.
5. A U-Net-style segmentation decoder fuses the HeMIS and SSL features to generate the final voxel-wise tumor segmentation output.

This design allows the same trained model to operate with complete MRI input as well as incomplete modality combinations. Missing modalities are replaced by zero tensors in the SSL input, while the HeMIS abstraction excludes missing modality features from mean and variance computation. This enables the model to handle all non-empty combinations of the four MRI modalities.

SSL Framework Used: The SSL framework used in the proposed hybrid model is Masked Image Modeling. During SSL pretraining, the model does not use segmentation masks. Instead, the four MRI modalities are concatenated into a single multi-channel 3D input volume and divided into non-overlapping 3D patches. A subset of these patches is randomly masked, and the SSL encoder-decoder branch is trained to reconstruct the missing patch information from the visible surrounding context.

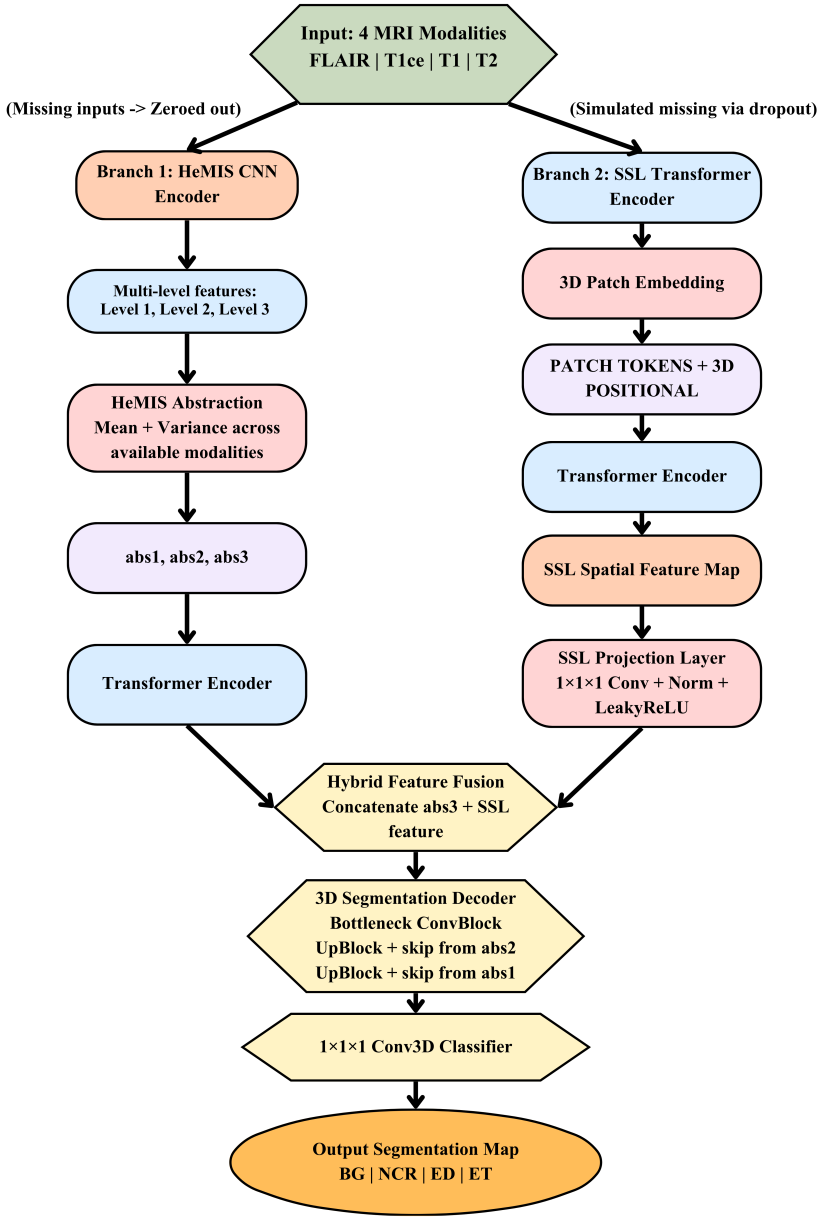


Figure 4.4: Hemis+SSL Architecture

Figure 4.4 illustrates the proposed hybrid HeMIS-SSL model for multimodal brain tumor segmentation. The model receives four MRI modalities: FLAIR, T1ce, T1, and T2. The HeMIS CNN branch extracts multi-level modality-specific features and applies mean–variance abstraction to handle missing modalities, producing abs1, abs2, and abs3. In parallel, the SSL Transformer branch converts the multimodal MRI input into 3D patch tokens with positional encoding and learns global spatial representations. The deepest HeMIS feature abs3 is fused with the projected SSL feature, while abs2 and abs1are used as decoder skip connections. Finally, the 3D segmentation decoder and 1×1×1 Conv3D classifier generate the output segmentation map for BG, NCR, ED, and ET tumor classes.

The SSL encoder begins with a 3D patch embedding layer implemented using a 3D convolution. This layer converts the input MRI volume into a sequence of patch tokens. A 3D sinusoidal positional encoding is then added to preserve spatial information along the depth, height, and width dimensions. The masked tokens and visible tokens are passed through a transformer encoder, allowing the model to learn long-range dependencies across the 3D MRI volume.

The SSL decoder is a light and simple linear reconstruction head. It takes in the encoded masked token representations, and it predicts the value of the voxels of the corresponding masked MRI patches. All those losses are relative to the target: the reconstruction target is normalized per masked patch, and the MIM loss is calculated as the mean squared error between the predicted masked patches and the normalized target patches. At the end of the SSL phase, a pretrained checkpoint with learned representation weights is obtained, and then it is used during the supervised segmentation fine-tuning phase.

SSL Pretraining (Masked Image Modeling): A Masked Image Modeling (MIM) strategy is used in the SSL pretraining stage. The model is pretrained on the entire training set from BraTS 2021 with 1,000 cases and no modalities are missing (0 modality missing) in each case. This enables the model to learn complete multi-modal MRI representations, before the fine-tuning phase of the missing modality scenarios.

The CNN modality encoders, the SSL transformer encoder and the SSL decoder are trained together with all other model parameters during pretraining. The CNN encoders learn meaningful MRI feature representations that are used in conjunction with the SSL encoder, instead of learning random feature maps to be compensated for by the SSL encoder.

The target for reconstruction for each masked patch is a set of original pixel values for the patch in all four modality channels. For enhancing the quality of the learning signal, aggregation of targets is performed per patch, i.e., each patch's target vector is normalized by subtracting out the mean and dividing by the standard deviation of that patch. Variance is clamped to be the smallest of 1e-6 so as not to divide by zero for uniform patches. This normalization puts more emphasis on learning relative spatial patterns among patches than absolute intensities so that better learned representations are obtained.

The reconstruction loss is the Mean Squared Error (MSE) calculated only on masked patch positions. The model is optimized with AdamW optimizer and learning rate (LR) of 1e-5 and weight decay (WD) of 0.05. The total number of pretraining epochs is used as Tmax in a CosineAnnealingLR scheduler. Pretraining is done without AMP in full float32 precision. The maximum norm for gradient clipping is 1.0. Batches with any NaN or Inf loss values are detected and skipped without causing the model weights to be corrupted.

The data split has been done at the subject level before any preprocessing, SSL pretraining, supervised fine-tuning, and evaluation in order to avoid data leakage. The BraTS 2021 data set was split into 5 folds. Folds 1–4 (1,000 subjects) served as the training set and Fold 0 (251 subjects) was held out for validation/evaluation. The same split was adhered to throughout both training stages. Thus, the self-supervised MIM encoder was pre-trained with the training subjects, and the supervised segmentation model was fine-tuned with the training subjects. Fold 0 was not applied in any of the learning phases.

Preprocessing was performed in each individual subject. Intensity clipping and z score normalization were performed separately for each volume and each modality based on non-zero foreground voxels of that same subject alone. Global normalization statistics were not calculated using the complete set of data. In this study, data augmentation and random modality dropout were only applied to the training set. Deterministic preprocessing was used with the validation/evaluation set. This prevented any validation data from aiding the learned SSL representation, supervised model weights, data augmentation procedure, or data statistics.

Training Strategy: The training strategy consists of two main stages: SSL pretraining and supervised segmentation finetuning.

In the SSL pretraining stage, all four MRI modalities are used, and the missing modality probability is set to 0.0. The model is trained using the MIM objective, where random 3D MRI patches are masked and reconstructed. This stage helps the SSL encoder learn general volumetric and multi-modal MRI representations before segmentation labels are introduced. SSL pretraining is conducted in full FP32 precision. AdamW optimization is used with weight decay, gradient clipping, and a warmup cosine learning-rate scheduler. The best SSL checkpoint is saved according to the lowest MIM reconstruction loss.

In the supervised finetuning stage, the model is trained for voxel-wise tumor segmentation using segmentation masks. During finetuning, missing modality simulation is applied by independently dropping modalities with a probability of 0.3. A safety rule ensures that at least one modality is always available. This trains the model to remain robust under incomplete MRI input conditions. Finetuning uses mixed precision training with automatic casting and gradient scaling. The model is validated after each epoch, and the checkpoint with the highest mean validation Dice score is saved as the best segmentation model.

During evaluation, the trained model is tested across all 15 possible non-empty modality combinations. This includes single-modality, two-modality, three-modality, and all-four-modality settings. The evaluation reports per-label Dice scores for NCR, ED, and ET, along with the mean foreground Dice. Clinical region scores for ET, Tumor Core, and Whole Tumor are also computed during validation.

Loss Functions: Two different loss functions are used in the two training stages:

- **For SSL pretraining:** The model uses Masked Image Modeling reconstruction loss. After the MRI input is divided into patches and selected patches are masked, the SSL decoder predicts the masked patch values. The target masked patches are normalized by subtracting the patch mean and dividing by the patch standard deviation. The final SSL loss is the mean squared error between the predicted masked patches and the normalized target patches.

$$\hat{x}_i = \frac{x_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}$$

$$\mathcal{L}_{\text{MIM}} = \frac{1}{|\mathcal{M}_{\text{mask}}|} \sum_{i \in \mathcal{M}_{\text{mask}}} \|\tilde{x}_i - \hat{x}_i\|_2^2$$

  The formula teaches the SSL encoder to learn useful MRI representations by reconstructing hidden patch content from the visible surrounding context.

  Here, x_i denotes the original target patch at masked patch position i, and $\hat{x}_i$ denotes the normalized version of that target patch. The term $\tilde{x}_i$ denotes the reconstructed patch predicted by the SSL decoder. The terms μ_i and σ_i^2 denote the mean and variance of the target patch x_i, respectively, and ϵ is a small numerical constant used for stability. The set $\mathcal{M}_{\text{mask}}$, written as calligraphic $\mathcal{M}$ with subscript mask, represents all masked patch positions selected during SSL pretraining, and $|\mathcal{M}_{\text{mask}}|$ denotes the number of masked patches. The notation $\| \cdot \|_2^2$ represents the squared L_2 reconstruction error. $\mathcal{L}_{\text{MIM}}$ is the final Masked Image Modeling reconstruction loss computed only over the masked patches.

- **For supervised segmentation finetuning:** The model uses a combined Cross-Entropy and Dice loss. The Cross-Entropy component provides voxel-wise classification supervision, while the Dice component directly optimizes region overlap. The two components are weighted equally, with a Cross-Entropy weight of 0.5 and a Dice weight of 0.5. To address class imbalance, the Cross-Entropy loss uses class weights of 0.1 for background, 2.0 for NCR, 1.5 for ED, and 2.0 for ET. The Dice loss is computed using softmax probabilities and excludes the background class.

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=0}^{C-1} w_c \, y_{i,c} \log(p_{i,c})$$

$$w = [0.1,\ 2.0,\ 1.5,\ 2.0]$$

The formula gives higher penalty to tumor-class errors than background errors, which helps the model learn small and imbalanced tumor regions.

Here, $\mathcal{L}_{\text{CE}}$ denotes the weighted Cross-Entropy loss. The symbol N denotes the total number of voxels used in the loss calculation, and C denotes the number of segmentation classes, which is 4 in this work. The index i denotes a voxel, and c denotes a class index, where 0 is background, 1 is NCR, 2 is ED, and 3 is ET. The term w_c denotes the weight assigned to class c. The term $y_{i,c}$ is the ground-truth class indicator, which equals 1 when voxel i belongs to class c and 0 otherwise. The term $p_{i,c}$ is the predicted softmax probability that voxel i belongs to class c. The weight vector $w = [0.1, 2.0, 1.5, 2.0]$ corresponds to background, NCR, ED, and ET, respectively.

The formula trains the model to classify each voxel correctly while also improving the overlap between predicted and ground-truth tumor masks.

$$\mathcal{L}_{\text{seg}} = 0.5\mathcal{L}_{\text{CE}} + 0.5\mathcal{L}_{\text{Dice}}$$

Here, $\mathcal{L}_{\text{Dice}}$ denotes the Dice loss used during supervised segmentation finetuning, and $\mathcal{L}_{\text{seg}}$ denotes the final combined segmentation loss. The term $\mathcal{L}_{\text{CE}}$ denotes the weighted Cross-Entropy loss defined above. The symbol C denotes the total number of segmentation classes. Since the background class is ignored in Dice loss, the summation starts from $c = 1$ and continues to $C - 1$, which corresponds to the foreground tumor classes.

The index i denotes a voxel, $p_{i,c}$ denotes the predicted softmax probability for voxel i and class c, and $y_{i,c}$ denotes the one-hot ground-truth label for voxel i and class c. The constant ϵ is the Dice smoothing constant; in the implementation, $\epsilon = 10^{-5}$. The coefficients 0.5 and 0.5 indicate that Cross-Entropy loss and Dice loss contribute equally to the final segmentation objective.

Comprehensive Missing Modality Strategy: There is a comprehensive missing modality strategy used in the training and the evaluation pipeline. The strategy is different for each of the three phases:

- **In the SSL pretraining phase:** `missing_prob=0.0`. The four modalities are always present. This avoids the difficult problem of missing channels in the SSL encoder during pretraining, where it learns full multi-modal MRI representations.
- **During segmentation finetuning (training):** `missing_prob=0.3`. The probabilities for the loss of each modality are 0.3, independent of the others. This allows for any modality to be present and another to be absent within a single training batch, which would be the case in the real world of MRI acquisitions.
- **During validation and evaluation:** `missing_prob=0.0`. All modalities are present and the model is assessed explicitly in all of the 15 possible modality combinations.

There is a safety guarantee: Always at least one modality is present. When all 4 modalities are randomly dropped at the same time, one is randomly chosen and restored. This will avoid degenerate empty inputs.

If a modality is not present in the inference, a zero tensor is used in its place which is of the same shape as a normal modality input. The HeMIS abstraction layer does not include the None value in the mean or variance calculation, and thus does not need to make any changes to the architecture.

Model Finetuning Training Setup: Once SSL is pretrained, the model then goes through supervised segmentation finetuning. The full model is initialized with SSL pretrained weights, which provide strong initial values for all parameters (including CNN encoders and SSL transformer). The SSL decoder is not used in the fine-tuning stage.

All model parameters are updated during finetuning, including the CNN modality encoders, SSL transformer encoder, SSL projection layer, HeMIS abstraction layers, and segmentation decoder. The finetuning configuration is as follows:

- **Optimizer:** AdamW with weight decay of 0.05
- **Learning rate:** 2e-4 with CosineAnnealingLR scheduler ($T_{\max}$ equal to total epochs)
- **Mixed Precision Training:** Automatic Mixed Precision (AMP) with GradScaler for float16 computation during forward and backward passes
- **Gradient unscaling:** `scaler.unscale_(optimizer)` is called before gradient clipping to ensure clipping operates on true gradient magnitudes rather than scaled ones
- **Gradient clipping:** maximum gradient norm of 1.0
- **NaN loss detection:** batches producing NaN loss are skipped without updating model weights
- **Validation:** performed after every epoch to track progress and save the best checkpoint
- **Checkpoint saving:** the model checkpoint with the highest mean validation Dice is saved

Evaluation Protocol: The trained model is evaluated comprehensively across all 15 combinations represent all non-empty subsets of the four BraTS 2021 modalities and are used for simulated missing-modality evaluation. The 15 combinations correspond to all non-empty subsets of the four modalities (FLAIR, T1ce, T1, T2).

Segmentation performance is reported using three clinically relevant metrics derived from the BraTS challenge evaluation protocol:

- **Enhancing Tumor (ET) Dice:** computed on label 3 directly.

- **Tumor Core (TC) Dice:** computed on the union of NCR (label 1) and ET (label 3), i.e., TC = (pred == 1) $\vee$ (pred == 3).
- **Whole Tumor (WT) Dice:** computed on the union of all tumor labels, i.e., WT = (pred $\geq$ 1), encompassing NCR, ED, and ET.

These clinical region metrics allow direct comparison with published methods as they align with the standard BraTS evaluation criteria. The formula gives a score between 0 and 1, where a higher value means better overlap between the predicted tumor segmentation and the ground truth.

$$\text{Dice}(P, T) = \frac{2|P \cap T|}{|P| + |T| + \epsilon}$$

Here, P denotes the predicted segmentation region for a particular tumor label or clinical tumor region, and T denotes the corresponding ground-truth region. The term $|P \cap T|$ represents the number of overlapping voxels between prediction and ground truth. The term $|P|$ represents the total number of predicted voxels in that region, and $|T|$ represents the total number of ground-truth voxels in that region. The constant ϵ is added to avoid numerical instability; in the evaluation code, $\epsilon = 10^{-5}$.

HeMIS variance is calculated for each modality combination as a feature-disagreement signal, in addition to the segmentation metrics. Variance is calculated at the deepest feature level (`enc3`) for all the modality feature maps that are present. If there is only one modality the variance is equal to 0.0 by definition. The variance is correlated with the disagreement between existing modality feature maps and is correlated with the lower Dice score but cannot be interpreted as a calibrated uncertainty score until validated. All results are saved to a CSV file for further analysis.

Hardware and Optimization Settings: All experiments were conducted on an NVIDIA GPU with CUDA support. Several PyTorch backend optimizations were enabled to maximize computational throughput:

- `torch.backends.cudnn.benchmark=True`: enables cuDNN auto-tuner to select the fastest convolution algorithms for fixed input sizes.
- `torch.backends.cuda.matmul.allow_tf32=True`: enables TF32 precision for matrix multiplications on Ampere and later GPUs.
- `torch.backends.cudnn.allow_tf32=True`: enables TF32 precision for convolutions.
- `torch.set_float32_matmul_precision('high')`: sets high precision mode for float32 matrix multiplications, balancing speed and accuracy.

Visualization and Qualitative Analysis: To provide qualitative insights into model behavior, a comprehensive visualization pipeline was implemented. All visualizations use the middle axial slice of each 3D volume for display. A consistent color coding scheme is applied: background in black, NCR in red, ED in green, and

ET in blue.

Two types of segmentation figures are produced for a subset of validation cases:

- **4-panel figure:** displays the FLAIR input (grayscale), ground truth segmentation, prediction using all four modalities, and prediction using FLAIR only. This directly illustrates the performance degradation when modalities are missing.
- **All-combinations figure:** a 17-panel figure displaying the FLAIR input, ground truth, and predictions from all 15 modality combinations. This provides a comprehensive view of how prediction quality degrades as fewer modalities are available.

An uncertainty calibration plot is also generated, showing two bar charts: mean Dice score versus number of available modalities, and mean HeMIS variance versus number of available modalities. This visualization demonstrates that HeMIS variance reliably signals prediction uncertainty: as the number of available modalities decreases, Dice scores drop and HeMIS variance increases, confirming that the variance is a valid uncertainty indicator.

Modality Encoder (CNN): Each of the four MRI modalities is encoded by an independent CNN encoder. The encoders do not share weights, allowing each one to specialize in the distinct intensity characteristics and contrast mechanisms of its respective modality. Each modality encoder follows a three-scale U-Net-style architecture:

- `enc1`: a ConvBlock operating at full spatial resolution, producing 32-channel feature maps at $128 \times 128 \times 128$.
- `enc2`: a ConvBlock operating at half resolution after MaxPool3D downsampling, producing 64-channel feature maps at $64 \times 64 \times 64$.
- `enc3`: a ConvBlock operating at quarter resolution after a second MaxPool3D, producing 128-channel feature maps at $32 \times 32 \times 32$.

Each ConvBlock consists of two sequential $3 \times 3 \times 3$ 3D convolutions, each followed by Instance Normalization (`InstanceNorm3D`) and Leaky ReLU activation with a negative slope of 0.2. Instance Normalization was chosen over Batch Normalization because it normalizes each sample independently, making it stable and effective even with very small batch sizes (down to batch size 1), which is common in 3D medical image segmentation tasks.

During SSL pretraining, the CNN modality encoders are trained jointly with the SSL encoder. All model parameters, including the CNN encoders, receive gradient updates from the MIM reconstruction loss. This joint training ensures that the CNN encoders learn meaningful MRI feature representations alongside the SSL encoder. When reconstruction quality improves, both the CNN encoders and the SSL encoder become better together, resulting in stronger learned representations that improve downstream finetuning convergence and final segmentation accuracy. This is in contrast to freezing the CNN encoders during pretraining, which would result in random feature maps being fed into the SSL encoder throughout pretraining.

HeMIS Abstraction Layer: The HeMIS abstraction mechanism provides the core missing-modality robustness of the proposed system. At each encoder scale (`enc1`, `enc2`, `enc3`), the HeMIS abstraction layer fuses the feature maps from all available modalities into a single unified representation.
It lets the model handle missing modalities by summarizing the available modality features using their average information and their disagreement.

$$\mu^s = \frac{1}{M} \sum_{m \in \mathcal{M}} f_m^s$$

$$(\sigma^s)^2 = \frac{1}{M} \sum_{m \in \mathcal{M}} (f_m^s - \mu^s)^2$$

$$F_{\text{HeMIS}}^s = \text{Concat}\left(\mu^s, (\sigma^s)^2\right)$$

Here, s denotes the encoder scale or feature level where HeMIS abstraction is applied. The symbol m denotes a modality index, and M denotes the number of available modalities for the current input. The set $\mathcal{M}$, written as calligraphic $\mathcal{M}$ in the equation, represents the available modality set; missing modalities are not included in this set.

The term f_m^s denotes the feature map extracted from modality m at encoder scale s. The term μ^s is the element-wise mean feature map across all available modalities, while $(\sigma^s)^2$ is the element-wise variance feature map across those same modalities. F_{HeMIS}^s denotes the final HeMIS abstraction feature at scale s, and Concat$(\cdot)$ means that the mean and variance feature maps are concatenated along the channel dimension. If only one modality is available, the variance tensor is set to zero with the same shape as the mean tensor.
For a given scale, let $F = \{f_i \mid \text{modality } i \text{ is present}\}$ be the set of available feature maps. The abstraction computes:

- **Mean:** the element-wise mean across all present modality feature maps, representing the average signal across available modalities.
- **Variance:** the element-wise variance across all present modality feature maps, representing the uncertainty or disagreement between modalities.

The mean and variance tensors are concatenated along the channel dimension, doubling the channel count at each scale. This concatenated representation is passed to the segmentation decoder as skip connections.

If there is only one modality, the variance is zero of the same shape as the mean, because the variance is meaningless for one sample. Only two modalities are available for variance computation so the`unbiased=False` is included to achieve numerical stability. The combination of these design decisions results in a numerically stable HeMIS abstraction for all 15 modality combinations.

The HeMIS variance is used as a design element and as an analytical signal for evaluation. The variance is a measure of this disagreement, and is positively correlated with Dice scores empirically (r = 0.51) and higher variance levels are associated

with more disagreement in the available modality feature maps. It is a feature-disagreement indicator, but not a calibrated confidence estimate, since there has been no calibration with uncertainty in the ground truth.

SSL Encoder (Patch Embedding): The SSL encoder starts with a 3D patch embedding module (`PatchEmbed3D`) which divides the input volume into non-overlapping 3D patches and maps each 3D patch into a high-dimensional embedding space. The patch embedding is realized as a 3D convolution with the kernel size of the patch size and stride of the patch size, which is actually equivalent to the non-overlapping extraction of patches and the linear projection at the same time.

The four MRI modalities are concatenated along the channel dimension into a single multi-channel volume of shape $[B, 4, 128, 128, 128]$. This 4-channel volume is treated as a single input by the patch embedding, with the 3D convolution operating across all four modality channels simultaneously within each patch. This is fundamentally different from processing each modality separately — the patch embedding learns joint cross-modal patch representations directly from the concatenated input. With a patch size of $16 \times 16 \times 16$, the $128 \times 128 \times 128$ volume is divided into $8 \times 8 \times 8 = 512$ patches. Each patch is projected to an embedding dimension of 384, producing a token sequence of shape $[B, 512, 384]$.

SSL Encoder (3D Sinusoidal Positional Encoding): To provide the transformer with spatial awareness of patch positions, 3D sinusoidal positional encodings are computed and added to the patch token embeddings. The positional encoding is computed separately for the depth (D), height (H), and width (W) spatial axes, then concatenated to form the full embedding-dimensional positional vector.

For each axis, the encoding uses sinusoidal functions at varying frequencies, following the approach of the original Transformer positional encoding but extended to 3D. The embedding dimension must be divisible by 6 (two sinusoidal components per axis, three axes), which is satisfied by the chosen dimension of 384 ($384/6 = 64$ per axis pair).

Among the design choices is the addition of a position encoding for the visible tokens as well as a position encoding for the mask tokens. This guarantees a representation that is spatially-aware at masked positions to help the decoder to predict the content of the patch at that position instead of a flat prediction for all of the masked patches. Without positional encoding on mask tokens, all the masked positions will have the same inputs for the decoder, introducing a learning limitation.

The positional encodings are stored in a dictionary (`_pos_cache`) which is indexed by the dimensions, the device and the dtype, meaning that they are not recalculated for each individual batch or epoch.

SSL Encoder (Masking Strategy): A random patch masking strategy is used during pretraining. Masking was performed at 0.6 (60% of patches were masked), which was deemed to provide enough challenges for reconstruction, while still maintaining enough context for meaningful learning. Random patch indices are generated

for each sample in a batch and the first 60% of these are set to be masked.

The mask token is a single learnable 384 dimension parameter initialized using a truncated normal distribution with standard deviation 0.02. This parameter is common for all the masked positions with different positional encoding per position. The `torch.where` operation is used to fill in masked positions with mask token embeddings, while leaving the original embeddings unchanged: this avoids in-place embedding operations that are not compatible with PyTorch's Automatic Mixed Precision (AMP) automatic differentiation.

The encoder feeds all the tokens (both visible and mask-replaced tokens) into the transformer, generating contextually enriched representations for all the patch positions, including the masked ones.

SSL Encoder (Transformer): The transformer encoder consists of 4 `TransformerEncoderLayer` blocks with the following configuration:

- Model dimension (d_{model}): 384
- Number of attention heads: 8 (each head has dimension 48)
- Feed-forward dimension: $384 \times 4 = 1,536$
- Dropout rate: 0.1
- Pre-normalization (`norm_first=True`): LayerNorm is applied before the attention and feed-forward sublayers, improving training stability
- LayerNorm applied after the final transformer block

The transformer uses a full sequence of 512 tokens (visible + mask tokens) so that all patches can look at each other via the self-attention mechanism. This global receptive field is an important feature over purely convolutional methods because it allows it to capture long-range spatial dependencies throughout the MRI volume.

SSL Decoder: A light-weight linear projection head exclusively for the purpose of pretraining, called SSL decoder. It has a single fully-connected layer (`nn.Linear`) which maps each token embedding 384-dimension to a patch reconstruction vector $4 \times 16 \times 16 \times 16 = 16,384$ dimensional representing all four modality channels of all voxels within a patch.

Only the positions covered by the masked token are used in the decoder and the reconstruction error is only calculated at these positions. This aligns with the Masked Image Modeling paradigm where the model is trained to predict a masked region from its surrounding context. The SSL decoder is discarded during the pretraining phase and not used during fine-tuning.

SSL Projection and Feature Fusion: After pretraining, the SSL encoder's spatial output is integrated into the segmentation pipeline through a projection and fusion mechanism. The SSL encoder produces a spatial feature map of shape $[B, 384, 8, 8, 8]$ after reshaping the token sequence back to spatial dimensions.

A projection module (`ssl_proj`) consisting of a $1 \times 1 \times 1$ 3D convolution (384 to 384 channels), followed by Instance Normalization and Leaky ReLU activation, is applied to this feature map. The projected SSL features are then trilinearly interpolated to match the spatial dimensions of the HeMIS bottleneck features (`abs3`, which has spatial dimensions of $32 \times 32 \times 32$). The interpolated SSL features are concatenated with `abs3` along the channel dimension, producing a fused bottleneck tensor. The total number of channels at the bottleneck is $2 \times 4 \times 32 + 384 = 640$, where $2 \times 4 \times 32 = 256$ comes from the HeMIS mean+variance concatenation at enc3 scale (128 channels mean + 128 channels variance), and 384 from the SSL projection.

Segmentation Decoder: The segmentation decoder follows a U-Net-style architecture with two upsampling stages and skip connections from the HeMIS abstraction layers at coarser scales.

- **Bottleneck ConvBlock:** reduces the 640-channel fused bottleneck tensor to 128 channels ($4 \times$ base_ch, where base_ch $= 32$).
- **UpBlock 1:** transposed convolution upsampling (stride 2) from 128 channels to 64 channels, followed by concatenation with `abs2` (HeMIS abstraction at enc2 scale, 128 channels: 64 mean + 64 variance), then a ConvBlock reducing to 64 channels.
- **UpBlock 2:** transposed convolution upsampling from 64 channels to 32 channels, followed by concatenation with `abs1` (HeMIS abstraction at enc1 scale, 64 channels: 32 mean + 32 variance), then a ConvBlock reducing to 32 channels.
- **Output head:** a $1 \times 1 \times 1$ 3D convolution projecting 32 channels to 4 output classes, producing the final segmentation logits of shape $[B, 4, 128, 128, 128]$.

Each UpBlock includes a size mismatch correction step: if the transposed convolution output spatial dimensions do not exactly match the skip connection dimensions, trilinear interpolation is applied to align them before concatenation. This prevents shape mismatch errors and ensures correct feature alignment.

Summary of Hyperparameters and Configuration: All configuration mappings, operational variables, and training bounds are organized into Table 4.4.

Parameter	Value
Dataset	BraTS 2021, 1251 cases, 5-fold split
Train / Val split	1000 / 251 (fold 0 as validation and test)
Input modalities	FLAIR, T1ce, T1, T2
Input spatial size	$128 \times 128 \times 128$ voxels
Patch size (SSL)	$16 \times 16 \times 16$
Number of patches	512 $(8 \times 8 \times 8)$
SSL embedding dimension	384
Transformer depth	4 layers
Attention heads	8
Mask ratio (pretraining)	0.75
Missing modality prob (finetune)	0.30
Base CNN channels	32
Bottleneck channels	640 (256 HeMIS + 384 SSL)
Segmentation classes	4 (BG, NCR, ED, ET)
Pretrain optimizer	AdamW, lr=1e-4, weight_decay=0.05
Pretrain scheduler	CosineAnnealingLR
Pretrain batch size	4
Finetune optimizer	AdamW, lr=2e-4, weight_decay=0.05
Finetune scheduler	CosineAnnealingLR
Finetune batch size	1
CE class weights	BG=0.1, NCR=2.0, ED=1.5, ET=2.0
CE / Dice loss weight	0.5 / 0.5
Gradient clipping	max norm = 1.0
Mixed precision (finetune)	AMP float16 with GradScaler
Pretrain precision	float32 (no AMP)

Table 4.4: Proposed HeMIS-SSL Fusion Hybrid Model Hyperparameters and Specifications

4.3 Dataset Details

The dataset is composed of 1,251 complete subjects, each representing one unique glioma case. All subjects include four MRI modalities that capture different tissue contrasts and provides complementary diagnostic information, together with a manually annotated segmentation mask defining tumor subregions. The four types of MRI modality are T1-weighted (T1), T1-weighted contrast-enhanced (T1ce), T2-weighted (T2), and FLAIR (Fluid-Attenuated Inversion Recovery). The T1-weighted modality is for anatomical reference of gray and white matter and for structural analysis as a baseline. After contrast agent injection with gadolinium, the T1ce modality shows areas in which there are disrupted blood–brain barriers, which are associated with active tumor growth and enhancement. T2-weighted images are sensitive to any fluid within the tumor and show cystic and edematous areas around the tumor, and the FLAIR modality provides suppression of the CSF signal which helps to better visualise the peritumoral edema and infiltrative tissue. Combined these modalities give a complete representation of the morphology and

pathological variation of the tumor allowing accurate discrimination of individual tumor components. A representative example of multimodal MRI images and the segmentation overlaid on the images is shown in Figure 4.5, which shows the complementary information provided by each imaging sequence, as well as the spatial distribution of tumor subregions between patients.

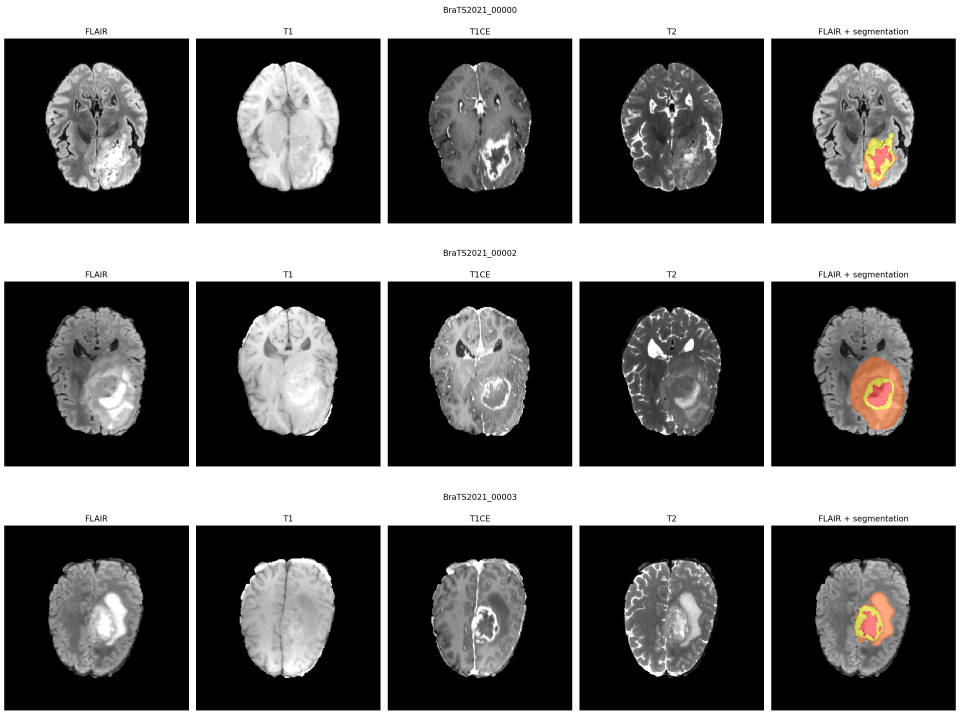


Figure 4.5: Multimodal MRI Examples(3 patients samples)

A segmentation label volume is also provided with voxel-level annotations for each subject in the BraTS 2021 dataset, validated by cross-institutional consensus and annotated by expert neuroradiologists. The four class indices in the label mask are assigned to different regions of the tumor or healthy tissues, labeled as follows: 0 for background and healthy tissues, 1 for necrotic and non-enhancing tumor core (NCR/NET), 2 for peritumoral edema (ED), and 3 for enhancing tumor (ET). These annotations are voxel-wise, allowing for fine-grained quantitative learning and evaluation of segmentation models. In addition these labels can be bundled in clinically Following the official BraTS evaluation protocol for segmentation performance, meaningful composite regions such as Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET) were calculated.

All images and segmentation labels are stored in a widely used standard for the storage of volumetric medical imaging data, the Neuroimaging Informatics Technology Initiative (NIfTI) format. The ability to preserve spatial orientation, voxel spacing, and affine transformation information maintains geometric consistency across modalities with the use of NIfTI. The dimension of each volume acquired by MRI is about $240 \times 240 \times 155$ voxels which is close to 1 mm^3. All the dataset providers have performed full skull stripping and co-registration of the data-set, which means that all modalities of a given subject are anatomically aligned, and only contain brain tissue relevant to the segmentation analysis.

The whole dataset is arranged following the standard BraTS directory structure. There are 5 NIfTI files for each patient file, one for each MRI modality and a segmentation mask. The naming convention uniquely identifies each subject and imaging sequence, allowing automated loading and pre-processing of the data in deep learning frameworks like MONAI and PyTorch. This organisation guarantees a perfect separation between input images and target labels and the consistency of all the subjects.

For the sake of reproducible experiments and robust evaluation, all 1251 subjects were split into five predefined folds. The cases in Fold 0 are used only for the validation and final test, and the cases in Folds 1-4 are used for the model training with 250, 250, 250, and 251 cases, respectively. This approach to partitioning guarantees that the training data and the evaluation data are completely separated, To prevent information leakage and give an unbiased assessment of model generalization performance. All subjects had full sets of the four MRI modalities, providing multimodal input across all experiments.

There is significant variation in tumor burden throughout the cohort, which is indicative of the diversity of glioma progression. Figure 4.6 shows the distribution of all 1,251 segmentation masks' total tumor volume. The mean tumor volume measured 96.0 cm^3 and the median tumor volume was 89.3 cm^3, suggesting a moderate right skew because of a few larger tumors. In most cases, the tumor volume is within the range of around 30 to 150 cm^3, although larger volumes (more than 300 cm^3) are also seen. This large discrepancy in tumor burden renders the segmentation task challenging and gives rise to a significant anatomical diversity that should be addressed by models that are able to recognize tumors across a wide range of sizes.

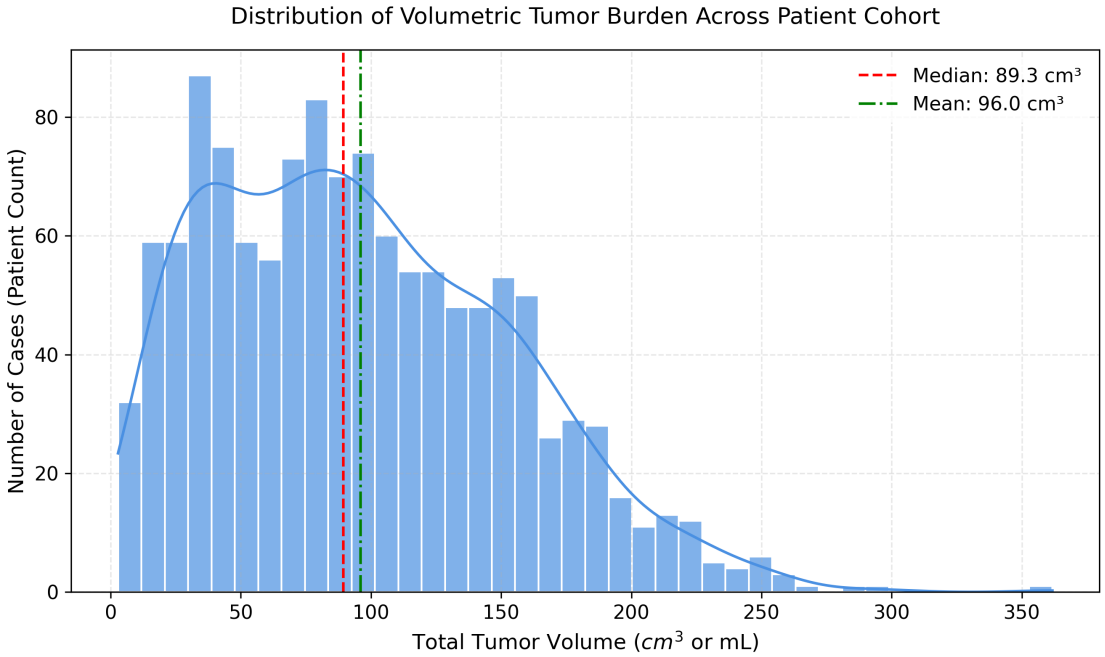


Figure 4.6: Tumor Volume Distribution

To further define tumor burden, each subject was classified into one of clinically interpretable volumetric categories according to the total tumor volume. As illustrated in Figure 4.7, 321 patients (25.7%) were categorized as Small (< 50 cm^3), 393 patients (31.4%) as Medium (50–100 cm^3), 309 patients (24.7%) as Large (100–150 cm^3), and 228 patients (18.2%) as Very Large (≥ 150 cm^3). The fairly even distribution throughout these volumetric tiers illustrates that there is a wide range of presentations of tumors in this data with not one size dominating. This diversity is especially useful for training segmentation models that should be robust to both the early and advanced disease stages.

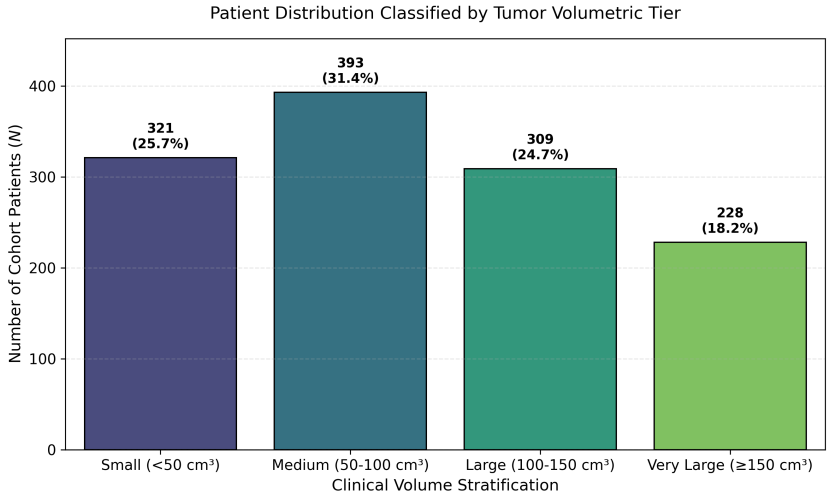


Figure 4.7: Tumor Size Categories

In addition to anatomical variations, there are also significant differences in modality specific intensity characteristics. The distribution of the voxel intensities for the four MRI sequences is shown in Figure 4.8. FLAIR, T1, T1ce, and T2 intensities exhibit distinct profiles which are related to the acquisition mechanisms and the tissue contrast properties of the modalities. Compared to T1 and FLAIR, which offer complementary structural information, T1ce and T2 modalities typically exhibit greater intensity ranges, which may be attributable to differences in sensitivity to contrast enhancement and fluid accumulation, respectively and pathological information. These modality-specific intensity distributions emphasize the value of the multimodal approach to learning because each sequence provides information that could enhance tumor localizations and boundary delineations.

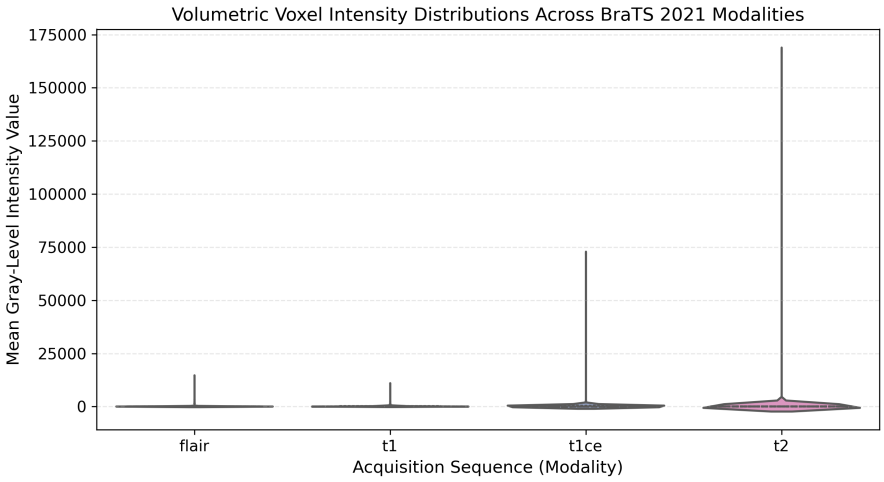


Figure 4.8: Modality Intensity Distribution

A crucial property of the BraTS 2021 dataset is the class imbalance that is present in voxel-wise segmentation tasks. The large proportion of voxels are classified in the background class, whereas the volume of voxels corresponding to structures associated with tumors is small compared to the imaging volume (as illustrated in Figure 4.9). Peritumoral edema is the largest tumor subregion, while necrotic core and enhancing tumor regions have significantly less voxels. This imbalance poses a big problem for deep learning models as optimization might be skewed towards the predominant background class. Therefore, the segmentation architecture and loss functions need to be well-designed to satisfy the requirement of highlighting minority tumor regions, for accurate delineation of clinically important structures.

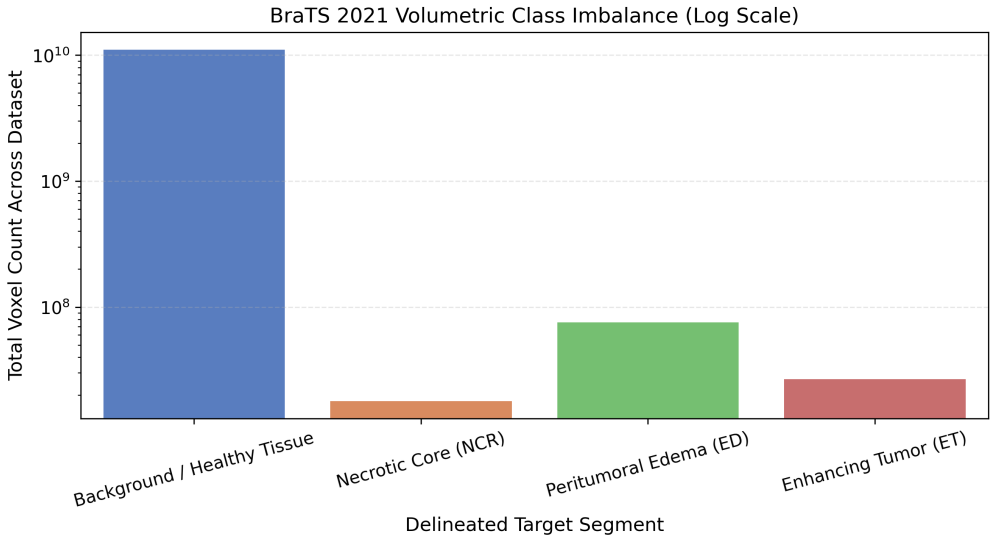


Figure 4.9: Tumor Class

The multi-institutional and multi-scanner diversity is one of the hallmarks of the BraTS 2021 data set. The MR images were obtained in various hospitals, vendors, and imaging sequences across the world with significant differences in the image contrast, noise properties, and acquisition conditions. Such a diversity makes the

dataset more clinically realistic and motivates modelling with the goal of learning robust and generalizable feature representations. This is especially useful for self-supervised and representation-learning methods since it encourages representation learning to be modality-invariant and anatomically meaningful.

In general, the BraTS 2021 dataset is a complete basis for this study. It is one of the most used multimodal benchmark datasets for brain tumor segmentation research due to its large cohort size, multimodal design, expert voxel-level annotations, volumetric diversity and heterogeneous acquisition sources. Standardized pre-processing, clinically relevant annotations, and significant interpatient variability make the models being evaluated under realistic conditions and the results can be meaningfully translated towards real-world imaging applications.

4.4 Pre-Processing

This section outlines the data preparation workflow implemented to standardize the volumetric inputs and targets before model training. The pipeline encompasses data validation, spatial transformation, and intensity normalization strategies.

4.4.1 Inputs and Data Convention

The BraTS 2021 dataset contains 1251 complete subjects that correspond to 1251 unique cases of glioma. For each subject, there are four MRI modalities (T1, T1-weighted contrast (T1ce), T2, and FLAIR (Fluid-Attenuated Inversion Recovery), along with a manually annotated segmentation mask. All images and labels are saved in the Neuroimaging Informatics Technology Initiative (NIfTI) format, which maintains the spatial orientation and voxel metadata. For each patient folder, there are five NIfTI files corresponding to the four MRI modalities and one segmentation label volume.

4.4.2 Data Validation and Integrity Screening

All subjects had full sets of four modalities for MRI data, resulting in consistent multimodal inputs for all experiments. The dataset providers provided co-registered and skull-stripped MRI volumes, so that the anatomical information was presented in the same way across the modalities for each individual and the non-brain information was removed before analysis. The data were also fivefold divided to ensure a strict separation of training and evaluation data, thus avoiding information leakage.

4.4.3 Spatial Standardization

The BraTS 2021 dataset consists of approximately $240 \times 240 \times 155$ voxels and a spatial resolution of nearly 1 mm^3 per voxel. A central crop is applied to all volumes to ensure that the inputs have the same dimensions and to help save memory. All images are resized to a fixed size of $128 \times 128 \times 128$ voxels, centered around the 3D spatial axes.

Cropping begins at position `"zero",` if the original volume is smaller than 128 voxels along any dimension. This will keep the dimensions of the tensors uniform without causing interpolation artifacts, due to resizing or resampling.

The BraTS 2021 dataset provides MRI volumes with dimensions of approximately $240 \times 240 \times 155$ voxels and spatial resolution close to 1 mm^3. To standardize input dimensions and reduce memory requirements, a central crop operation is applied to all volumes. Each image is cropped to a fixed size of $128 \times 128 \times 128$ voxels, centered along all three spatial axes.

If the original volume is smaller than 128 voxels along any dimension, cropping begins at position zero. This approach maintains consistent tensor dimensions without introducing interpolation artifacts through resizing or resampling.

4.4.4 Label Normalization

BraTS segmentation labels are remapped to a unified four-class representation. Label 4, which corresponds to Enhancing Tumor (ET) in some BraTS versions, is converted to label 3. The resulting label space consists of:

- **0:** Background
- **1:** Necrotic and Non-Enhancing Tumor Core (NCR/NET)
- **2:** Peritumoral Edema (ED)
- **3:** Enhancing Tumor (ET)

This remapping ensures consistent label definitions throughout the dataset and simplifies the segmentation objective.

4.4.5 Foreground Masking

Only non-zero voxels are intensity normalized. A foreground brain mask is implicitly defined by choosing voxels that contain non-zero intensity values, excluding background and skull regions that contain no relevant information for diagnosis. The statistics computed for normalizations are done only within this foreground mask.

4.4.6 Intensity Normalization (Per-Volume, Per-Channel Z-Score)

Each MRI modality is separately intensity normalized. First percentile clipping at 0.05th and 99.95th percentile is performed within the foreground mask for each volume to remove the extreme intensity outliers. After clipping, the z-score normalization method is used to subtract off the mean foreground intensity and divide by the standard deviation of the same foreground. To avoid numerical instability, a small constant $\epsilon = 1 \times 10^{-8}$ is added to the denominator.

The formula removes extreme intensity values and then converts each MRI modality into a normalized intensity distribution suitable for model training.

$$p_{\text{low}} = P_{0.05}(X_m(\Omega)), \quad p_{\text{high}} = P_{99.95}(X_m(\Omega))$$

$$X_m^{\text{clip}}(i) = \text{clip}(X_m(i), p_{\text{low}}, p_{\text{high}})$$

$$X_m^{\text{norm}}(i) = \frac{X_m^{\text{clip}}(i) - \mu_m}{\sigma_m + 10^{-8}}$$

Here, X_m denotes the input MRI volume of modality m, where m may represent FLAIR, T1ce, T1, or T2. The symbol Ω denotes the set of non-zero brain voxels, meaning voxels where $X_m(i) > 0$, and i denotes a voxel location inside the 3D MRI volume. $P_{0.05}$ and $P_{99.95}$ represent the 0.05th and 99.95th percentile intensity values computed only from the non-zero brain voxels. The terms p_{low} and p_{high} are the lower and upper clipping thresholds, respectively. $X_m^{\text{clip}}(i)$ is the clipped intensity value at voxel i, and $X_m^{\text{norm}}(i)$ is the final normalized intensity value. The terms μ_m and σ_m represent the mean and standard deviation of the clipped non-zero voxels for modality m. The constant 10^{-8} (ϵ) is added for numerical stability in the pre-processing code.

4.4.7 Data Augmentation

Data augmentation is applied during training only, not during validation or evaluation. The augmentation pipeline is applied after loading from the cache, ensuring that each epoch sees a different augmented version of the data. The following augmentation strategies are employed:

- **Random Spatial Flips:** Flips along all three spatial axes (depth, height, width) with a probability of 0.5 per axis. Flipping is applied consistently across all modalities and the segmentation mask to preserve spatial correspondence.
- **Random Brightness Jitter:** A uniform random offset sampled from the range $[-0.1, 0.1]$ is independently added to each modality volume.
- **Random Contrast Jitter:** Each modality volume is independently multiplied by a uniform random factor sampled from the range $[0.9, 1.1]$.

These augmentations increase the effective diversity of the training set and improve the model's robustness to variations in MRI acquisition parameters and scanner differences across clinical sites.

4.4.8 Divisibility and Padding Policy

The pre-processing pipeline adds constraints on spatial size, such as center cropping to $128 \times 128 \times 128$ voxels. When the original volume is less than 128 voxels in any axis, a crop operation will begin at position zero. No further policy was given in regard to padding.

4.4.9 Tensor Packaging and Label Representation

The volumes of the images are pre-processed and cropped and translated into PyTorch tensors. The tensors are processed and written on disk as .pt files, with a caching mechanism, the initial time the dataset is processed. These tensor files can in turn be loaded directly during the next epochs of training thereby saving the unnecessary NIfTI loading and pre-processing. The segmentation scheme applied is a four class segmentation scheme which is normalized.

4.4.10 Reproducibility and Data Consistency

The data set is divided into five folds that are intentionally defined to enable repeatable experimentation. Cross validation and final testing will be conducted on Fold 0, and training will be performed on Folds 1–4. This is a fixed partitioning strategy which has the benefit of uniform evaluation throughout experiments and eliminates information leakage between training and testing sets. The pre-processing is also standardized and the cropping size is fixed which further contributes to the uniformity of the data.

4.4.11 Numerical and Performance Safeguards

The following precautions are taken in the pre-processing pipeline. Using percentile clipping at 0.05th and 99.95th percentile decreases the impact of outliers with extreme intensities. In z-score normalization, the standard deviation is increased by a small value of 1×10^{-8} to avoid having the standard deviation value equal to zero, which can cause instability in the calculation. Additionally, a disk-based caching system is used to store preprocessed tensors in .pt format, resulting in a substantial decrease of CPU overhead and data loading time during training.

4.4.12 Common Pitfalls Avoided

The pre-processing pipeline avoids several common issues in medical image analysis:

- Inconsistent label definitions are prevented through label remapping.
- Extreme intensity outliers are mitigated through percentile clipping.
- Background voxels are excluded from normalization statistics through foreground masking.
- Information leakage is avoided through predefined train-test fold separation.
- Interpolation artifacts are avoided by using center cropping instead of resizing or resampling.
- Repeated pre-processing overhead is minimized through tensor caching.

4.4.13 Outcome of the pre-processing Pipeline

The pre-processing pipeline generates standardized multimodal MRI volumes that have the same spatial dimensions in voxels ($128 \times 128 \times 128$), normalized intensity distributions, and a common segmentation label. The combination of foreground masking, intensity normalization, label remapping, and tensor caching ensure that all 1,251 patients are consistent and generate high-quality, computation-efficient inputs for training deep learning models.

Chapter 5

Result Analysis

5.1 Performance Evaluation

The fusion model proposed in the paper was tested under various MRI input conditions and evaluated regarding its segmentation accuracy, missing-modality robustness and reliability. The evaluation was not limited to the full four modalities input, since the main goal of this research is to develop a model that can work even if some of the MRI modalities are unavailable. Rather, the trained model was evaluated on all 15 possible non-empty combinations of the four MRI modalities (FLAIR, T1ce, T1 and T2).

The validation/ test split of the BraTS 2021 was used for the evaluation. Dice-based segmentation metrics were used to evaluate the model output. The main metric used for evaluation was scored according to the dice scores, due to the highly imbalanced nature of brain tumor segmentation; tumor regions account for a much smaller proportion of volume than background class. This means that even models that are extremely accuracy at the voxel level can be misleading, if they correctly predict background for the majority of voxels. The advantage of this is that it directly measures the overlap between the predicted tumor region and the ground-truth tumor mask, making it more suitable than the dice score.

Two sets of Dice scores were used for the evaluation of the model. The first group are the label-wise Dice scores of the three tumor classes: necrotic/non-enhancing tumor core (NCR/NET), peritumoral edema (ED), and enhancing tumor (ET). Label means Dice is the average of these three scores. The second group is the Enhancing Tumor (ET), Tumor Core (TC), and Whole Tumor (WT) scores from the BraTS-style clinical region Dice. These three clinical region scores are used to compute the clinical mean Dice. This dual assessment offers class and clinically relevant evaluations of performance.

Combination	**SimCLR (3D)**	**Med3D**	**SSL+HeMIS**
F	0.235	0.2017	0.418
T1c	0.298	0.0718	0.532
T1	0.001	0.0121	0.214
T2	0.060	0.1102	0.424
F + T1c	0.622	0.4612	0.745
F + T1	0.306	0.2689	0.486
F + T2	0.288	0.2845	0.539
T1c + T1	0.376	0.2437	0.654
T1c + T2	0.421	0.4180	0.719
T1 + T2	0.013	0.1276	0.468
F + T1c + T1	0.706	0.6468	0.754
F + T1c + T2	0.681	0.5582	0.763
F + T1 + T2	0.334	0.3215	0.551
T1c + T1 + T2	0.450	0.5918	0.732
All 4	0.724	0.7044	0.766

Table 5.1: Comparison of SimCLR, Med3D, and SSL+HeMIS across 15 modality combinations

Both the greatest increase in performance and the highest performance overall was obtained when all four modalities were available. The model performed well in this setting with a label mean Dice score of 0.766 and a clinical mean Dice score of 0.851. The clinical region scores were particularly high, with ET Dice 0.791, TC Dice 0.859 and WT Dice 0.902. This is a validation that the multimodal input will give the most reliable segmentation output as each MRI sequence will provide complementary information about the tumor.

But the results also indicate that the model is also usable if some modalities are not present. The mean Dice scores for the combined FLAIR + T1ce + T2 were 0.763 and 0.847 for label and clinical, respectively, and very close to the all-four-modality result. This means that the model is not fully reliant on all four modalities and can yield good segmentation performance without the availability of any one modality. An important result since not all real clinical MRI data may include all sequences.

T1ce outperformed the other single-modality methods with a label mean Dice score of 0.532 and clinical mean Dice score of 0.611. This is the expected result, as T1ce is known to accentuate the enhancing parts of the tumor. FLAIR and T2 were also reasonably good single modalities as they give good visibility of edema and abnormal tissue regions. The weakest result was achieved by T1 alone with a label mean Dice score of 0.214. Thus, it appears that T1 offers valuable anatomical information, but this alone is insufficient for accurate tumor subregion segmentation.

Number of Available Modalities	Average Label Mean Dice	Average Clinical Mean Dice
1 modality	0.397	0.517
2 modalities	0.602	0.719
3 modalities	0.700	0.798
4 modalities	0.766	0.851

Table 5.2: Average model performance by number of available modalities

The results show a clear improvement as more modalities become available. The average label mean Dice increased from 0.397 with a single modality to 0.602 with two modalities, 0.700 with three modalities, and 0.766 with all four modalities. A similar pattern is observed for clinical mean Dice, which increased from 0.517 with one modality to 0.851 with all four modalities. This trend demonstrates the importance of multimodal information in brain tumor segmentation.

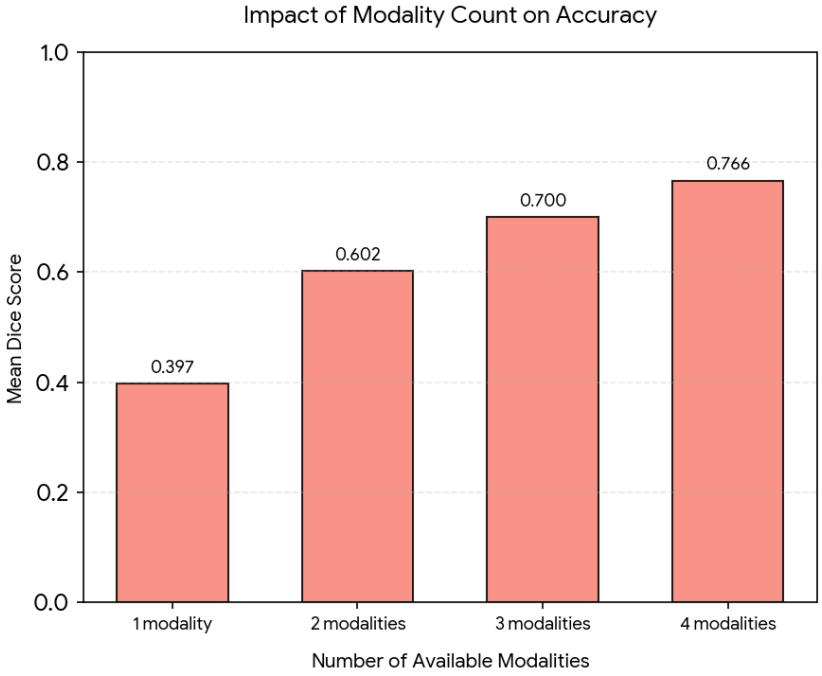


Figure 5.1: Impact of Modality Count on Mean Dice Score

Figure 5.1 shows that segmentation performance improves as the number of available MRI modalities increases.

To assess the balance of correct tumor detection and false prediction, the Dice score, precision, recall and F1-score were also calculated. In the all four modalities setting, the model's performance results in macro precision, macro recall, and macro F1-score of 0.824, 0.841, and 0.831, respectively. The class-wise F1-scores were 0.784 for NCR, 0.847 for ED, and 0.863 for ET. The results indicate that the model performs consistently across the tumor classes, though NCR is more challenging than ED and ET. This makes sense because NCR regions tend to be more complex in terms of size and shape and may not be easily identified in the vicinity of healthy

or abnormal tumor tissue.

In addition to quantitative segmentation outputs, qualitative segmentation outputs were produced to visually check the model predictions. These figures show the input MRI slice, ground truth segmentation mask, and predicted segmentation mask. The qualitative results confirm the quantitative results, predictions based on all 4 modalities have better agreement with ground truth, predictions based on single modalities have more missing or incomplete tumor regions. The segmentation quality also changes dynamically depending on which modalities are available, as shown by the all-combination prediction figures.

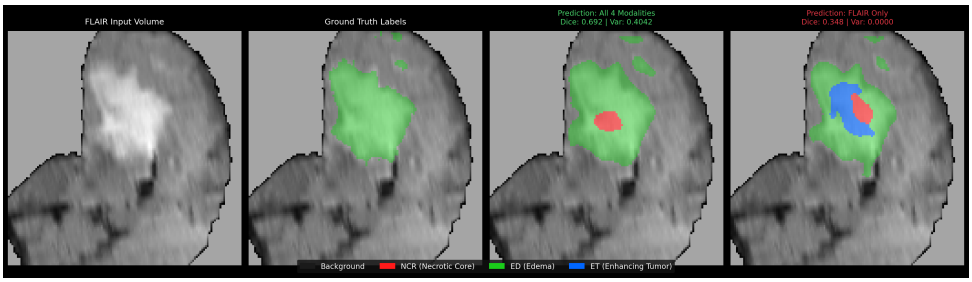


Figure 5.2: Representative Segmentation Output Using All Four Modalities and FLAIR-only Input

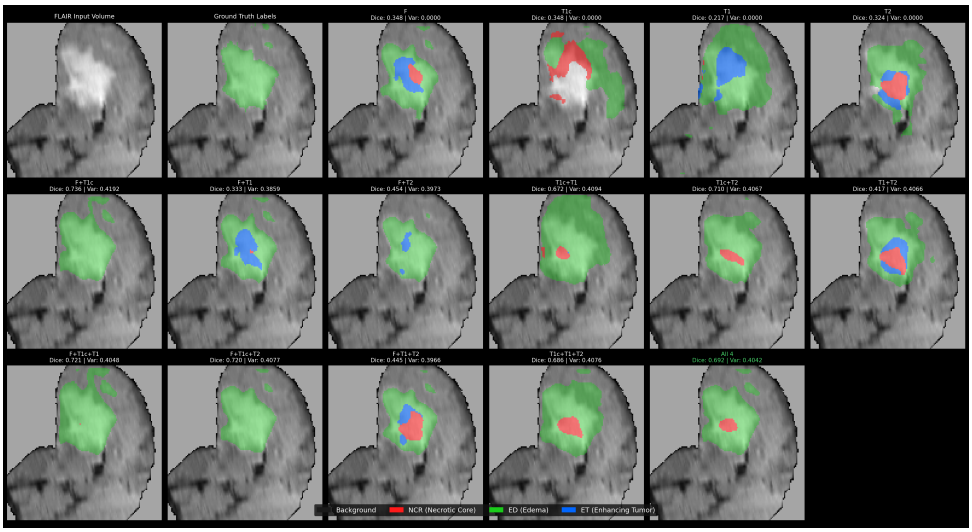


Figure 5.3: Prediction Outputs Across All 15 Modality Combinations for a Representative Case

Overall, the results of the performance evaluation demonstrate that the main goal of the proposed SSL+HeMIS model is fulfilled, namely good segmentation performance in the presence of all modalities and acceptable performance in the case of incomplete modalities. The model is expected to be best when inputted with all modalities, but the performance difference between all four modalities and the combination of FLAIR, T1ce, and T2 is small, suggesting good robustness in the absence of one modality.

The results further validate the effect of modality identity, rather than modality number. Combinations with T1ce usually lead to better performance for ET and TC segmentation, whereas the FLAIR and T2 are good for whole tumor and edema-related segmentation. Hence, the proposed model is appropriate for the target application of this research, that is, the missing-modality setting.

5.2 Analysis of Design Solutions

The main goal of this research was not only to obtain high segmentation performance under complete MRI input, but also to develop a model that remains usable when one or more MRI modalities are missing. Therefore, the design solutions were analyzed using four main criteria: segmentation performance, missing-modality support, architectural feasibility, and suitability for practical medical imaging conditions.

Three design solutions were considered in the final analysis: SimCLR, Med3D, and SSL+HeMIS. SimCLR was included as a self-supervised learning baseline because it can learn useful representations from unlabeled medical images. Med3D was included as a 3D medical imaging baseline because it is designed to process volumetric medical data. SSL+HeMIS was considered as the proposed final design because it combines self-supervised representation learning with modality-aware feature fusion. The design-level comparison is summarized in Table 5.3.

Design Solution	**Main Strength**	**Main Limitation**	**Final Decision**
SimCLR	Learns useful SSL-based visual representations from 3D MRI data	Does not include a direct mechanism for handling missing MRI modalities	Used as baseline
Med3D	Processes volumetric medical images and learns 3D spatial features	Depends on stable complete-input conditions and lacks modality-aware fusion	Used as baseline
SSL+HeMIS	Combines SSL feature learning with HeMIS-style missing-modality fusion	More architecturally complex than the baselines	Selected as final model

Table 5.3: Design-level comparison of considered solutions

SimCLR is useful because self-supervised learning can reduce dependence on labeled data. However, its contrastive learning design does not directly solve the missing-modality problem. It can learn strong image-level or feature-level representations, but it does not explicitly fuse only the available MRI modalities during inference. For this reason, it is suitable as a representation-learning baseline but not as the final solution for the missing-modality objective.

Med3D is useful because brain tumor segmentation is a 3D volumetric task, and Med3D-style models are better suited than 2D models for learning spatial rela-

tionships across MRI slices. However, the Med3D baseline also lacks an explicit missing-modality fusion strategy. It is therefore more appropriate as a supervised 3D medical imaging baseline than a missing-modality solution.

The proposed SSL+HeMIS model better matches the problem requirements. Its SSL component supports representation learning from MRI data before supervised segmentation training, while the HeMIS-inspired fusion component allows the model to combine information dynamically from the available modalities. This makes the design more suitable for incomplete MRI input conditions. Instead of requiring a separate model for every missing-modality case, the same trained framework can evaluate all non-empty modality combinations.

Model	**Mean Dice Across 15 Combinations**	**Dice with All 4 Modalities**	**Selection Status**
SimCLR	0.368	0.724	Baseline only
Med3D	0.335	0.704	Baseline only
SSL+HeMIS	0.584	0.766	Selected

Table 5.4: Aggregate performance comparison of final design candidates

The results of the aggregated modality combination indicate that the performance of the SSL+HeMIS was the best mean Dice score in the 15 modality combinations. It was also best in the presence of all four modalities. This means that the proposed design is not only more robust when dealing with missing-modality but also when dealing with the complete-input.

The choice of SSL+HeMIS on the basis of the suitability of its design and experimental performance is thus justified. Although both SimCLR and Med3D are helpful baselines, these do not directly tackle the central task of the project of strong segmentation in the face of incomplete MRI input. The choice of SSL+HeMIS is motivated by its ability to deliver a strong segmentation performance, while being flexible in terms of modality and being practical in the context of medical imaging applications.

5.3 Final Design Adjustments

Once the model was tested, both in the case of complete and missing modalities, several final design decisions were confirmed to enhance the reliability and clarity of the proposed system. Such changes were done based upon the observed performances, missing modality behavior, and the practical goal of having a robust brain tumor segmentation model for incomplete MRI inputs.

The first big change was to retain SSL+HeMIS as the final selected architecture. This model was then evaluated in a performance test, and was found to yield the most consistent results when evaluated under various modality combinations. Most importantly, it wasn't only the all four modality settings. The model worked well in the absence of one, two, or three modalities. Hence, the final design was decided around the self-supervised representation learning of MRI with modality fusion in

the line of HeMIS.

The second change was to make sure that the missing-modality evaluation method was part of the final design. Owing not merely to the complexity of the task, but also to the need to evaluate the final performance based on the 15 possible combinations of non-empty modalities, the result of the entire input was not reported but rather the final evaluation was organized around all these 15 combinations. This adjustment was important, as the main problem of the research is not just segmentation accuracy in ideal situations, but also robustness in the case of missing MRI sequences. Therefore, the final model is assessed in a simulated missing-modality robustness evaluation.

The third adjustment was to the input handling during inference. If a modality is not available, a zero-filled placeholder is used for the SSL input branch when calculating fusion statistics; the HeMIS abstraction layer only computes fusion statistics from the available modality features. This eliminates the need for training models based on different combinations of modalities. It can also be used to handle varying input availability from a single trained model without making changes to the model architecture.

The fourth adjustment was the use of Dice-based evaluation as the main performance criterion. The main reason for not using accuracy as the primary metric was that the brain MRI is a 3D volume of voxels from the background. Models can have high accuracy by predicting the majority of backgrounds. Accordingly, the last evaluation is aimed at Dice scores on the label level as well as BraTS-like clinical region Dice scores for ET, TC and WT. This is a better way to evaluate the segmentation results of tumors.

The fifth adjustment was the incorporation of qualitative visual examination with quantitative examination. All predictions from each modality as well as all predictions from the 15 modality combinations were presented in the segmentation output to complement the numerical results. These visual outputs illustrate the impact of removing important modalities on the quality of segmentation. In addition, they also make the evaluation more interpretable since the reader has the ability to see the real differences in tumor masks instead of just relying on numerical scores.

The sixth adjustment relates to the meaning of "variance" in HeMIS. In this work, HeMIS variance is solely used as a feature-disagreement signal: it quantifies the deviation of the feature maps across the modalities and empirically shows to be predictive of the segmentation performance when less informative modalities are provided. It's not a calibrated uncertainty score and cannot be used directly as a substitute for clinical confidence without reliability analysis, such as calibration curves or expected calibration error evaluation. The primary performance indicator continues to be the dice score, with variance a secondary analytical indicator.

Design Aspect	Final Adjustment	Reason
Final model choice	SSL+HeMIS retained as the selected model	It showed the strongest overall robustness across complete and incomplete modality settings
Evaluation protocol	All 15 non-empty modality combinations retained	Matches the research objective of missing-modality segmentation
Missing input handling	Zero-filled SSL input with HeMIS fusion over available modalities	Allows one model to handle different modality combinations
Main metric	Dice-based evaluation prioritized	More suitable than accuracy for imbalanced tumor segmentation
Qualitative analysis	Segmentation figures included	Supports numerical findings with visual evidence
Variance interpretation	Treated as feature-disagreement signal	Avoids overclaiming clinical uncertainty without calibration

Table 5.5: Summary of final design adjustments

Overall, these final adjustments make the proposed system more consistent with the research objective. The final design is not optimized only for the ideal case where all MRI modalities are present. Instead, it is designed and evaluated as a modality-flexible segmentation framework. This makes the final model more appropriate for the practical problem such as: brain tumor segmentation under incomplete MRI modality conditions.

5.4 Statistical Analysis

The results are analyzed to gain insights into the behavior of the proposed SSL+HeMIS model under varying modality conditions in four ways: descriptive statistics by number of modalities, comparison with baseline models, correlation between evaluation variables, and class-wise predictive behavior.

Results were first tabulated on the basis of the number of MRI modalities available and descriptive statistics were computed for the resulting groups. It is helpful to determine if the model performance gets better with the increasing number of modalities. Table 5.6 presents the summary of the mean and standard deviations of label mean Dice, clinical mean Dice, and macro F1-score.

No. of Available Modalities	No. of Combinations	Label Mean Dice	Clinical Mean Dice	Macro F1-score
1 modality	4	0.397 ± 0.133	0.516 ± 0.130	0.464 ± 0.133
2 modalities	6	0.602 ± 0.120	0.719 ± 0.090	0.669 ± 0.119
3 modalities	4	0.700 ± 0.101	0.798 ± 0.075	0.768 ± 0.103
4 modalities	1	0.766	0.851	0.831

Table 5.6: Statistical summary by number of available modalities

The descriptive results show a clear upward trend as the number of available modalities increases. The average label mean Dice improves from 0.397 for single-modality input to 0.766 for all four modalities. The same pattern is observed for clinical mean Dice and macro F1-score. This confirms that multimodal information improves segmentation performance.

However, the standard deviation values also show that performance is not determined only by the number of modalities. The specific identity of the available modality is also important. For example, combinations containing T1ce generally perform better than combinations without T1ce because T1ce provides strong contrast information for enhancing tumor and tumor core regions.

Second, a configuration-level paired comparison was performed between SSL+HeMIS and the two baseline models, SimCLR and Med3D. The comparison was based on the 15 modality-combination Dice scores. Since these are aggregate modality-combination scores rather than independent patient-level samples, the statistical result should be interpreted as configuration-level evidence rather than a clinical significance test.

Model	Mean Dice Across 15 Combinations	Standard Deviation	Mean Difference from SSL+HeMIS
SimCLR	0.368	0.239	+0.217
Med3D	0.335	0.219	+0.250
SSL+HeMIS	0.584	0.166	—

Table 5.7: Configuration-level comparison between SSL+HeMIS and baseline models

The paired comparison shows that SSL+HeMIS achieved a higher Dice score than both baselines across the modality-combination settings. The average improvement was 0.217 over SimCLR and 0.250 over Med3D. A paired t-test across the 15 combination-level scores produced $p < 0.001$ for SSL+HeMIS compared with both SimCLR and Med3D. A Wilcoxon signed-rank test also showed $p < 0.001$.

These results support the observation that the proposed model is consistently stronger across different modality availability conditions. However, because the test is performed on modality-combination averages, this should be presented as supporting statistical evidence, not as a replacement for patient-level external validation.

Third, correlation analysis was performed to examine relationships among modality count, Dice score, and HeMIS variance. Pearson correlation was used for this analysis. Moreover, The correlation analysis showed that label mean Dice and clinical mean Dice were strongly associated, with a Pearson correlation of $r = 0.92$. The number of available modalities showed a weak-to-moderate positive relationship with label mean Dice ($r = 0.28$) and clinical mean Dice ($r = 0.25$). This indicates that performance generally improves with additional modalities, but modality identity also plays an important role. The number of modalities was strongly correlated with global HeMIS variance ($r = 0.96$), which is expected because variance is computed across available modality features. ROI-level variance showed a moderate positive relationship with label mean Dice and clinical mean Dice, both approximately $r = 0.51$. Therefore, HeMIS variance should be interpreted as a feature-disagreement signal rather than a directly calibrated uncertainty score.

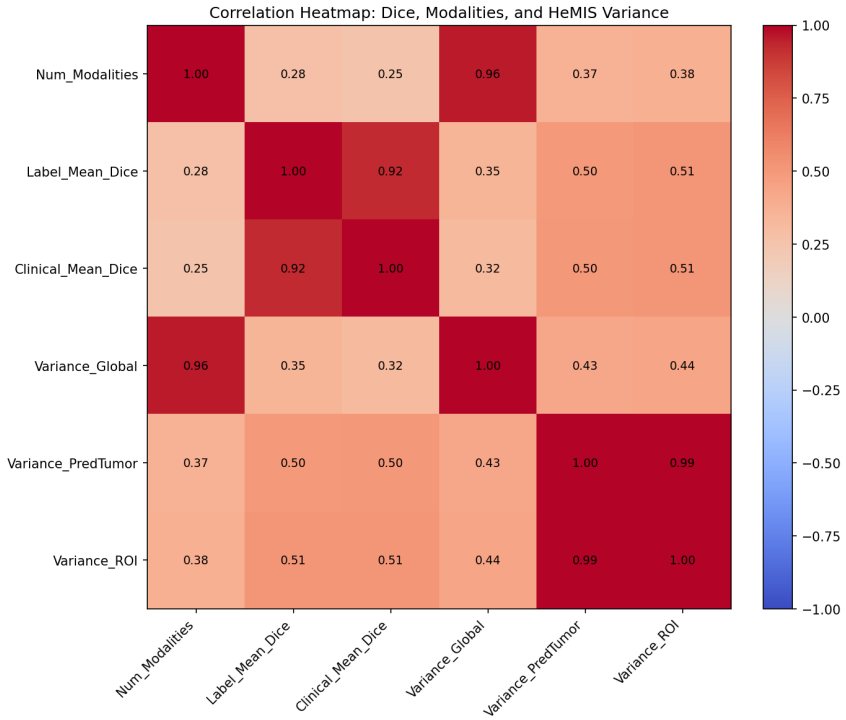


Figure 5.4: Correlation Heatmap between Modality Count, Dice Scores, and HeMIS Variance

Fourth, class-wise classification metrics were analyzed using precision, recall, and F1-score. For the all-four-modality setting, the model achieved a macro precision of 0.824, macro recall of 0.841, and macro F1-score of 0.831. Among the three tumor classes, ET achieved the highest F1-score of 0.863, followed by ED with 0.847 and NCR with 0.784. This indicates that the model was strongest in detecting enhancing tumor and edema regions, while NCR remained the most difficult tumor class.

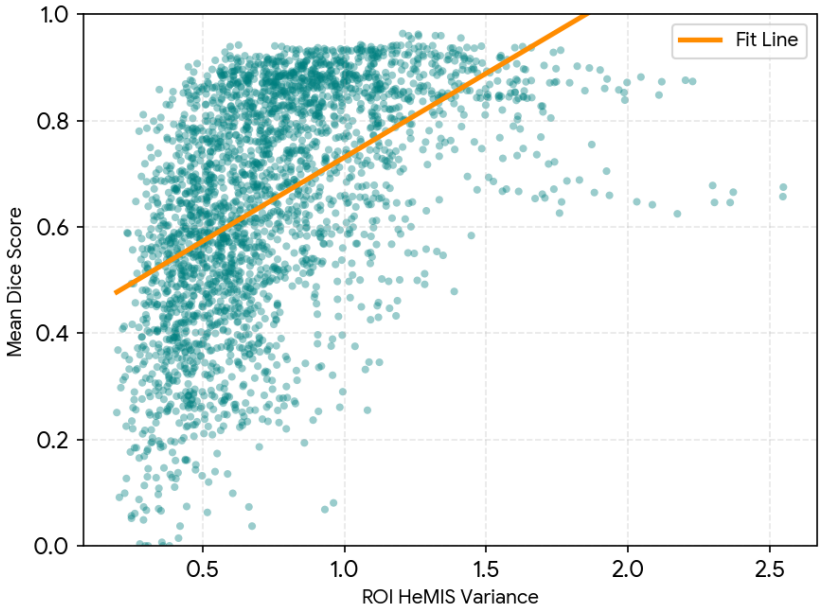


Figure 5.5: Per-Case ROI Variance vs Dice (r=0.513)

From the raw tumor-only confusion matrix, row-wise percentages were calculated to interpret class-level behavior. The confusion matrix showed that 95.32% of ED voxels and 89.89% of ET voxels were correctly classified. NCR had a lower correct classification rate of 77.23%, with some NCR voxels being confused with ED and ET. This suggests that NCR is the most challenging subregion because it has less consistent visual contrast and may be more difficult to separate from neighboring tumor tissues.

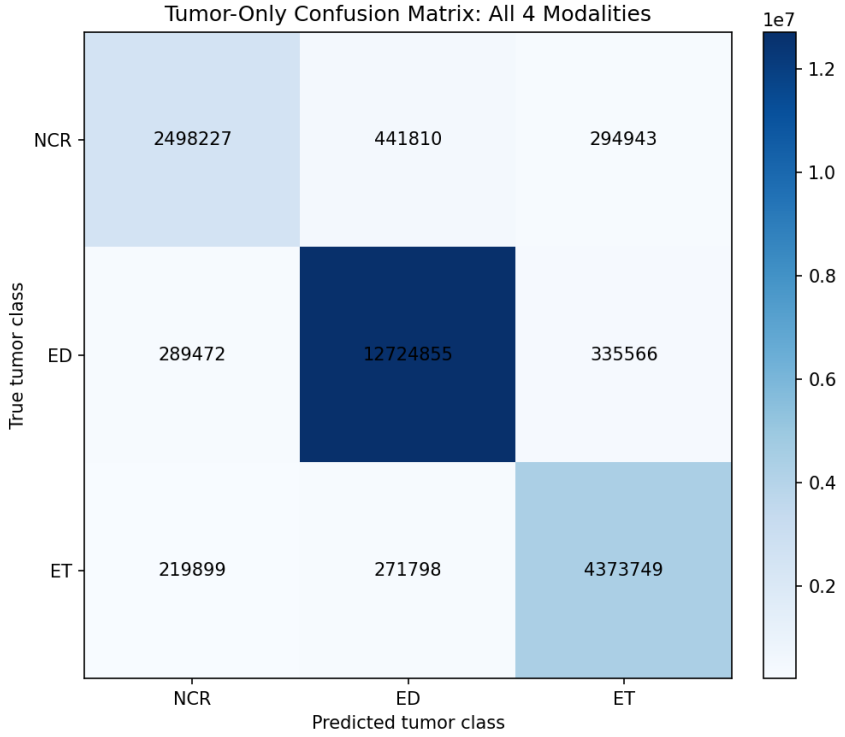


Figure 5.6: Tumor-only Confusion Matrix for All-Four-Modality Prediction

Overall the statistical analysis verifies three important findings. First, the more MRI modalities that are available the better the segmentation performance. Second, the average Dice performance of SSL+HeMIS surpasses that of SimCLR and Med3D in the 15 different modality-combination settings. Thirdly, class-wise analysis reveals that the segmentation of ED and ET is more reliable than NCR. These statistical results support the final design SSL+HeMIS to be suitable for accurate segmentation of brain tumors in incomplete MRI modality.

5.5 Comparisons and Relationships

This section contrasts the behaviour observed for different model designs and modality combinations. The previous sections provided some numerical performance and statistical summary but this section concentrates on the relationships between model behavior, availability of MRI modality, and reliability of segmentation.

The first is between the model architecture and missing-modality robustness. The SimCLR and Med3D offer useful baseline performance but are not specifically designed to deal with variable MRI input availability. SimCLR is trained in a self-supervised contrastive learning fashion and Med3D is trained in a supervised 3D medical imaging pipeline. Both models, though, explicitly do not contain a modality-aware fusion mechanism. Consequently, they are unstable if one or more modality is not present.

In contrast, SSL+HeMIS is based on the problem of missing modality itself. The HeMIS-inspired fusion mechanism enables the flexible combination of the modality features available, and the SSL component enhances the feature representations used for subsequent supervised segmentation. This is likely the reason for the improved performance in complete and incomplete modality settings of SSL+HeMIS.

The second important relationship is between the number of available modalities and segmentation performance. In general, performance improves when more MRI modalities are available. Single-modality inputs provide the least information and produce the weakest average segmentation results. Two-modality and three-modality inputs improve performance because they combine complementary tumor information from different MRI sequences. The best result is obtained when all four modalities are available. This pattern confirms that multimodal MRI is important for brain tumor segmentation because no single sequence captures all tumor characteristics equally well.

However, the results also show that modality count alone does not fully determine performance. The identity of the available modality is equally important. T1ce-containing combinations generally perform better than combinations without T1ce. This relationship is clinically meaningful because T1ce highlights contrast-enhancing tumor tissue and is especially important for identifying enhancing tumor and tumor core regions. For example, combinations such as FLAIR + T1ce, T1ce + T2, and FLAIR + T1ce + T2 show strong performance because they include contrast-enhanced information together with edema-sensitive modalities.

FLAIR and T2 also show a strong relationship with whole tumor and edema-related segmentation. These modalities are useful because they highlight abnormal fluid-sensitive regions and peritumoral edema. When FLAIR or T2 is combined with T1ce, the model receives both tumor-core information and surrounding abnormal-tissue information. This helps explain why FLAIR + T1ce + T2 performs almost as well as the all-four-modality setting. In contrast, T1 alone performs poorly because it mainly provides anatomical structure and has weaker direct contrast for tumor subregion separation.

The relationship between modality availability and qualitative segmentation output also supports the numerical findings. In the all-four-modality predictions, tumor boundaries are more consistent with the ground truth, and the model identifies the main tumor regions more completely. When only one modality is available, the segmentation becomes less complete and may miss parts of the tumor.

The all-combination visualization figures are especially useful here because they show how predictions change when specific modalities are removed. These figures demonstrate that the model does not fail completely under missing-modality input, but the quality of segmentation depends strongly on which modalities remain available.

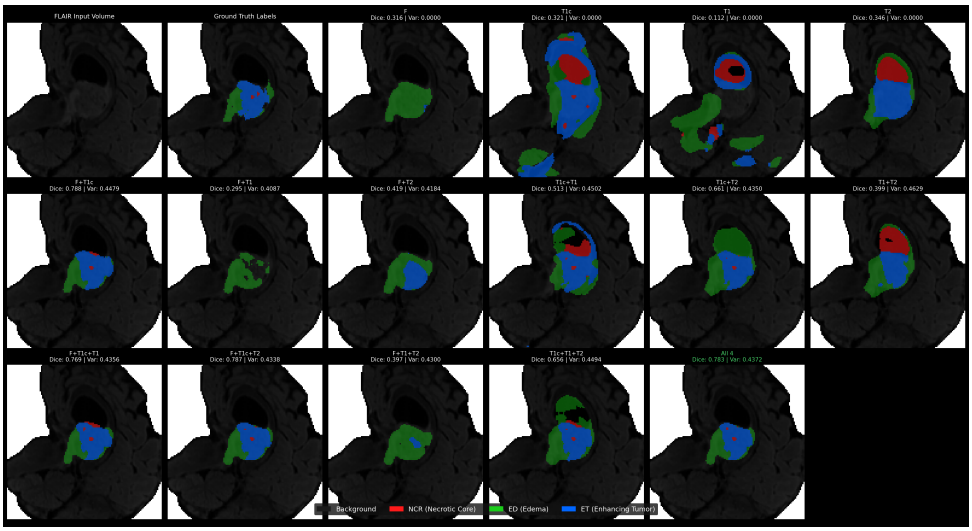


Figure 5.7: Visual comparison of predictions under different modality combinations

The proposed approach is practical and in a fusion direction, while the existing missing modality ideas in the literature are based on modality synthesis direction. There are some techniques which try to produce missing MRI modalities prior to segmentation. Unlike SSL+HeMIS, however, scans are not created when the data is absent. On the contrary, it directly handles the modalities available and combines the features of the modalities statistically. This is easier for implementation in the experimental pipeline, as there is no need for an extra stage of generating images before they are segmented. It also does not show synthetic MRI images as actual clinical images.

Consideration of performance vs. practicality is also important. A model that works

well only with the four modalities is suitable for benchmark datasets, but not as useful in hospital settings. The proposed model can be easily applied to any nonempty modality combinations and thus is more similar to real-world incomplete MRI data. That is not to say that the absence of modalities is a harmless thing, but that there is always an improvement in performance if more modalities are added. The model is more flexible however, since it is also capable of segmentation output if the input protocol is incomplete.

The main comparative relationships are summarized in Table 5.8.

Relationship	**Observation**	**Interpretation**
SSL+HeMIS vs. SimCLR and Med3D	SSL+HeMIS performs better across modality combinations	Modality-aware fusion improves robustness
Modality count vs. Dice score	Dice generally increases as more modalities are available	Multimodal information improves segmentation
T1ce presence vs. ET/TC segmentation	T1ce containing combinations perform strongly	T1ce is important for enhancing tumor and tumor core detection
FLAIR/T2 presence vs. WT segmentation	FLAIR and T2 support edema and abnormal-region detection	These modalities contribute strongly to whole tumor segmentation
T1-only input vs. segmentation quality	T1 alone gives the weakest performance	Anatomical structure alone is insufficient for reliable tumor subregion segmentation
Quantitative results vs. visual outputs	Better Dice scores correspond to more complete visual masks	Qualitative figures support the numerical evaluation

Table 5.8: Summary of key comparisons and relationships

Overall, the comparisons and relationships show that the proposed SSL+HeMIS model is not only the best-performing design among the tested solutions, but also the most suitable for the research objective. The results demonstrate that segmentation quality depends on both the number and type of available MRI modalities. T1ce is highly valuable for tumor core and enhancing tumor segmentation, while FLAIR and T2 provide important information for whole tumor and edema-related regions. The model performs best with all four modalities, but its ability to remain functional under incomplete modality combinations makes it more practical than models that assume complete MRI input.

5.6 Discussion

The experimental findings show that the proposed SSL+HeMIS fusion model successfully addresses the central problem of brain tumor segmentation under incomplete MRI modality conditions. It should be emphasized that missing modalities are simulated by withholding modalities from complete BraTS 2021 cases, not obtained

from real hospital missing-data cases. The results demonstrate that the model performs best when all four MRI modalities are available, but more importantly, it remains functional and comparatively stable when one or more modalities are missing. This is the main strength of the proposed framework because real clinical MRI data may not always contain a complete set of FLAIR, T1ce, T1, and T2 sequences.

The all-four-modality outcome corroborates the significance of comprehensive information obtained by the multimodal MRI. In the event that all modalities are available, the model gets complementary structural, contrast-enhanced, fluid-sensitive, and edema-related information. This enables it to achieve the best performance across all segments. The results for the FLAIR + T1ce + T2 combination, however, were very close to the all-four-modality result, indicating that the model can be fairly robust with loss of T1. This implies that not all modalities are equally useful for segmentation. Both T1 and T2 are found to be useful for anatomical reference, while T1ce, FLAIR, and T2 are found to be more contrastive for the segmentation task, especially when it comes to tumors.

The modality-wise behaviour also has an important clinical interpretation. T1ce was the best single modality as it revealed the tumor core and showed tumor tissue with contrast enhancement. The other two that were important were FLAIR and T2, which highlight edema and abnormal tissue areas that are useful for whole tumor segmentation. On the contrary, the weakest performance was obtained using T1 only, revealing that this anatomical structure is insufficient for reliable segmentation of tumor subregions. Thus, the results corroborate the hypothesis that the performance of the model is dependent on both the number of modalities that are available and the clinical importance of the specific modalities that are present.

Comparing the final design to SimCLR and Med3D further supports the argument for the final design. The two baselines can be used to gain insight into other design directions, but they are not specifically developed for segmentation using the missing-modality. Self-supervised feature learning is performed by SimCLR, and volumetric 3D medical image representation is performed by Med3D, however, there is no dedicated mechanism to fuse only the available MRI modalities. The proposed SSL+HeMIS model is a combination of representation learning and modality-aware fusion, hence its high performance under the 15 modality combinations. This helps to confirm the final choice of SSL+HeMIS as the most appropriate design for this problem addressed in this thesis.

It is also worth pointing out that the qualitative outputs are in line with the results obtained from the quantitative approach. Complete or strong modality combinations predict a better tumor coverage and consistent tumor boundaries. In contrast, weaker combinations, particularly those that lack T1ce and/or those that use only T1, exhibit greater segmentation that is not complete. The results of the Dice-based analyses are consistent with this visual behaviour and give an idea of the meaning of the numerical scores. The segmentation figures (Figure 5.2 & Figure 5.3) are thus important in demonstrating that the model is not just performing better numbers, but is also creating more discernable tumor masks in the presence of more informative modalities.

The results also reveal the importance of evaluating the medical AI models in realistic settings. The evaluation would have demonstrated good segmentation performance, but would not have evaluated the real world missing-modality problem if the model had only been tested with all four modalities. The study compares all 15 non-empty modality combinations, giving a more comprehensive view of the model between different modality combinations behavior in situations of partial stimulus. This makes the analysis more applicable to the clinical settings where imaging protocols can vary between hospital, patient, machine and resources.

The results, however, should be taken with the necessary caution. The proposed model is tested on a well-known strong benchmark dataset, the BraTS 2021 dataset, but still a controlled research dataset. The findings are not clinical direct proof for a clinical use. There are other potential issues that might arise from real hospital data, including the differences in scanners, motion artifacts, incomplete acquisition protocol, unusual tumor appearances, and demographic differences. So, there would be a need for external validation on independent clinical data sets before any claim towards deployment.

Another restriction is that the uncertainty analysis performed using HeMIS variance should not be misinterpreted. The use of variance is useful since it reflects dis-agreement between the modalities available, but it is not yet a clinically calibrated uncertainty score. It should be noted that variance in this experiment is not an alternative to the confidence calibration or expert review but is a feature-disagreement signal that is used to assist analysis. In a future version of this work, more robust methods for uncertainty calibration, such as reliability diagrams, expected calibration error or patient-level uncertainty validation using segmentation failure cases should be incorporated.

Limited model comparison is also a concern. While SimCLR and Med3D were added as meaningful baseline designs, they were not designed with the same goal of missing modality as SSL+HeMIS. Thus, the comparison should be understood as a robustness stress test and not a true architectural comparison. The proposed SSL+HeMIS still preserves its appropriateness for the research objective, but further research can be conducted to compare it with other missing-modality specific models such as the models of modality synthesis, transformer-based multi-modal fusion models, and other variants of HeMIS.

The proposed model from a practical aspect is very relevant as it is learned with a single framework to process different modality combinations. This would be more efficient than having individual models for all possible missing-modality scenarios. Further, it makes the system more flexible for practical applications where availability of the MRI sequence can vary among patients. Concurrently, the results indicate quite clearly that whenever possible, full imaging is still desirable. The model should thus be interpreted as a robustness-oriented system of support, rather than anything to justify omitting clinically useful MRI sequences.

In general, the discussion suggests the proposed SSL+HeMIS framework is mean-

ingful in terms of maintaining a balance between segmentation performance and missing modality robustness. The model is best when trained on full multimodal inputs, and when trained with partial inputs it is usable, and more robust than baselines evaluated. The results validate that self-supervised representation learning with HeMIS-based modality fusion can enhance the applicability of 3D brain tumor segmentation tasks.

Chapter 6

Conclusion

6.1 Summary of Findings

This research developed and evaluated an SSL+HeMIS framework for 3D brain tumor segmentation using multimodal MRI data. The main objective was to assess whether self-supervised pretraining combined with modality-aware fusion could support segmentation when one or more MRI modalities are artificially withheld during evaluation. The model was trained and evaluated using the BraTS 2021 dataset with four MRI modalities: FLAIR, T1ce, T1, and T2.

The experimental results show that the proposed model sustained its performance when multiple modalities were unavailable. This confirms that complete multimodal MRI input provides the strongest segmentation outcome. However, the model also maintained usable performance under missing-modality conditions. Among the incomplete combinations, FLAIR + T1ce + T2 produced results very close to the all-four-modality setting, indicating that the absence of T1 caused only a small reduction in performance.

The findings also show that segmentation performance depends not only on the number of available modalities but also on the type of modality present. T1ce was the strongest single modality because of its importance in identifying enhancing tumor and tumor core regions. FLAIR and T2 were also useful for edema and whole tumor segmentation, while T1 alone produced the weakest result. This indicates that anatomical information alone is not sufficient for reliable tumor subregion segmentation.

SSL+HeMIS outperformed the other methods in terms of average Dice on the 15 modality combinations. According to the results, the HeMIS fusion strategy improved the model's stability in the presence of varying input signals. The results, overall, validate that the proposed SSL+HeMIS model was effective for the desired task, particularly when the modality is missing. More validation with other data and real clinical cases is needed before a practical deployment.

6.2 Contributions to the Field

This work in the field of medical image segmentation focuses on brain tumor segmentation in the presence of missing modality MRI. Many segmentation methods make the assumption that both training and inference are performed with all modalities of MRI available. In a real medical situation, however, it may be impossible to use a modality because of differences in imaging protocols, acquisition limitations, patient movements, or resource constraints. In this study, this practical challenge is investigated by testing the model on all 15 non-empty combinations of FLAIR, T1ce, T1, and T2.

Primarily, this paper introduces a novel work on the fusion of modality information, through the concept of HeMIS, with Self-Supervised Learn-ing (SSL). The self-supervised pretraining stage is used to pre-train a useful representation of the volumetric MRI data prior to supervised segmentation training, while the HeMIS fusion component allows the network to fuse only the available modality features. This results in a single, integrated segmentation system that works well across different modalities where one or more modalities is missing, without the need for a specific model for each modality combination.

The work also has important contributions to offer by providing detailed comparisons to the baselines of SimCLR and Med3D. As observed, SSL+HeMIS has better performance than the other methods and the average Dice values is higher with the 15 modality combinations, thus showing the benefit of the missing-modality-aware fusion strategy used for this task. In addition, the results of the modality combination analysis indicate the relative importance of individual MRI sequences. T1ce is found to be the most important sequence for better segmentation of tumor and tumor core, while FLAIR and T2 are the most important sequences for better segmentation of the edema and whole tumor regions.

Still another contribution is that both quantitative and qualitative evaluation is used to better understand model behavior within limited input configurations. The results of the study are presented as label-wise Dice scores, clinical region Dice scores, precision, recall, F1-score, correlation analysis, confusion matrix analysis and visual segmentation results. This provides a more general and comprehensive understanding of the effect of removing different modalities on overall segmentation quality.

Finally, Our contribution is a single SSL+HeMIS framework that performs brain tumor segmentation across all 15 non-empty modality combinations of FLAIR, T1ce, T1, and T2 using one trained model. The framework combines masked image modeling based self-supervised pretraining with HeMIS-inspired mean-variance modality fusion, and it is evaluated on BraTS 2021 under simulated missing-modality conditions.

6.3 Recommendations for Future Work

There are several directions of future research that could benefit the current study. The current evaluation could be enhanced with the use of full 5-fold cross validation

instead of just a single fold (fold 0), for a more robust and reliable estimation of the model generalization performance over the entire BraTS 2021 dataset. There should also be a structured division of the data set for training, validation and final testing, which further ensures that the data is properly separated from experiments and reduces bias in the evaluation.

Second, cross-dataset evaluation should be performed on newer or other datasets, such as those in BraTS 2022, to evaluate the potential of the proposed SSL+HeMIS framework to generalize its learning beyond the distribution used for training. Third, the number of training epochs for the self-supervised pretraining stage can be increased in order to reduce the reconstruction loss of the masked image modelling task, thereby obtaining a better MRI representation. In addition, higher capacities can be achieved using a deeper transformer encoder, higher embedding dimensions, and more attention heads with the availability of more computational resources.

Fourth, the framework can be extended to incorporate tumor grading together with the segmentation framework. The existing model identifies subregion segmentation masks, but does not explicitly perform feature extraction of directly relevant clinical features to the future clinical grading, such as exact subregion volumes, tumor texture, or internal tissue heterogeneity. A possible study for the future is to use the global features of the self-supervised learning (SSL) encoder to feed into a classification head to perform segmentation and tumor grading simultaneously in a multi-task learning approach.

Finally, simulation should be done with a large number of real clinical missing-modality cases, including MRI sequences that are missing because of a patient's motion artifacts, stringent imaging time constraints, or the availability of scanners. This validation still needs to be done prior to clinical use.

Bibliography

[1] I. Havaei, H. Larochelle, P. Jodoin, G. Hamarneh, and A. Mccallum, "Brain tumor segmentation on mri with missing modalities," *arXiv preprint arXiv:1904.07290*, 2019. [Online]. Available: https://arxiv.org/pdf/1904.07290.

[2] A. Mondal, N. Islam, M. H. Shovon, M. M. Rahman, and M. M. Reza, "Ssclnet: A self-supervised contrastive loss-based pre-trained network for brain mri classification," *IEEE Access*, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/10018340.

[3] S. Shurrab and R. Duwairi, "Brain tumor segmentation and edge detection using self-supervised learning," *PeerJ Computer Science*, vol. 10, e1045, 2022. [Online]. Available: https://peerj.com/articles/cs-1045.pdf.

[4] Y. Li, Y. Gao, D. Zhang, and S. Pan, "Advancing medical image segmentation with self-supervised learning: A 3d student-teacher approach for cardiac and neurological imaging," *OpenReview Preprint*, 2023. [Online]. Available: https://openreview.net/pdf?id=PBdGho6avn.

[5] B. H. Menze and S. Bakas, "The brain tumor segmentation–metastases (brats-mets) challenge 2023: Brain metastasis segmentation on pre-treatment mri," *arXiv preprint arXiv:2501.11755*, 2023. [Online]. Available: https://arxiv.org/abs/2501.11755.

[6] D. M. H. Nguyen, H. Nguyen, N. T. Diep, T. N. Pham, T. Cao, B. T. Nguyen, P. Swoboda, N. Ho, S. Albarqouni, P. Xie, D. Sonntag, and M. Niepert, "Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching," in *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/58cc11cda2a2679e8af5c6317aed0af8-Paper-Conference.pdf.

[7] K. Wang, Z. Li, H. Wang, S. Liu, M. Pan, M. Wang, S. Wang, and Z. Song, "Improving brain tumor segmentation with anatomical prior-informed pre-training," *Frontiers in Medicine*, vol. 10, p. 1 211 800, 2023. DOI: 10.3389/fmed.2023.1211800. [Online]. Available: https://doi.org/10.3389/fmed.2023.1211800.

[8] W. Zhang, M. Li, H. Li, M. Xu, and Y. Yang, "Self-supervised wavelet-based attention network for semantic segmentation of mri brain tumor," *Sensors*, vol. 23, no. 5, p. 2719, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/5/2719.

[9] M. Alam, M. I. Khan, M. T. Rahman, F. Ahmed, and M. S. Rahman, "Advancements in deep learning techniques for brain tumor segmentation: A survey," *Biomedical Signal Processing and Control*, vol. 87, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914824001321.

[10] M. Chen, M. Zhang, L. Yin, L. Ma, R. Ding, T. Zheng, Q. Yue, S. Lui, and H. Sun, "Medical image foundation models in assisting diagnosis of brain tumors: A pilot study," *European Radiology*, vol. 34, no. 12, pp. 6667–6679, 2024. DOI: 10.1007/s00330-024-10728-1. [Online]. Available: https://doi.org/10.1007/s00330-024-10728-1.

[11] M. Karagoz and G. Fox, "Residual vision transformer (resvit) based self-supervised learning model for brain tumor classification," *arXiv preprint arXiv: 2411.12874*, 2024. [Online]. Available: https://arxiv.org/pdf/2411.12874.

[12] C. Lin, Y. Wang, Y. Lu, H. Li, Y. Chen, and W. Zhang, "Enhanced self-supervised learning for multi-modality mri segmentation and classification," *arXiv preprint arXiv:2407.10377*, 2024. [Online]. Available: https://arxiv.org/pdf/2407.10377.

[13] R. Shah and S. Rathi, "Cu-net: A u-net architecture for efficient brain-tumor segmentation on brats 2019 dataset," *arXiv preprint arXiv:2406.13113*, 2024. [Online]. Available: https://arxiv.org/pdf/2406.13113.

[14] P. Singh, A. Das, P. Bhardwaj, R. Jain, and N. Saxena, "A review of self-supervised, generative, and few-shot deep learning methods for data-limited mri segmentation," *NMR in Biomedicine*, 2024. [Online]. Available: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/nbm.5143.

[15] J. Zhao, J. Cai, C. Sun, H. Yang, Q. Zheng, and Z. Zhou, "Self-supervised learning framework application for medical image analysis: A review and summary," *BioMedical Engineering OnLine*, vol. 23, no. 1, 2024. [Online]. Available: https://link.springer.com/content/pdf/10.1186/s12938-024-01299-9.pdf.

[16] M. Bhattacharya, S. Gupta, A. Singh, C. Chen, G. Singh, and P. Prasanna, "Brainmrdiff: A diffusion model for anatomically consistent brain mri synthesis," *arXiv preprint arXiv:2504.04532*, 2025. [Online]. Available: https://arxiv.org/pdf/2504.04532.

[17] M. Rahman, S. Islam, M. Khan, M. Mia, M. Shovon, and N. Rifat, "Self-supervised learning approach for early detection of rare neurological disorders in mri data," *Journal of Information Security and Intelligence Systems*, vol. 25, no. 1, 2025. [Online]. Available: https://jisis.org/wp-content/uploads/2025/04/2025.I1.010.pdf.

[18] L. Robinet, A. Berjaoui, and E. C.-J. Moyal, "Multimodal masked autoencoder pre-training for 3d mri-based brain tumor analysis with missing modalities," *arXiv preprint arXiv:2505.00568*, 2025. [Online]. Available: https://arxiv.org/abs/2505.00568.

[19] X. Wang, C. Li, X. Zhang, C. Su, and X. Li, "Feasibility study of detecting and segmenting small brain tumors in a small mri dataset with self-supervised learning," *Diagnostics*, vol. 15, no. 3, p. 249, 2025. [Online]. Available: https://www.mdpi.com/2075-4418/15/3/249.

[20] Y. Xu, L. Yu, S. Wang, and X. Xu, "A generalizable 3d framework and model for self-supervised learning in medical imaging," *arXiv preprint arXiv:2501.11755*, 2025. [Online]. Available: https://arxiv.org/pdf/2501.11755.